



European  
Commission

# New Techniques and Technologies for Statistics 2015

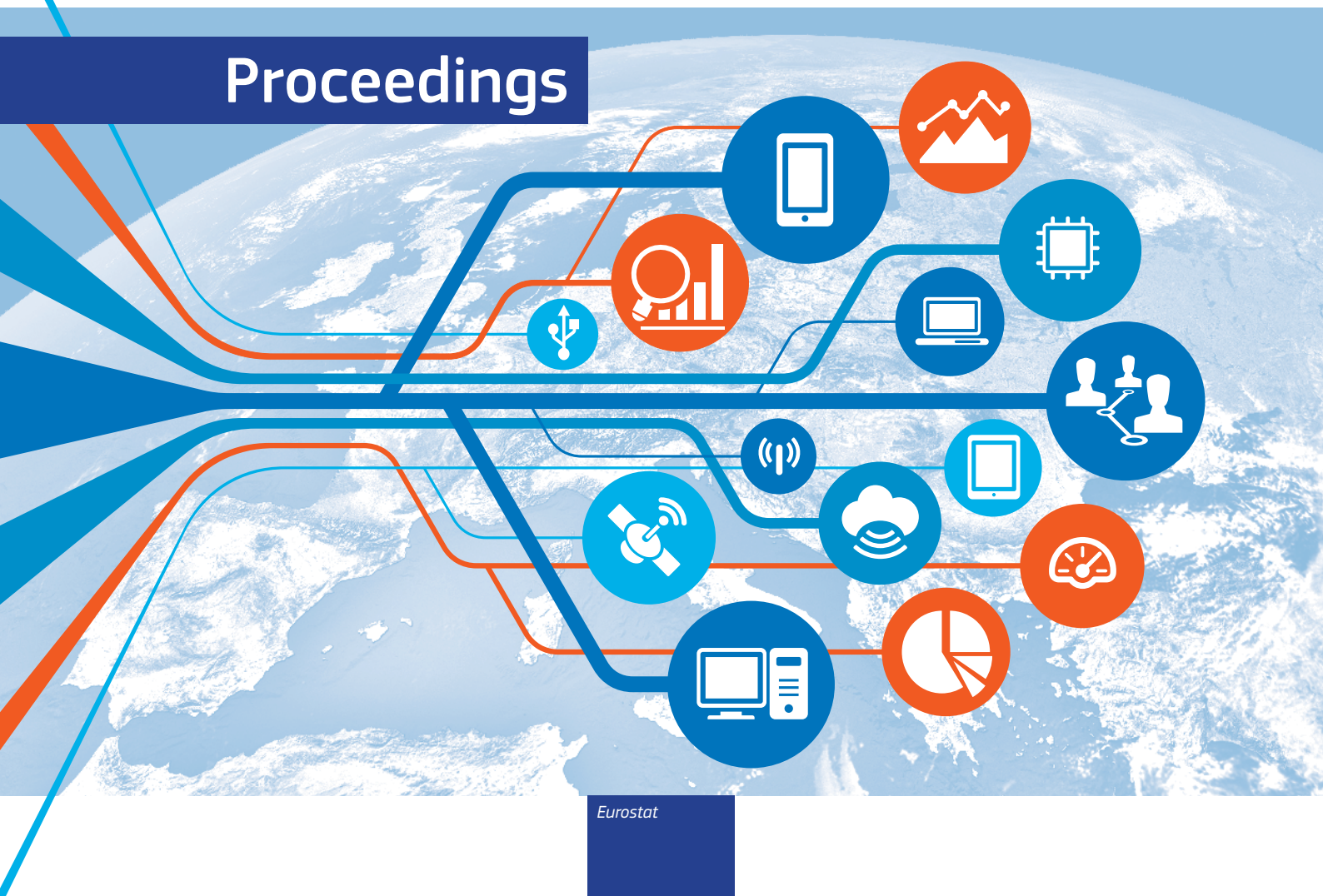
Reliable Evidence for a Society in Transition

Brussels

**9-13 March 2015**

[www.NTTTS2015.eu](http://www.NTTTS2015.eu)

Proceedings



Eurostat

## Monday 9 March 2015

13:00–18:00

### SATELLITE WORKSHOP ON BIG DATA

JENK

**Workshop organisers:** Piet Daas, Statistics Netherlands; Matjaz Jug, CBS; Pilar Rey del Castillo, Eurostat; Marina Signore, Istat; Peter Struijs, Statistics Netherlands; Carlo Vaccari, Istat; Steven Vale, UNECE; Antonino Virgillito, Istat  
**Workshop chair:** Steven Vale, UNECE

## Tuesday 10 March 2015

08:30–09:30

### Welcome coffee

L2

09:30–12:00

### Session 1A – OPENING SESSION

GASP

**Chair:** Walter Radermacher, Eurostat

**Moderators:** Roberto Barcellan & Martin Karlberg, Eurostat

Opening address by **Walter Radermacher**, Director-General, Eurostat

Research and statistics – the JRC perspective.

Keynote address by **Professor Vladimír Šucha**, Director-General, Joint Research Centre, European Commission

Official Statistics in the Next Decade, Some Outstanding Issues.

Keynote address by **Professor Danny Pfeffermann**, University of Southampton and Central Bureau of Statistics in Israel

Data Journalism and Visualization: A Truthful Art.

Keynote address by **Alberto Cairo**, Knight Chair in Journalism at the School of Communication of the University of Miami

12:00–13:30

### Lunch break

L2

### Session 2P – POSTER PRESENTATION

L2

Big Data: Google searches predict unemployment in Finland

Joonas Tuhkuri, The Research Institute of the Finnish Economy and the University of Helsinki

Collecting, storing and managing fuzzy data in statistical relational databases

Miroslav Hudec, Dušan Praženka, University of Economics in Bratislava, Faculty of Economic Informatics, Bratislava, Slovakia

Self-employment – a complementary solution for full use of EU labour potential

Cristescu Amalia, Dorel Ailenei, Bucharest University of Economic Studies

Using a Poisson Regression Model to analyze woman's Labor Force data

Reem Ismail Mohamed Elsybaey, Central Agency for Public Mobilization and Statistics, Cairo, Egypt

Non-Extensive Entropy Econometrics and CES production Model: Some EU Country Case Study

Second Bwanakare, University of information Technology and Management of Rzeszow, Poland

The evolution of the disparities among the EU member states regarding FDI – the case of the former communist countries

Vasile Alecsandru Strat, Bucharest University of Economic Studies, Department of Statistics and Econometrics

13:30–14:15

### Session 3A – CALIBRATED BAYES: AN ATTRACTIVE FRAMEWORK FOR OFFICIAL STATISTICS IN THE 21<sup>ST</sup> CENTURY

GASP

**Keynote address by Roderick J. Little, Professor of Biostatistics, University of Michigan**

**Chair:** Caterina Giusti, University of Pisa

**Moderator:** Dario Buono, Eurostat

14:30–15:30

### Session 4A – CORPORATE, INTERNATIONAL AND CROSS-SECTOR BIG DATA INITIATIVES

GASP

**Chair:** Albrecht Wirthmann, Eurostat

**Moderator:** Christophe Demunter, Eurostat

Guidelines for statistical organisations when forming Big Data partnerships

Task Team on Big Data Partnerships

A Shared Computation Environment for International Cooperation on Big Data

Matjaz Jug, UNECE; Carlo Vaccari, Istat; Antonino Virgillito, UNECE, Istat

The Office for National Statistics – Big Data Project

Jane Naylor, Nigel Swier, Susan Williams, ONS

### Special 4B – NON-PROBABILITY SAMPLING Session

MANS

**Session organiser/chair:** Silvia Biffignandi, University of Bergamo

**Panel:** Roderick J. Little, University of Michigan; Dan Hedlin, Stockholm University; Beat Hulliger, University of Northwestern Switzerland (FHNW)

On Model-Representativeness

Beat Hulliger, University of Northwestern Switzerland (FHNW), Switzerland

Is random sampling necessary?

Dan Hedlin, Stockholm University

Round table discussion : What is the role of data from non-probability samples in official statistics in the future?

Roderick J. Little, University of Michigan; Dan Hedlin, Stockholm University; Beat Hulliger, University of Northwestern Switzerland (FHNW)

**Session 4C – DISCLOSURE CONTROL, SYNTHETIC DATA AND RECORD LINKAGE**

JENK

**Chair:** Michael Carlson, Stockholm UniversityDisclosure Risk Measurement with Entropy in Sample Based Frequency Tables  
László Antal, Natalie Shlomo, Mark Elliot, University of Manchester, UKPrivacy Preserving Probabilistic Record Linkage  
Duncan Smith, Natalie Shlomo, University of Manchester, UKDevelopment of pseudonymised matching methods for linking multiple administrative datasets  
Pete Jones, ONSGenerating synthetic geocoding information for public release  
Monika Jingchen Hu, Duke University; Jörg Drechsler, Institute for Employment ResearchAutomated methods for providing bespoke synthetic data for the UK Longitudinal Studies  
Beata Nowok, Gillian M. Raab, Chris Dibben, Administrative Data Research Centre – Scotland (ADRC-S), University of Edinburgh

15:30–16:00

**Coffee break**

L2

16:00–17:00

**Special 5A – THE ROLE OF VISUALISATION IN DATA DISSEMINATION Session**

GASP

**Session organiser:** Marina Signore, Istat**Chair:** Emanuele Baldacci, Istat**Moderator:** Britta Gauckler, EurostatAssembling information: dynamic dashboards for actionable data analytics  
Cesare Furlanello, Fondazione Bruno KesslerA more visual dissemination to attract new users  
Philippe Bautier, Bernard Le Goff, EurostatUncharted Territories – Data driven graphics beyond the basics  
Michael Neutze, DestatisAn enhanced visualization service based on geospatial and statistical linked open data  
Monica Scannapieco, Pina Grazia Ticca, Istat**Session 5B – SURVEY AND QUESTIONNAIRE DESIGN; MANAGEMENT OF NON-RESPONSE**

MANS

**Chair:** Jan Bjørnstad, Statistics NorwayTargeted designs in surveys and longitudinal studies  
Annamaria Bianchi, Silvia Biffignandi, University of Bergamo, ItalyImproving the response rates in business surveys. The case of LCS 2012  
Ciro Baldi, Marilena Angela Ciarallo, Stefano De Santis, Rossana Renzi, Graziella Spera, IstatImputation under edit restrictions and known totals  
Ton de Waal, Statistics Netherlands and Tilburg University; Wieger Coutinho, Loket Aangepast-Lezen; Natalie Shlomo, University of ManchesterCalibration for Nonresponse Treatment: in One or Two Steps?  
Per Gösta Andersson, Stockholm University; Carl-Erik Särndal, Professor Emeritus at Statistics Sweden**Special 5C – INFORMATION STANDARDS FOR STATISTICS: THE 2020 VISION Session**

JENK

**Session organiser:** Marco Pellegrino, Eurostat**Chair:** Roberto Barcellan, EurostatProgress in sharing statistical data and metadata using international standards  
Francesco Rizzo, IstatA new international standard for data validation and processing  
Marco Pellegrino, EurostatA metadata-driven process for handling statistical data end-to-end  
Denis Grofils, Eurostat

17:15–18:15

**Session 6A – BIG DATA SOURCES: WEB SCRAPING AND SMART METERS**

GASP

**Chair:** Pilar Rey del Castillo, Eurostat**Moderator:** Fernando Reis, EurostatUsing Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies  
Giulio Barcaroli, Monica Scannapieco, Donato Summa, Istat; Marco Scarnò, CinecaAutomatic price collection on the Internet (web scraping)  
Ingolf Boettcher, Statistics AustriaUsing web scraped data to construct consumer prices indices  
Robert Breton, Nigel Swier, Robert O'Neill, ONSModelling sample data from smart-type meter electricity usage  
Susan Williams, ONS

## Special 6B – PROGRESS IN THE PRODUCTION AND POLICY USE OF GDP Session AND BEYOND INDICATORS

MANS

**Session organisers:** Marianne Paasi, Directorate-General Research and Innovation, European Commission & Marina Signore, Istat  
**Chair:** Marianne Paasi, Directorate-General Research and Innovation, European Commission

From the construction to the usage of statistics beyond GDP

Marina Signore, Donatella Fazio, Istat

The potential of Web2.0 communities for statistics

Donatella Fazio, Istat; Katherine Scrivens, OECD; Maria Grazia Calza, Istat

Recent Advances in the measurement of Intangible assets

Mary O'Mahony, King's College London; Carol Corrado, The Conference Board; Jonathan Haskel, Imperial College London; Cecilia Jonan Lasinio, LUISS and Istat

## Session 6C – ENTERPRISE ARCHITECTURE AND INTEGRATION

JENK

**Chair:** Jean-Marc Museux, Eurostat

A European Statistical System Enterprise Architecture Reference Framework

Task Force on ESS Enterprise Architecture

CORE: a concrete implementation of the CSPA architecture

Mauro Bruno, Rolando Duma, Monica Scannapieco, Marco Silipo, Giulia Vaste, Istat

On the Development of a CSPA Error Correction Service: Design and Implementation Issues

Donato Summa, Marco Silipo, Monica Scannapieco, Diego Zardetto, Mauro Bruno, Istat

Reversing the flow: from an integrated system of administrative microdata to an infrastructure for the users

Simone Ambroselli, Giuseppe Garofalo, Istat

Dealing with measurement and integration errors in administrative data: the case of the Italian multi-source system on small and medium enterprises

Orietta Luzi, Marco Di Zio, Ugo Guarnera, Roberta Varriale, Istat

18:30-21:00

Welcome dinner in the "Piazza" restaurant in the Berlaymont Building

# Wednesday 11 March 2015

09:00-09:45

## Session 7A – HUMAN BEHAVIOR MODELING FROM (BIG) MOBILE DATA Keynote address by Nuria Oliver, Scientific Director, Telefonica Research, Spain

GASP

**Chair:** Michail Skaliotis, Eurostat

**Moderator:** Fernando Reis, Eurostat

09:45-10:15

Coffee break

L2

10:15-11:15

## Special 8A – MOBILE PHONE DATA AS A SOURCE FOR OFFICIAL STATISTICS Session

GASP

**Session organiser:** Rein Ahas, University of Tartu

**Moderator:** Bogomil Kovachev, Eurostat

Defining usual environment with mobile positioning data

Rein Ahas, Janika Raun, University of Tartu; Margus Tiru, Positium LBS, Estonia

Using mobile positioning data for official statistics: daydream nation or promised land?

Demunter Christophe, Reis Fernando, Eurostat

Using Passive Mobile Positioning Data in Tourism and Population Statistics

Laura Altin, Positium, University of Tartu; Margus Tiru, Erki Saluveer, Positium; Anniki Puura, University of Tartu

## Session 8B – RECORD LINKAGE AND STATISTICAL MATCHING

MANS

**Chair:** Natalie Shlomo, University of Manchester

Indicator for the representativeness of linked sources

Dingeman Jan van der Laan, Bart F. M. Bakker, Statistics Netherlands

New proposals for linkage error estimation

Tiziana Tuoto, Niki Stylianidou, Istat

The use of uncertainty to choose the matching variables in statistical matching

Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Istat

Environmental conditions / behaviour and income – statistical matching of EU-SILC and micro-census Environment

Alexandra Wegscheider-Pichler, Statistics Austria

The role of the auxiliary information in Statistical Matching Income and Consumption

Gabriella Donatiello, Marcello D'Orazio, Doriana Frattarola, Antony Rizzi, Mauro Scanu, Mattia Spaziani, Istat



## Session 8C – SEASONAL ADJUSTMENT, TEMPORAL DISAGGREGATION AND FORECASTING

JENK

**Chair:** Dario Buono, Eurostat

Improved Time-varying Day Adjustment in SEASABS

Jonathan Campbell, Lujuan Chen, Australian Bureau of Statistics

1 out of 20 possible scenarios: how to perform temporal disaggregation of annual sector accounts data

Filippo Gregorini, Dario Buono, Enrico Infante, Eurostat

Forecasting Evaluation with JDemetra+

De Antonio Liedo, Jean Palate, National Bank of Belgium

Simple forecasting techniques can reduce forward-series bias and keep revisions low for benchmarked Quarterly National Accounts estimates

Geoffrey Brent, Alex Stuckey, Australian Bureau of Statistics

Does web anticipate stocks? Analysis for a subset of systemically important banks

Michela Nardo, Erik van der Goot, Joint Research Centre, European Commission

11:30-12:30

## Session 9A – LUNCHTIME SPEED DATING PART I: LIGHTNING BOLT PRESENTATIONS

GASP

**Chair:** Marina Signore, Istat

Building a Cross Border Data Access System for Improved Scholarship and Policy: The Case of the German IAB Network of RDCs

Jörg Heining, IAB; William C. Block, Warren A. Brown, Cornell Institute for Social and Economic Research; Stefan Bender, IAB

Microaggregation for the masses: non-confidential enterprise-level data for analytical and research purposes

Sebastien Perez-Duarte, Pierre Lamarche, European Central Bank

Showing Uncertainty of Official Statistics

Edwin de Jonge, D. Jan van der Laan, Jessica Solcer, Statistics Netherlands

Bayesian estimation approach in frameworks, integration of compilation and analysis

Jan W. van Tongeren, previously statistics UN division; Ruud Picavet, previously Tilburg University

From Hombar to Ear. Evolution of data processing in Hungary

Eva Laczka, Hungarian Central Statistical Office

Robust Variable Estimation by Combining Administrative data sources

Guy Vekeman, Statistics Belgium

Constructing Structural Earnings Statistics from Administrative Datasets

Kevin McCormack, Mary Smyth, Central Statistics Office Ireland

The sensitivity of job reallocation measures to longitudinal linkage problems

Karen Geurts, University of Leuven

Occupational mismatch impact on earnings

Monica-Mihaela Maer Matei, Bucharest University

New Approach to Gross Domestic Product Decomposition

Ante Rozga, Elza Jurun, University of Split; Ivan Šualto, Zagreb School of Economics and Management

Backward recalculation of labor force indicators: a case of Turkey

Necmettin Alpay Kocak, Özlem Yigit, Enes Ertad Uslu, Turkstat

Backcalculating MIP (Macroeconomic Imbalances Procedure) indicators to improve data coverage: an empirical approach

Rosa Ruggeri Cannata, Ferdinando Biscosi, Dario Buono, Eurostat

Use of register data for prior waves of EU-SILC in Austria: the case of "back-calculation"

Richard Heuberger, Thomas Glaser, Statistics Austria

Pitfalls of regression modelling with complex survey data

Florian Ertz, Ralf Thomas Münnich, University of Trier

## Session 9B – PRIVACY AND ACCESS TO BIG DATA AS WELL AS OTHER CONFIDENTIAL DATA

MANS

**Chair:** Aleksandra Bujnowska, Eurostat

A Suggested Framework for National Statistical Offices for Assessing and Managing Privacy Risks Related to the Use of Big Data  
Task Team on Big Data Privacy

Using Research Data Centres (RDCs) to access Big Data

David Schiller, Anja Burghardt, Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute of Employment Research (IAB)

CASD-TeraLab, a secure remote access system to confidential Big Data: description, demonstration and use cases

Alexandre Marty, GENES; Frank Cotton, INSEE; Kamel Gadouche, Nawrès Guédria, GENES

A gateway to European Research Services

Anja Burghardt, David Schiller, IAB

## Special 9C – PRINCIPAL EUROPEAN ECONOMIC INDICATORS (PEEIs) Session

JENK

**Session organiser:** Gian Luigi Mazzi, Eurostat

**Chair:** Roberto Barcellan, Eurostat

An integrated system for euro area and member states turning point detection

Gian Luigi Mazzi, Eurostat; Monica Billio, University of Venice; Laurent Ferrara, Banque de France

A new graphical tool for business cycle monitoring

Jacques Anas, Analysing Cycles in Economies; Ludovic Calès, HENDYPLAN; Gian Luigi Mazzi, Eurostat

Towards a daily indicator of economic conditions

Massimiliano Marcellini, Bocconi University Milan; Claudia Foroni, Norges Bank; Gian Luigi Mazzi, Eurostat;

Fabrizio Venditti, Bank of Italy

12:30-14:00

## Lunch break

L2

12:30-14:00

## Session 10P – LUNCHTIME SPEED-DATING PART II: POSTER PRESENTATIONS

L2

For a list of posters presented please refer to session 9A

13:00-13:55

## Special 10B – REMOTE ACCESS Session

MANS

**Session organiser/chair:** Roxane Silberman, Réseau Quetelet

Remote Access to European Microdata

Maurice Brandt, Destatis

14:00-14:45

## Session 11A – RECENT ADVANCES IN SURVEY SAMPLING

GASP

**Keynote address by Yves Tillé, Professor, University of Neuchâtel**

**Chair:** Beat Hulliger, University of Northwestern Switzerland (FHNW)

**Moderator:** Jean-Marc Museux, Eurostat

15:00-16:00

## Special 12A – THE WORLDWIDE MODERNISATION OF POPULATION CENSUSES Session

GASP

**Session organisers:** Fritz Scheuren, NORC at the University of Chicago & Anders Holmberg, Statistics New Zealand

**Chair:** Fritz Scheuren, NORC at the University of Chicago

**Moderator:** Britta Gauckler, Eurostat

Use of Administrative Sources for Censuses – Merits and Challenges

Lars Thygesen, Statistics Denmark

The Irish Statistical System and the emerging Census Opportunity

John Dunne, Central Statistics Office Ireland

How to change a traditional census? Options and progress on the use of administrative data in New Zealand's census transformation programme

Christine Bycroft, Statistics New Zealand

## Session 12B – ESTIMATION

MANS

**Chair:** Didier Dupré, Eurostat

Variance Estimation in Complex Sampling Designs: The Finite Population Bootstrap Using Pseudo-Populations

Quatember Andreas, Johannes Kepler, University Linz

An R Library to construct empirical likelihood confidence intervals for complex estimators

Yves Berger, University of Southampton

Design-based confidence intervals and significance test for regression parameters using an empirical likelihood approach

Melike Oguz Alper, Yves G. Berger, University of Southampton

A multivariate Regression Estimator for Rotating Surveys

Karen Caruana, Yves G. Berger, University of Southampton

Estimation from Contaminated Multi-Source Data Based on Latent Class Models

Ugo Guarnera, Roberta Varriale, Istat

**Session 12C – NATIONAL ACCOUNTS AND INDICATORS****JENK****Chair:** John Verrinder, Eurostat

A web-semantic data-warehouse approach to the compilation of national accounts: a test case on European National Accounts  
 Francois Libeau, Hendyplan; Roberto Barcellan, Eurostat; Bo Sundgren, Stockholm University; Dominique Ladiray, INSEE; Boris Motik, Oxford University

An Analysis of Household Debt using the Linkage between Micro and Macro Balance Sheet data  
 Juha Honkkila, Ilja Kristian Kavonius, Statistics Finland

Households in Europe in years of economic crisis  
 Leonidas Akritidis, Filippo Gregorini, Eurostat

Summarizing Data using Partially Ordered Set Theory: An Application to Fiscal Frameworks in 97 Countries  
 Julia Bachtrögl\*, Harald Badinger\*, \*\*; Aurélien Fichet de Clairfontaine, Wolf Heinrich Reuter\*  
 \*Vienna University of Economics and Business, \*\*Austrian Institute of Economic Research

A step towards communicating with indicators  
 Justyna Gustyn, Central Statistical Office of Poland

**16:00–16:30****Coffee break****L2****16:30–17:45****Session 13A – BIG AND GEOSPATIAL DATA****GASP****Chair:** Piet J.H. Daas, Statistics Netherlands**Moderator:** Albrecht Wirthmann, Eurostat

Geostatistics Portal – a platform for statistical data geovisualization  
 Mirosław Migacz, Central Statistical Office of Poland

Experiences using LUCAS data in Finnish Land Cover monitoring – Current activities and future plans  
 Markus Törmä, Elise Järvenpää, Pekka Härmä, Lena Hallin-Pihlatie, Suvi Hatunen, Minna Kallio, Finnish Environment Institute SYKE

Forecasting skyrocketing unemployment with big data  
 Maria Rosalia Vicente, Ana Jesús López, Rigoberto Pérez, University of Oviedo

High frequency road sensor data for official statistics  
 Marco Puts, Piet Daas, Martijn Tennekes, Statistics Netherlands

Projection of road sensors to the Dutch road network  
 Martijn Tennekes, Marco Puts, Statistics Netherlands

An exercise in producing flows statistics from big data sources  
 Pilar Rey del Castillo, Vidal Miguel Lázaro Toribio, Eurostat

**Session 13B – CENSUS APPLICATIONS****MANS****Chair:** Michael Neutze, Destatis

Weighting classes versus Dual System Estimation for population estimates using a census or administrative sources  
 Owen Abbott, Helen Ross, ONS

Population size estimation with different imputation techniques for incomplete covariates  
 Susanna Gerritse, Utrecht University; Bart F.M. Bakker, Statistics Netherlands, VU University Amsterdam; Peter G.M. van der Heijden, Utrecht University, University of Southampton

Measuring Uncertainty in ONS population estimates: capturing variability in statistics from combinations of census, administrative and survey sources  
 Katy Stokes, ONS

Creating a new framework for census workplace data  
 David Martin, Samantha Cockings, Andrew Harfoot, University of Southampton; Bruce Mitchell, Ian Coady, ONS

Uncertain population forecasting: A case for practical uses  
 Jakub Bijak, Isabel Alberts, Juha Alho, John Bryant, Thomas Buettner, Jane Falkingham, Jonathan J. Forster, Patrick Gerland, Thomas King, Luca Onorante, Nico Keilman, Anthony O'Hagan, Darragh Owens, Adrian Raftery, Hana Ševčíková, Peter W.F. Smith

Mass appraisal at the Census Level-Israeli Case  
 Larisa Fleishman, Yury Gubman, Central Bureau of Statistics in Israel

**Session 13C – SURVEY INTEGRATION, COORDINATION AND ALIGNMENT****JENK****Chair:** Martin Axelson, Statistics Sweden

A European toolbox for a modular design and pooled analysis of social survey programmes  
 Martin Karlberg, Fernando Reis, Cristina Calizzani, Fabrice Gras, Eurostat

Sampling coordination of business surveys: a new method implemented at INSEE  
 Emmanuel Gros, INSEE

Avoiding duplicate collection of flow data: estimating intra-EU inbound tourism using partner data  
 Christophe Demunter, Krista Dimitrakopoulou, Eurostat

The harmonisation of mirror data using simultaneously estimated accuracies  
 Arie ten Cate

Aligning estimates from different surveys using Empirical Likelihood methods  
 Ewa Kabzinska, Yves G. Berger, University of Southampton

18:00-20:00

**Session 14A – HORIZON 2020 NETWORKING SESSION**

GASP

**Chair:** Martin Karlberg, Eurostat

Receptivity and knowledge transfer in statistical governance issues in international context  
 Marika Pohjola, University of Tampere

Webdatanet & Master in Webdatametrics Web based Data Collection and Analyses  
 Pablo de Pedraza, University of Amsterdam (UVA)

MINTSE-NET (Minimize Total Survey Error Network)  
 Marc Plate, Statistics Austria

Linked Open Statistics Infrastructure  
 Efthimios Tambouris, University of Macedonia and ITI-CERTH

Linking spatial monitors: Geospatial indicators as Linked Open Data  
 Anuja Dangol, SADL (KU Leuven)

Cohesion Policy, Regions, and the Perception of Europe  
 Maria Rosalia Vicente, University of Oviedo

The unit problem in official economic statistics  
 Boris Lorenc, The unit problem in official business statistics

Tailored motivation of business respondents (Improving data collection by soft computing)  
 Miroslav Hudec, University of Economics in Bratislava

Estimation of Behavioral Parameters of CGE Models For the 28 EU Countries  
 Second Bwanakare, University of Information Technology and Management in Rzeszow

**Session 14P – POSTER NETWORKING SESSION**

L2

Informal settlements in Egypt, 2011-The case of Al-Duwika Zone  
 Dalia Galal ElAbady, Central Agency for Public Mobilization and Statistics

Visualisation of macroeconomic indicators in maps with R  
 Jan-Philipp Kolb, Gesis – Leibniz Institute for the Social Sciences

Towards better communication channels  
 Agnieszka Mróz, Central Statistical Office Poland

On estimation of polish real estate market characteristics using Internet data sources  
 Maciej Beręsewicz, Poznań University of Economics, Statistical Office in Poznań

EMIR data from trade repositories as a new source of OTC CDS data  
 Grzegorz Skrzypczyński, Linda Fache-Rousová, Małgorzata Osiewicz, European Central Bank

Data Integration: an Application of a Spatially-Adjusted Regression Tree Model  
 Lisa Borsi, University of Trier; Rebecca C. Steorts, Carnegie Mellon University; Ralf Münnich, University of Trier

Quality, analytic potential and accessibility of linked administrative, survey and publicly available data  
 Manfred Antoni, Alexandra Schmucker, Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

Challenges of linking statistical data and phonetic pronunciation software. Case study: problem of regular statistics establishments' frames in Egypt  
 Nehall Ahmed Farouk Mohamed, CAPMAS, Egypt

Estimation of Economic Performance Indicators for the Agricultural-Food Sector Through Integration of Surveys and Administrative Sources  
 Gismondi Roberto, Alfredo Cirianni, Paolo Righi, Istat

Early material deprivation statistics  
 Sigita Grundiza, Eurostat

Weighting procedure in SILC and its effect on variance estimation  
 Emanuela Di-Falco, Emilio Di Meglio, Eurostat

Use of new technique to measure wellbeing index for North Africa countries (A comparison study)  
 Mahmoud Mohamed Elsarawy, Central Agency for Public Mobilization and Statistics

Estimation of the at-risk-of-poverty rate from interval data  
 Simon Lenau, Ralf Münnich, University of Trier

Optimum allocation of variables in a modular survey architecture  
 Evangelos Ioannidis, Athens University of Economics and Business; Fernando Reis, Cristina Calizzani, Fabrice Gras, Martin Karlberg, Eurostat; Takis Merkouris, Athens University of Economics and Business; Michalis Petrakos, Photis Stavropo, Agilis SA; Li-Chun Zhang, University of Southampton and Statistics Norway

Unit non-response in household wealth surveys: experience from the Eurosystem's Household Finance and Consumption Survey  
 Guillaume Osier, European Central Bank

18:30-20:00

**Networking Cocktail**

L2



# Thursday 12 March 2015

09:00-09:45

## Session 15A – THE IMPACT OF THE DATA REVOLUTION ON OFFICIAL STATISTICS

GASP

**Keynote address by Rob Kitchin, Professor at the National Institute of Regional and Spatial Analysis, National University Ireland Maynooth**

**Chair:** Evelyn Ruppert, Goldsmiths University of London

**Moderator:** Albrecht Wirthmann, Eurostat

09:45-10:15

Coffee break

L2

10:15-11:15

## Special 16A – LEVERAGING THE POTENTIAL OF LINKED (OPEN) DATA IN STATISTICS

GASP

**Session organiser/chair:** Carola Carstens, Directorate-General CONNECT, European Commission

**Moderator:** Bogomil Kovachev, Eurostat

ICT Tools for statistical linked open data: The OpenCube toolkit

Efthimios Tambouris, Evangelos Kalampokis, Konstantinos Tarabanis, University of Macedonia and ITI-CERTH, Thessaloniki, Greece

Official Statistics meets the Semantic Web: how SDMX and RDF can live together

Raffaella Maria Aracri, Stefano De Francisci, Andrea Pagano, Monica Scannapieco, Istat

## Special 16B – ADVANCES IN SMALL AREA ESTIMATION WITH APPLICATIONS – Session PART I

MANS

**Session organiser/chair:** Risto Lehtonen, University of Helsinki

Poverty mapping at a local level with suitable modelling of income

Isabel Molina, Universidad Carlos III de Madrid

Small area estimates of income: means, medians and percentiles

Alison Whitworth, Kieran Martin, ONS; Nikos Tzavidis, University of Southampton; Marie Cruddas, Christine Sexton, Alan Taylor, ONS

Comparing small area estimation methods for poverty indicators in the municipalities of Minas Gerais State

Solange Correa, University of Southampton; Debora Souza, Nicia Brandolin, Viviane Quintaes, Djalma Pessoa, Brazilian Institute of Geography and Statistics

## Session 16C – QUALITY OF ADMINISTRATIVE AND MULTISOURCE DATA

JENK

**Chair:** Jean-Pierre Poncelet, Eurostat

Analysing whether sample survey data can be replaced by administrative data

Arnout van Delden, Reinder Banning, Arjen de Boer and Jeroen Pannekoek, Statistics Netherlands

Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables

Sander Scholtus, Bart F. M. Bakker, Arnout van Delden, Statistics Netherlands

Measuring the quality of multisource statistics

Sorina Văju, Mihaela Agafitei, Fabrice Gras, Wim Kloek, Fernando Reis, Eurostat

An ontology-based approach to administrative data sources' documentation and quality evaluation

Giovanna D'Angiolini, Pierina De Salvo, Andrea Passacantilli, Edoardo Patrino, Teresa Saccoccio, Istat

11:30-12:30

## Session 17A – ASSESSMENT OF BIG DATA

GASP

**Chair:** Fernando Reis, Eurostat

**Moderator:** Christophe Demunter, Eurostat

Profiling big data sources to assess their selectivity

Piet Daas, Joep M.S. Burger, Statistics Netherlands

Analysis of the potential of selected big data repositories as data sources for official statistics

Michalis Petrakos, Anais Santourian, Gregory Farmakis, Photis Stavropoulos, Georgia Oikonomopoulou, Eleni Ntakou, Alexandra Trampeli, Marina Koumaki, Agilis SA

A suggested framework for national statistical offices for assessing the quality of big data

Task Team on Big Data Quality

Proposal for an accreditation procedure for big data sources

Albrecht Wirthmann, Eurostat; Photis Stavropoulos, Michalis Petrakos, George Petrakos, Agilis SA

## Special 17B – ADVANCES IN SMALL AREA ESTIMATION WITH APPLICATIONS – Session PART II

MANS

**Session organiser:** Risto Lehtonen, University of Helsinki

**Chair:** Monica Pratesi, University of Pisa

Estimating regional ARPRs in a border region

Ralf Münnich, University of Trier

Theoretical and practical aspects of mapping poverty in Poland using SAE methods

Marcin Szymkowiak, Statistical Office Poland, Poznań University of Economics

Estimation of poverty rate for small areas by model calibration and "hybrid" calibration methods

Risto Lehtonen, University of Helsinki; Ari Veijanen, Statistics Finland

## Session 17C – MIXED (WEB) MODE INTERVIEWER EFFECTS IN DATA COLLECTION

JENK

**Chair:** Britta Gauckler, Eurostat

Towards an integrated Consumer Expenditure Survey -Combining Multi-mode Data Collection and Big Data Extracts

Gustav Haraldsen, Sverre Amdam, Statistics Norway; Li-Chun Zhang, University of Southampton/Statistics Norway

Integrating the Web Mode in the Austrian Household Budget Survey 2014/15 – First Experiences With a New IT System for Data Collection in a Mixed Mode Design

Marc Plate, Romana Riegler, Statistics Austria

Adapting Labour Force Survey questions from interviewer-administered modes for web self-completion in a mixed-mode design

Peter Betts, Ben Cubbon, ONS

Recommended practices for the design of business surveys questionnaires

Stefania Macchia, Istat

Interviewer Effects in Real and Falsified Interviews. Results from a large scale experiment

Peter Winker, Karl-Wilhelm Kruse, University Giessen; Natalja Menold, Uta Landrock, GESIS

12:30-14:00

Lunch break

L2

12:30-14:00

## Session 18P – POSTER PRESENTATION

L2

User-friendly framework for metadata and microdata documentation based on international standards and PCBS Experience

Haitham Zeidan, Palestinian Central Bureau of Statistics; Geoffrey Greenwell, OECD

Sampling design data file

Seppo Laaksonen, University of Helsinki

Pros and cons of Using CSpro in Designing a Data Entry Applications For the Statistical Survey at CAPMAS

Waleed Mohammed, Central Agency for Public Mobilization and Statistics

Development of a dynamic Survey Platform for Complex Questionnaires: The case of Oman

Jaffar H Mansour, RealSoft

Quality in the web data collection: Standardising online questionnaires, integration with administrative sources and development of bias control mechanisms

Cristina Prado, Patxi Pizarro, Carmen Guinea, Basque Statistics Office

13:15-13:55

## Special 18B – SKILLS FOR TOMORROW'S OFFICIAL STATISTICIANS Session

MANS

**Chair:** Annika Näslund, Eurostat

EMOS-European Master in Official Statistics

Zivile Aleksonyte-Cormier, Markus Zwick, Eurostat

Going beyond GDP: a new challenge also for training statisticians. A new Master proposal and experience

Filomena Maggino, University of Florence; Maria Pia Sorvillo, Istat

A new job for statisticians: the data scientist. Which skills, how to build them

Ludovico Antonio Ottaiano, Istat

14:00-15:00

## Special 19A – COMBINING STATISTICAL AND OTHER SOURCES FOR ENVIRONMENTAL ANALYSIS Session

GASP

**Session organiser/chair:** Jan-Erik Petersen, European Environment Agency

**Discussant:** Steven Vale, Statistical Division, UNECE

**Moderator:** Pilar Rey del Castillo, Eurostat

Geo-spatial data and statistics to support the knowledge base for monitoring natural capital

Jan-Erik Petersen, European Environment Agency; Anton Steurer, Eurostat

Web tools for accessing and disseminating data of different formats

Mauro Michielon, European Environmental Agency

The European bird monitoring programmes as examples of citizen science relevant to policy and research

Petr Voříšek, Czech Society for Ornithology, Czech Republic; Ruud Foppen, Dutch Centre for Field Ornithology, Netherlands;

Richard Gregory, Royal Society for Protection of Birds, UK

## Special 19B – ADVANCES IN SMALL AREA ESTIMATION WITH APPLICATIONS – Session PART III

MANS

**Session organiser:** Risto Lethonen, University of Helsinki

**Chair:** Marcin Szymkowiak, Statistical Office Poland, Poznań University of Economics

Small Area Estimation models with outliers in covariates

Monica Pratesi, Caterina Giusti, Stefano Marchetti, Nicola Salvati, University of Pisa

Accounting for Hyperparameter Uncertainty in Small Area Application Based on a State-Space Model: the Case of the Dutch Labour Force Survey

Oksana Bollineni-Balabay, Jan van den Brakel, Statistics Netherlands, Maastricht University School of Business and Economics; Franz Palm, Maastricht University School of Business and Economics

Reliable estimates in groups with small samples

Agne Bikauskaitė, Dario Buono, Eurostat

## Session 19C – CENTRES OF EXCELLENCE AND ESSNETS

JENK

**Chair:** Amerigo Liotti, Eurostat

The Seasonal Adjustment Center of Competence-Missions and First Achievements

Dominique Ladiray, INSEE

The use of statistical services in the European System of interoperable statistical Business Registers

Susanne Maus, Eurostat

Standardisation in the European Statistical System: inventory of normative documents and the standard-setting process – results of the ESSnet on Standardisation

Csaba Ábry, Zoltán Vereczkei, Hungarian Central Statistical Office

Conclusions of the ESSNet – DCSSon web and mixed mode data collection in official social surveys

Annemieke Luiten, Statistics Netherlands; Karen Blanke, Destatis

Manual for statistics on energy consumption in households

Duncan Millard, Department of Energy and Climate Change; Cristian Fetic, Eurostat

15:00–15:30

## Coffee break

L2

15:30–18:00

## Session 20A – CLOSING SESSION

GASP

**Chair:** Mariana Kotzeva, Eurostat

**Moderators:** Roberto Barcellan & Martin Karlberg, Eurostat

The ever changing landscape of statistics.

Closing address by **Professor Jelke Bethlehem**, Leiden University

Official Statistics in the New Data Ecosystem.

Closing address by **em. Professor David Hand**, Senior Research Investigator and Chief Scientific Advisor, Winton Capital Management

Big Data in the context of Official statistics in Horizon 2020.

Closing address by **Mr. Zoran Stančič**, Deputy Director General, DG for Communications Networks, Content & Technology, European Commission

Closing address by **Mariana Kotzeva**, Deputy Director General, Eurostat

18:00–20:30

## Session 21B – SATELLITE WORKSHOP ON LINKED STATISTICS

MANS

**Workshop organisers:** Efthimios Tambouris, Konstantinos Tarabanis, University of Macedonia and ITI-CERTH; Bill Roberts, Swirrl; Andriy Nikolov, fluidOps

**Workshop chair:** Efthimios Tambouris, University of Macedonia and ITI-CERTH

# Friday 13 March 2015

09:00–13:00

## SATELLITE EVENT ON THE INGRID PROJECT

MANS

**Chairs:** Guy Van Gyes, HIVA-KU Leuven, Belgium, coordinator InGRID; Ralf Münnich, University of Trier, Germany

09:00–16:30

## Tutorial: R IN THE STATISTICAL OFFICE

JENK

**Instructors:** Matthias Templ, Vienna University of Technology, Statistics Austria, Palacký University Olomouc, data-analysis OG; Valentin Todorov, United Nations Industrial Development Organization (UNIDO)

= Room Alcide de Gasperi  
**GASP** - Level 2

= Room Sicco Mansholt  
**MANS** - Level 0

= Room Lord Roy Jenkins  
**JENK** - Level 0

= Lobby on level 2  
**L2** - Level 2

### NTTS 2015 Scientific Committee

Chairman: **Martin Karlberg**, European Commission – Eurostat

**Rein Ahas**, University of Tartu

**Silvia Biffignandi**, University of Bergamo

**Carola Carstens**, European Commission – Directorate-General Communications Networks,  
Content & Technology

**Piet Daas**, Statistics Netherlands

**Patrick Deboosere**, VUB

**Anders Holmberg**, Statistics New Zealand

**Beat Hulliger**, University of Northwestern Switzerland (FHNW)

**Risto Lehtonen**, University of Helsinki

**Ralf Münnich**, University of Trier

**Marianne Paasi**, European Commission – Directorate-General Research and Innovation

**Giuditta de Prato**, European Commission – Institute for Prospective Technological Studies

**Pilar Rey del Castillo**, European Commission – Eurostat

**Evelyn Ruppert**, Goldsmiths University of London

**Fritz Scheuren**, NORC at the University of Chicago

**Natalie Shlomo**, University of Manchester

**Marina Signore**, Istat

**Roxane Silberman**, Réseau Quetelet

### NTTS 2015 – Reliable Evidence for a Society in Transition

New Techniques and Technologies for Statistics (NTTS) is an international biennial scientific conference series, organised by Eurostat, on new techniques and methods for official statistics, and the impact of new technologies on statistical collection, production and dissemination systems.

The purpose of the conference is both to allow the presentation of results from currently ongoing research and innovation projects in official statistics, and to stimulate and facilitate the preparation of new innovative projects (by encouraging the exchange of views and co-operation between researchers – including the possible building of research consortia) with the aim of enhancing the quality and usefulness of official statistics and to prepare activities related to research in statistics within the European Framework Programme for Research and Development (Horizon 2020).



# Big Data: Google Searches Predict Unemployment in Finland

Joonas Tuhkuri ([joonas.tuhkuri@etla.fi](mailto:joonas.tuhkuri@etla.fi))<sup>1</sup>

**Keywords:** big data, Google, Internet, forecasting, unemployment

## 1. INTRODUCTION

There are over 100 billion searches on Google every month. [1] Could data from Google searches help to predict the unemployment rate in Finland?

Predicting the present and the near future is of interest, as the official records of the state of the economy are published with a delay. Furthermore, the data are subject to revisions, sampling variation, and measurement error. It would be helpful to have more timely information on unemployment, especially during an economic crisis. From a policy perspective, more accurate knowledge could inform better decisions that might help to reduce the unemployment.

Data on Google searches are publicly available in real-time. Real-time information might help to *nowcast* the present unemployment rate, which is uncertain. Furthermore, Google search queries might be associated with the future expectations and thus help to *forecast* the future unemployment.

To answer these questions, this paper compares a simple univariate autoregressive model of unemployment to a model that contains a variable, Google Index, formed from Google data. In addition, cross-correlation analysis and Granger non-causality tests are performed. Furthermore, to study the robustness of the results, I explore the sensitivity of the findings to the selected search terms. The Google Index is constructed from the Google data using approximately 2 million [2] search queries that are related to search for unemployment benefits. The underlying idea is that Google searches in these topics might be related to actual filings for unemployment benefits. Moreover, the Internet plays an important role in the labor market [3–5]. That is why Google searches might be able to offer information especially on the unemployment rate. To be clear, I do not claim any clear causality in this paper. However, Google searches might offer a signal on the future unemployment rate. A new data set could also offer new insights on the unemployment.

Previous literature suggests that Google searches could be useful in predicting the unemployment rate in Germany [6], the United Kingdom [7], the United States [8], and predicting the initial claims for unemployment in the United States [9]. This paper offers an extensive review of the literature on forecasting with Internet search data. In summary, the previous studies on Internet searches hint that the variation in volumes of Internet search terms could reveal intentions or sentiment of the population that uses the Internet. However, the topic is still relatively little studied. Previous results serve as a motivation to study further the possibility to use Google searches for predicting the unemployment rate, in this case in Finland. This is the first paper to use Google data to study the Finnish economy.

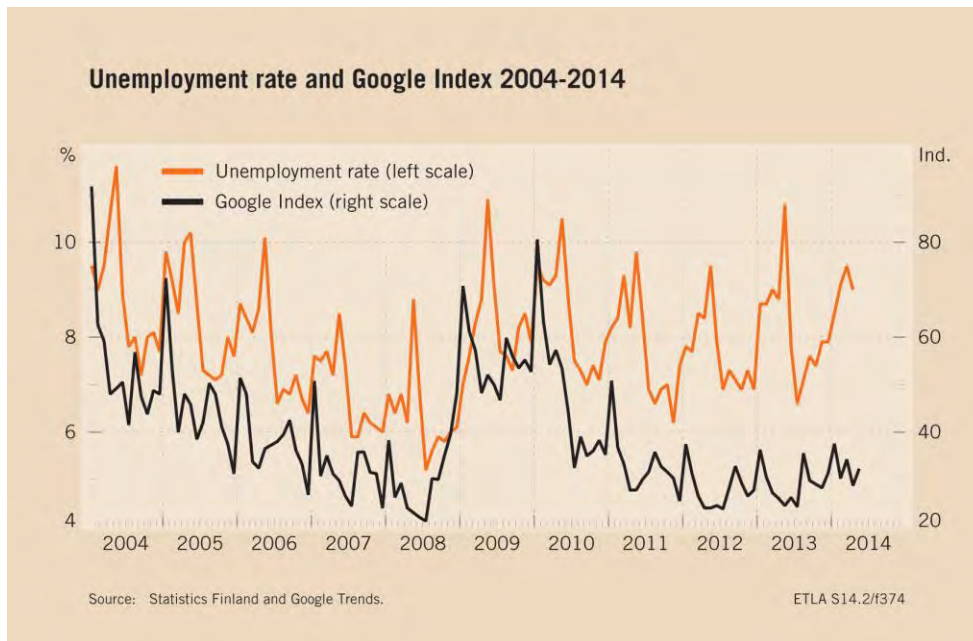
---

<sup>1</sup> The Research Institute of the Finnish Economy and the University of Helsinki.

## 2. METHODS

The primary data sources for this paper are *Google Trends* database by *Google Inc.* and the Labor Force Survey published by the *Statistics Finland*. *Google Trends* measures the incidence of specific search terms on Google compared to other searches terms. To my knowledge, this is the first paper that additionally provides statistics on the actual search volumes on Google. Although there have been efforts to use the data on Google searches in economic research, the Google data has not been well documented. This paper fills this gap by offering a careful documentation and a discussion on the Google data.

This paper uses an extended version of the methods initially suggested by Choi and Varian [9] and Goel et al. [10] to find out whether Google searches predict the unemployment rate in Finland. In summary, I construct a one single Google Index that will simultaneously describe the evolution of several search terms that are related to unemployment, such as “unemployment benefits”. Figure 1 describes the evolution of Google Index and the unemployment rate 2004–2014. The search terms are selected based on prior knowledge of the labor market. Google Index is available in real time, while the unemployment rate is only available after the end of each month. This gives the Google data a meaningful forecasting lead.



**Figure 1. Evolution of Google Index and the unemployment rate in Finland 2004–2014. Source: Statistics Finland and Google Trends.**

I apply standard ARIMA model selection procedures and select seasonal AR(1) model as a benchmark for predicting the unemployment rate. Then I add the most recent value of Google Index to the model as an additional predictor. Finally I compare the properties of the two models. In specific, I study how the out-of-sample forecasts improve, measured by mean absolute error using a rolling window forecast. For each prediction I compare dynamic forecasts that contain the most recent information available at the date of prediction. I study particularly the turning points since they are hardest to forecast. Last I run Granger non-causality tests and study the cross-correlation function.

One concern would be that the results were very sensitive to the choice of the set of search terms. I explore this sensitivity by estimating the models with different search terms.

### 3. RESULTS

The results tell that a simple time series model extended with Google data predicts the unemployment rate better than the same model without data on Google search volumes.

Table 1 summarizes the out-of-sample nowcasting and forecasting accuracy of the benchmark (0.0) and the extended models. Compared to a simple benchmark, Google search queries improve the prediction of the present by 10% measured by mean absolute error. Moreover, predictions using search terms perform 39% better over the benchmark for near future unemployment 3 months ahead. The paper also suggests that Google search queries tend to improve the prediction accuracy around turning points.

**Table 1. Nowcasting and forecasting accuracy of the benchmark and the extended models.**

	Model	MAE	$\Delta$
$t$	(0.0)	7.8 %	10.0 %
	(1.0)	7.0 %	
$t+1$	(0.0)	9.3 %	16.9 %
	(1.1)	7.7 %	
$t+2$	(0.0)	10.5 %	32.9 %
	(1.2)	7.0 %	
$t+3$	(0.0)	11.1 %	39.2 %
	(1.3)	6.7 %	
$t+4$	(0.0)	11.3 %	30.5 %
	(1.4)	7.7 %	
$t+5$	(0.0)	11.3 %	25.3 %
	(1.5)	8.4 %	
$t+6$	(0.0)	11.4 %	20.5 %
	(1.6)	9.0 %	

MAE = mean absolute error  
 $\Delta$  = improvement in forecasting accuracy

The estimation results of the models support the findings. The coefficient of the present Google Index is statistically significant at 1% level and it has a positive sign, which means that the searches related to unemployment benefits are positively connected to the present unemployment rate. More to the point, the coefficient is 0.0056, which means that 1% increase in current search intensity is associated with 0.5% increase in current unemployment rate.

Extending the benchmark model with *Google Trends* data increases the coefficient of determination by 14.8% and decreases the values of both Akaike and Bayesian information criteria. These findings suggest that the Google searches offer useful information in explaining variation of the unemployment rate.

The correlation coefficient between monthly unemployment and Google Index is 0.87. I observe that the cross-correlation coefficients between present unemployment volumes and past Google searches appear to be larger than the coefficient of the opposite case. In other words, the Google Index presents a classic pattern of a leading indicator. According to the Granger non-causality test, Google searches offer useful information in predicting the unemployment rate. In contrast, the lagged values of unemployment rate do not seem to offer useful information in predicting the Google searches.

The results indicate that Google searches might offer genuinely new information on the unemployment that is not already included in the unemployment series itself. Robustness checks suggest that the results are not sensitive to the selected search terms.

#### **4. CONCLUSIONS**

I have found that a simple seasonal first-order autoregressive model, which includes relevant Google variables, tends to outperform models that exclude these predictors in predicting the unemployment rate in the short run. Moreover, joint analysis of the series suggests that the changes in Google searches, which are related to unemployment benefits, more often than not precede the changes the unemployment rate. The results are in line with the previous findings on Google searches and the unemployment [6–9].

As a result, the Research Institute of the Finnish Economy has launched a trial for a real-time forecast model ETLAnow that automatically predicts the unemployment rate for three months ahead using data from Google Trends and Statistics Finland, publishing the results every morning. Currently we are building a model that would predict the unemployment rate for each EU-28 country in real-time using big data.

The results suggest that Google searches could offer useful information for predicting the Finnish unemployment rate. The results also demonstrate that big data can be utilized to forecast official statistics.

#### **REFERENCES**

- [1] Google Internal Data, (2014).
- [2] Google Adwords, (2014).
- [3] B. Stevenson, The Internet and Job Search, NBER Working Paper 13886 (2008).
- [4] P. Kuhn and H. Mansour, Is Internet Job Search Still Ineffective? The Economic Journal 124(581) (2014), 1213–1233.
- [5] K. Kroft and D. G. Pope, Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist, Journal of Labor Economics 32(2) (2014), 259–303.
- [6] N. Askitas and K. F. Zimmermann, Google Econometrics and Unemployment Forecasting, Applied Economics Quarterly 55(2) (2009), 107–120.
- [7] N. McLaren and R. Shanbhogue, Using internet search data as economic indicators, Bank of England Quarterly Bulletin 1 (2011) 134–140.
- [8] F. D’Amuri and J. Marcucci, The Predictive Power of Google Searches in Forecasting Unemployment, Bank of Italy Working Paper 891, (2012).
- [9] H. Choi and H. R. Varian, Predicting the Present with Google Trends, Economic Record 88(s1) (2012), 2–9.
- [10] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, Predicting consumer behavior with Web search, Proceedings of the National Academy of Sciences of the United States of America 107(41) (2010), 17486–90.



# Collecting, storing and managing fuzzy data in statistical relational databases

Miroslav Hudec (miroslav.hudec@euba.sk)<sup>1</sup>, Dušan Praženka<sup>1</sup>

**Keywords:** fuzzy data, fuzzy information, relational database

## 1. INTRODUCTION

The values of attributes are not always known with sufficient precision to justify the use of traditional relational databases to store these data [1]. Many data are fuzzy either by their nature [2] or caused from the non-ideal measuring [3]. The fuzziness is amplified when both types appear. Therefore, we should not neglect these facts. The same holds for collecting information and knowledge. Knowledge collected by experts often contains elements of uncertainty [4]. This information should be considered in knowledge management systems and recommender systems.

Our main objective is constructing environment for efficiently storing and re-using fuzzy data. Full fuzzy databases [5] are sophisticated, but we see the lack of practical applications and tools. On the other hand, relational database management systems are well developed and broadly used. Hence, adding fuzziness into statistical relational databases is a promising way. In our work we have started with analyzing approach initially suggested in [6] and continued with creation of adjustments and improvements in order to be useful for official statistics. Keeping integrities of the relational model, we could be able to store, update, disseminate and delete data by SQL like queries [7].

## 2. FUZZY DATA

The fuzzy set theory [8] provides a robust framework for systematically handling uncertainty based on fuzziness. In the fuzzy set theory belonging to a set is a matter of degree. A fuzzy set  $A$  over the universe of discourse  $X$  is defined by function  $\mu_A(x)$  that matches each element of the universe of discourse  $X$  with its membership degree to the set  $A$  in the following way:

$$\mu_A(x): X \rightarrow [0,1]$$

Concepts like *medium value* or value close to  $a$ , where  $a$  is a real number are expressed by triangular or trapezoidal fuzzy sets (Figure 1a and Figure 1b respectively). *High value* is expressed by linear gamma fuzzy set (Figure 1c) whereas *small value* is explained by L fuzzy set (Figure 1d). Finally, we should consider single value of fuzzy data as singleton fuzzy set (Figure 1e).

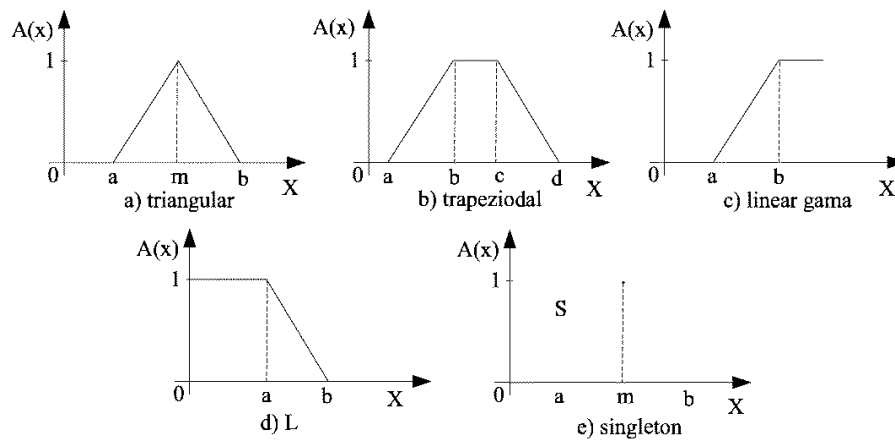
## 3. FUZZINESS IN THE REAL WORLD

Many data are fuzzy either by their nature, caused from the tolerance level of instruments for measuring and as a result of respondents' estimation. For example "environmental data, quality of life data and measurements of continuous one-dimensional quantities cannot be adequately expressed by crisp (sharp) numbers." [2] A good example of the first type is the flooded level marked on a wall illustrated in Figure 2. Where exactly to

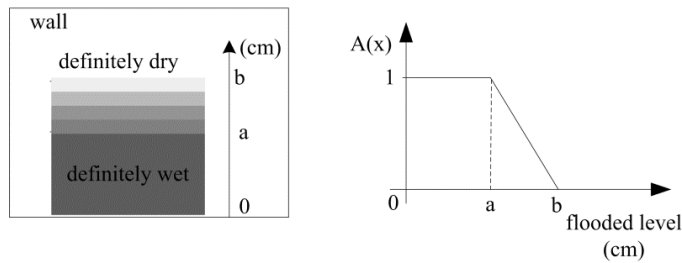
---

<sup>1</sup> University of Economics in Bratislava, Faculty of Economic Informatics, Bratislava, Slovakia

measure this value in order to ensure that similar entities will be similarly treated? By fuzzy data (right side in Figure 2) we could avoid this problem.



**Figure 1. Fuzzy sets**



**Figure 2. Level of flood as a fuzzy set**

We should keep in mind that the measurement made by a measuring instrument is usually approximate due to the tolerance interval. It means that the precise value is somewhere in the (small) interval  $[a, b]$ , i.e.  $\mu(x) = 1$  for  $x \in [a, b]$ .

People measure (guess) values often by estimation. For example someone could declare that speed was approximately 100 km/h but for sure not lower than 90 km/h and not higher than 110 km/h. This uncertainty could be managed by triangular fuzzy set (Figure 1a). Respondents often estimate answers even for open ended questions. Furthermore, it could be useful to allow skilled interviewers to remark relevance of answers or the credibility of surveyed person.

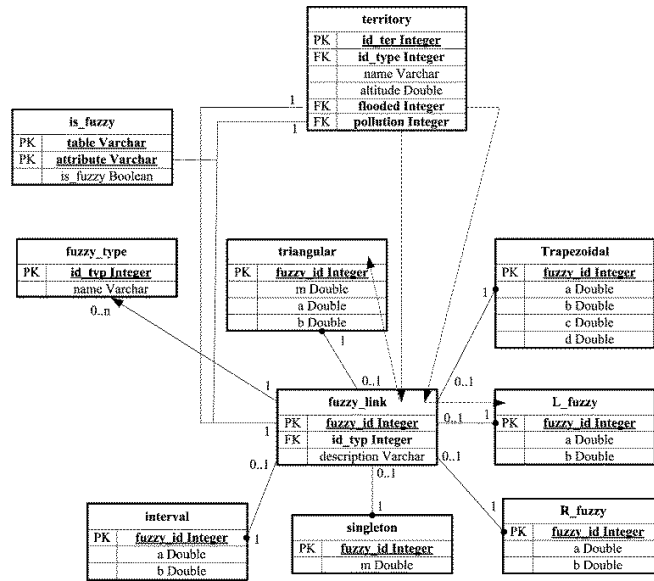
Another example is expressing staff knowledge about solving tasks and how to reveal appropriate solutions for new tasks. Similarity of task and collected information is under consideration in the fuzzy recommender systems [9].

The following question arises: Why we should manage fuzzy data? We could just select one crisp value as a representative of measured value. In could cause some unexpected results and the differently treatment of similar entities [1], [5], [10].

#### 4. FUZZY DATA IN RELATIONAL DATABASES

Fuzzy meta model keeps all relevant fuzzy data and manages links to relational tables of real entities [1], [6]. The logical model of a municipal statistics database is shown in Figure 3. Each fuzzy value or linguistic term is represented by membership function. For example, the table trapezoidal has the following attributes (fuzzy\_id, a, b, c, d) in order to manage storing trapezoidal fuzzy data (Figure 1b). In the same way we construct other

relational tables for storing other types of fuzzy data. In the table territory columns flooded and pollution do not store real values but foreign keys to the respective tables where real values are stored as fuzzy. In this way relational database is capable to store fuzzy data and meet integrity rules.



**Fig.3. Fuzzy data of territorial units in relational database [1]**

The size of stored data is affected by number of fuzzy attributes (columns) and types of fuzzy sets. In the extreme situation all  $n$  attributes (excluding primary key) for all  $m$  entities are fuzzy of trapezoidal form. Therefore, the size of database is  $9(m \cdot n)$  [1]. Anyway, user should carefully decide which attributes should be stored as fuzzy and in which form. The validation rules warn users if they want to input values in inappropriate fuzzy set. Keeping in mind this deduction, we could say that the size of fuzzy relational database could be significantly lower than this extreme situation.

Possibility and necessity functions [5] allow us to e.g. calculate the possibility that fuzzy data *pollution about 20g* belongs to the concept *small pollution*. It is useful in querying, disseminating and data processing. In mining summarizing information we could keep in mind disclosure control [10].

Moreover, relational databases could be straightforwardly extended to manage fuzzy data. In case of fuzzifying existing attributes having already collected values, they should be migrated to fuzzy data (of singleton type). New values could be collected as fuzzy.

We are now considering storing and managing fuzzy information expressed by linguistic terms for knowledge management. Merging this approach and approach discussed in [11] might be promising but definitely, further research is required.

## 5. FUZZY DATA IN STANDARDS FOR DATA EXCHANGE

Another interesting question is exchange of uncertain data and information among statistical organizations if needed. Standards like e.g. SDMX and DDI [12] could be utilized for this purpose. If we drew analogy between model explained in Figure 3 and the above standards then it implies that we could add tags on two levels. On the attributes level we should add tag which informs users that this attribute contains fuzzy information. For each observation we should add tags explaining type of fuzzy information (fuzzy set) and parameters of fuzzy set. Currently, these tags does not exist but they construction could be possible if institutions decide to use fuzzy data.

## 6. CONCLUSION

It should not be neglected that many real data of statistical interest are fuzzy. Crisp values cannot be always ideally measured or estimated without loss of relevant information. Therefore, we need an efficient way for storing them and re-using in variety of analyses. Account on that, our work was focused on managing fuzzy data in relational databases.

We have solved the task by additional tables (fuzzy meta model) and have created validation rules. Each fuzzy data is represented by parameters of respective membership function. Therefore, mapping data into standards for data and metadata exchange could be promising. Number of fuzzy attributes and types of fuzzy sets affects size of the database and search time. Therefore, users should reasonably decide which attributes should be managed as fuzzy.

## REFERENCES

- [1] M. Hudec, Fuzzy Data in Traditional Relational Databases, IEEE conference of Neurel, Belgrade (2014), 195-199.
- [2] R. Viertl, Fuzzy data and information systems, WSEAS International Conference on Systems, Corfu (2011), 83-85.
- [3] F. Pavese, Why should correction values be better known than the measured true value? Journal of Physics: Conference series 459, (2013).
- [4] L. Portinale and A. Verrua, Exploiting Fuzzy-SQL in Case-Based Reasoning, Fourteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS) (2001)
- [5] J. Galindo (ed.), Handbook of Research on Fuzzy Information Processing in Databases, Hershey: Information science reference (2008).
- [6] S. Škrbić, Using fuzzy logic in relational databases, Ph.D. dissertation, University of Novi Sad (2008)
- [7] M. Hudec, What could fuzzy logic bring to statistical information systems? Statistika, 48(1) (2011), 58-70.
- [8] L.A. Zadeh, Fuzzy sets, Information and Control, 8(3) (1965), 338-353.
- [9] E. Herrera-Viedma, C. Porcel, A. López-Herrera, S. Alonso, A Fuzzy Linguistic Recommender System to Advice Research Resources in University Digital Libraries, Studies in Fuzziness and Soft Computing Volume 220, Springer, (2008), 567-585.
- [10] M. Hudec, Fuzzy database queries in official statistics: Perspective of using linguistic terms in query conditions, Statistical Journal of the IAOS, 29(4) (2013), 315-323.
- [11] M. Vučetić and M. Vujošević, A literature overview of functional dependencies in fuzzy relational database models, Techniques Technologies Education Management, 7(4), (2012), 1593-1604.
- [12] D. Praženka and P. Boško, Combining technical standards for statistical business processes from end-to-end, New Techniques and Technologies for Statistics (2011).



# Self-employment - a complementary solution for full use of EU labour potential

Amalia Cristescu (cristescuamalia@gmail.com)<sup>1</sup>, Dorel Ailenei (dorel\_ailenei@yahoo.com)<sup>2</sup>

**Keywords:** labour market, self-employment, logistic regression

## 1. INTRODUCTION

The issue of the full use of the labour has been widely debated in economics. From *The General Theory of Employment, Interest and Money* ([1]) to the *Human Capital Theory* ([2]) and to the modern theories of economic growth ([3], [4]) there has been built a quasi-consensus regarding the importance of labour force. Following this dominant trend in the literature, the economic policies made a priority of the full exploitation of human resources<sup>3</sup>. However, rarely labour markets in developed countries were close enough to the goal of full employment. Beyond the multiple nuances and definitions of the "full employment" objective, the literature raises a multitude of reasons: wage rigidities, labour market failures, macroeconomic structural imbalances, conflicting objectives of the economic policy mix, poor labour market regulations and/or excess/deficit of regulations etc. And yet, there is still one fundamental question that cannot be avoided: why do some countries manage to record significantly higher employment rates than others?

The authors analyze a set of data taken from Eurostat on the extent and characteristics of the self-employment phenomenon in the EU, comparing the performance of the Member States in this field. In order to validate the results of this analysis there was designed a logistic regression model based on data from a Eurobarometer survey. Thus, the key factors that support the development of the self-employment phenomenon were identified, analyzing a few ways to stimulate them. Special attention was paid to education, where the EU also records a poor performance compared to the USA and Japan. The analyses were customized for Romania, as a new member of the EU, rigorously assessing its chances to achieve the objectives of the Europe 2020 Strategy in the field of employment.

## 2. METHODS

The starting point in our analysis is given by the employment gap between the EU and the U.S.A. and Japan. In Europe 2020 Strategy, the E.U. recognizes that "only two-thirds of our working-age population is currently employed, compared to over 70% in the US and Japan." To cover this deficit, the EU sets an ambitious target for employment of 75% by 2020. At the moment, there are significant differences between the Member States regarding the actual employment rate and the target proposed for 2020. Only three countries (Denmark, the Netherlands and Sweden) took a target for employment rate of 80%, as they reach this performance at present, too (79.8% - Sweden, 76.5% - the Netherlands and 75.6% - Denmark). The good part of this picture is that in 2020, 22 of the 28 EU Member States will exceed the current level of employment rate in the USA

---

<sup>1</sup> Bucharest University of Economic Studies

National Scientific Research Institute for Labour and Social Protection

<sup>2</sup> Bucharest University of Economic Studies

<sup>3</sup> Europe 2020 Strategy, „Europe needs to make full use of its labour potential to face the challenges of an ageing population and rising global competition.”

and Japan. Considering that self-employment could be an alternative to improve the employment rate, the authors analyzed the factors that stimulate this phenomenon.

## 2.1 Testing incentive factors of self-employment

The analysis is based on data obtained from a survey – the Euro-barometer "Quality of Life in European Cities" - conducted during November-December 2012. The survey was conducted on a sample of 22,683 people in 83 cities in the Member States of the European Union and Croatia, Turkey, Iceland, Norway and Switzerland. In Romania people were interviewed in Bucharest, Cluj-Napoca and Piatra Neamț. The following characteristics were selected from the Euro-barometer database: *age, gender, residence, level of education and employment status*. The econometric analysis applied to the sample data was to identify the impact that the variables selected have on the self-employment option. The professional status was determined based on the variable that refers to the respondent's occupation – he/she was asked about his/her occupation at the time of the interview. The original variable was post-coded, yielding only three categories: employed, self-employed and unemployed. The unemployed respondents were eliminated, and the econometric estimation was carried out on the employed sample in terms of labour market. Thus, a new variable was designed – a binary variable that takes value 1 if the person is self-employed and 0 if the person is employed. Since the dependent variable is a binary variable, we used a logistic regression model. Basically, the logistic regression model describes a non-linear relationship between the binary variable Y and k explanatory variables  $X_1, X_2, \dots, X_k$ .

The logit model is described by the relationship:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^k \beta_j x_{i,j})}}$$

If a logistic transformation is applied to the equation above the result is the linear relationship between  $\text{logit}(p_i)$  and the explanatory variables:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^k \beta_j x_{i,j}$$

This last equation is called the logit form of the model where  $\text{logit}(p_i)$  is the logarithm of the likelihood of an event to happen with respect to the explanatory variables.

In our analysis we considered the binary variable as a dependent variable corresponding to the status on the labour market (SE), which takes value 1 if the person is self-employed and 0 if the person is employed. The chosen explanatory variables are: *age group, area of residence, gender and the last school graduated*. All variables are binary.

## 3. RESULTS

The binary variables used with the logistic regression were coded as follows:

**se** = self – employed (dependent variable)

**male** = masculine gender, compared to the feminine gender

**age\_15\_24, age\_35\_44, age\_45\_54, age\_55\_6, over\_age\_65** = age groups, compared to the 25-34 age group

**urban** = urban area, compared to rural area

**iscd 1\_2, iscd 5\_6** = education levels, compared to the isced 3\_4 education level.

**still** = still in school

**Table 1. The results of the logistic regression analysis**

Logistic regression	Number of obs	=	22683
	LR chi2(10)	=	814.26
	Prob > chi2	=	0.0000
Log likelihood = -10528.812	Pseudo R2	=	0.0372

	se	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
male		.656179	.035576	18.44	0.000	.5864513	.7259068
age_15_24		-.2123424	.09055	-2.35	0.019	-.3898171	-.0348676
age_35_44		.3272443	.0516622	6.33	0.000	.2259884	.4285003
age_45_54		.5092524	.0518912	9.81	0.000	.4075476	.6109573
age_55_64		.6469205	.0587183	11.02	0.000	.5318348	.7620062
over_age_65		1.599684	.0942749	16.97	0.000	1.414909	1.78446
urban		.1504594	.0744938	2.02	0.043	.0044543	.2964645
iscd 1-2		-.0083872	.0869229	-0.10	0.923	-.178753	.1619785
iscd 5-6		.3065565	.0393846	7.78	0.000	.2293641	.383749
still		.2532513	.1218862	2.08	0.038	.0143588	.4921437
cons		-2.537502	.0882487	-28.75	0.000	-2.710466	-2.364537

The logistic regression analysis (table 1) highlighted the following aspects:

- In terms of gender, men are about 65.6% more involved in self-employment than women. The result is natural since women are more involved in household chores, raising children or part-time jobs that do not affect the household time.
- In terms of age group it can be observed that the phenomenon of self-employment is increasing with age. Compared with those in the 25-34 age group those in the 15-24 age group have less chances of being self-employed because they are less present on the labour market since they are focused on graduating and they do not have accumulated enough knowledge to become self-employed. The age groups 45-54 and 55-64 have more than 50% chances of being self-employed, and this could signal an adaptation to the employers' behaviour who tend to substitute the expensive labour force (seniors) with the cheaper one (juniors). This type of behaviour has increased during the economic crisis and it is perfectly rational considering the optimization paradigm used by the producer (substitution effect of the production factors). At the age group of over 65 the chances are even higher because of the need to supplement income, which is a reaction to the undesirable tendency to extend the active age (retirement).
- In terms of the area of residence, those in urban areas are more likely to be self-employed, due to the fact that urban areas offer more opportunities for self-employment development.
- Depending on the level of education, it is clearly that those with higher education (ISCED 5-6) are more likely to be self-employed, indicating that self-employment activities are more often than not intensive in human capital.

#### 4. CONCLUSIONS

The issue of employment is a key strategic goal of the E.U. However, the Europe 2020 objective to reach an employment rate of 75% is rather challenging, since very few countries have assumed it. Based on the analysis of the Eurostat data on employment the authors concluded that the starting level in the new period for the E.U. (2013-2020) is very different in terms of employment rate, since only 9 of the 28 member states exceed

the level of employment of 70%. As an alternative to increase the employment rate, self-employment varies greatly by age, gender, field of activity and level of education. Thus, the complex nature of self-employment causes considerable challenges in developing economic policy measures and requires a clear understanding of the mechanisms behind self-employment. This knowledge is essential to allow an accurate definition of a coherent set of measures in order to induce and support the transition to self-employment while ensuring an efficient allocation of public resources.

The analysis based on logistic regression performed on the sample of the Euro-barometer "Quality of Life in European Cities" shows that the most important factors in the trend towards self-employment are: age (experience), gender (male), education level (ISCED 5 -6) and area of residence (urban).

The countries are different with respect to the phenomenon of self-employment. We make an analysis based on data obtained from a survey – the Euro-barometer "Quality of Life in European Cities" and these types of data not allow us to analyze each country. We will intent to extend our analyze using other type of data from Eurostat database, to capture the differences for each country.

## REFERENCES

- [1] Keynes, J. M., *The General Theory of Employment, Interest and Money*. (2008), Chapter 19: BN Publishing.
- [2] Becker, G. S., *Human Capital. A Theoretical and Empirical Analysis, with Special Referees to Educationa*.(third editon), (1993), The University of Chicago Press, Chicago 60637.
- [3] Romer, D., *Advanced Macroeconomics*, (2011), McGraw-Hill
- [4] Seyfried, W., Examining the Relationship between Employment and Economic Growth in the Ten Largest States, *Southwestern Economic Review*, (2011), pp. 13-24

# Non-Extensive Entropy Econometrics and CES production Models: Some EU Country Case Study

Second Bwanakare, University of information Technology and Management of Rzeszow,  
[sbwanakare@wsiz.rzeszow.pl](mailto:sbwanakare@wsiz.rzeszow.pl)

**Keywords:** inverse problem; q-generalization of K-L information divergence; constant elasticity of substitution functions; econometrics.

## 1. INTRODUCTION

This paper applies the non-extensive entropy econometrics approach [1] for estimating parameters of constant elasticity of substitution (CES) production models of Germany (GER), France (FR), and the United Kingdom (UK). Conceptually, CES models belong to the class of stochastic non-linear inverse problems displaying power law (PL) properties. The applied technique is based upon the q-generalized Kullback-Leibler (K-L) information divergence under constraining information related to the Bayesian information processing optimal rule. This work considers PL-related non-extensive entropy to remain valuable even in the case of low frequency series since the outputs provided by Gaussian law correspond to the limiting case of Tsallis entropy when the Tsallis q-parameter equals unity. Since a number of complex phenomena exist involving long-range correlations, still observable when data is time (or space) scale-aggregated—this may be the case for CES functions—this represents another argument in favour of applying Tsallis non-extensive entropy formalism. The approach leads to robust estimates for these nonlinear—analytically intractable—functions in each of the considered countries. Outputs from some traditional econometric techniques like the nonlinear least squares(NLLS) or the generalized method of moments(GMM) approaches are also presented.

## 2. METHODOLOGY

### 2.1. Relation between PL and a CES production function

As shown in [1], the connection between a two factor CES production model[2] and PL is direct. Explaining the gross domestic product ( $VA_t$ ) using two classical factors: labor ( $L_t$ ) and capital ( $K_t$ ) and aggregating components of a classical CES model into one variable without conserving additivity, one gets a generic case of a PL of the form:

$$va_t = \alpha \left[ \lambda k_t^{-p} + 1 \right]^{-\frac{v}{p}} e^{\varepsilon_t} \equiv \beta k_t^{\eta} e^{\varepsilon_t} \quad (1)$$

where, in this case, the endogenous variable  $va_t$  is the product per capita. Parameter  $\beta$  represents a general level of technology. The variable  $k_t$  stands for a capital coefficient. The exponent  $\eta$  belongs within the interval  $(-1, +\infty)$  and defines a per capita product elasticity

with respect to the capital coefficient. The random term  $\varepsilon_t$ , itself, is assumed to follow PL structure. Index  $t$  means time period. As far as relationships between a PL and non-extensive Tsallis entropy is concerned, see e.g. [3]. The proposed model generalizes the statistical theory of information [4] approach to non-ergodic systems, i.e. those with dynamically correlated micro-states, thereby suggesting the Tsallis  $q$ -parameter different to unity.

## 2.2. The model estimation

We directly present below the model under a reparameterized form. In Tsallis statistics, there exist a few similar forms of constraint distribution in moments. If we use the one proposed by Curado-Tsallis[5], the non-extensive cross-entropy (NCEE) econometric model can be stated as:

$$\text{Min}H_q(a//a^0, p//p^0, r//r^0) \equiv \alpha \left[ \sum a_{jm} \frac{[a_{jm}/ao_{jm}]^{q-1} - 1}{q-1} + \dots + \sum p_{km} \frac{[p_{km}/po_{km}]^{q-1} - 1}{q-1} \right] + \beta \sum r_{nj} \frac{[r_{nj}/ro_{nj}]^{q-1} - 1}{q-1} \quad (2)$$

s.t.:

$$\ln VA = \ln \left( \sum_{j=1}^J g_j a_j^q \right) - \frac{\sum_{h=1}^H v_h w_h^q}{\left( \sum_{m=1}^M v_m p_m^q \right)} \ln \left[ \sum_{i=1}^I (t_i b_i^q) L^{\left( \sum_{m=1}^M z_m p_m^q \right)} + \left( 1 - \sum_{i=1}^I (t_i b_i^q) K^{\left( \sum_{m=1}^M z_m p_m^q \right)} \right) \right] + \sum_{n=1}^N \sum_{j=1}^J z_{nj}^q \quad (3)$$

$$\sum_{j>2\dots M} a_j = 1 \quad \sum_{m>2\dots M} p_m = 1 \quad (4)$$

$$\sum_{i>2\dots I} b_i = 1 \quad (5)$$

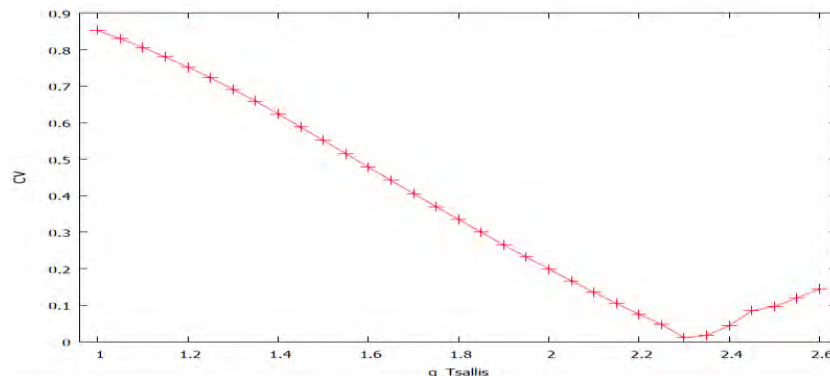
$$\sum_{h>2\dots I} w_h = 1 \quad (6)$$

For reasons of formal presentation, the criterion function (Eq. 2) does not include probabilities  $w_h$ , explaining degree of economy changing to scale, and  $b_i$ , the parameter of distribution between factors. In order to improve the estimated parameter quality through a convex solution space and depending on existing economic theory, additional a priori information should be added to (2)-(6).  $G, Z, T$  support spaces are included in the general support space for all the parameters of the constraining equation system. Here one presents how has been specified this particular model and not a general rule of specification. Depending on error distribution, the weights  $\alpha, \beta$  introduced in the above dual objective function may exercise a significant impact on the model's optimal outputs through the Lagrange multipliers. Next, the parameter confidence area related to NCEE is a normalized entropy index with the numerator representing the present entropy of the system and the denominator displaying its maximum entropy [1]. Finally, as explained in [4], estimated statistics fulfill the basic Fisher-Rao-Cramer information index properties, including continuity, symmetry, maximum, and additivity. In fact, the q-generalization of K-L information divergence (Eq. 2) keeps the basic estimator properties of the K-L information divergence model [6] unchanged.

## 3. RESULTS

The observed original statistical data comes from Eurostat. They have been deflated using value added deflator over a nineteen-year period for three selected countries. We use a GAMS (General Algebraic Modelling System) code to compute the model. Table 1 presents outputs from NCEE. The index of error coefficient (EC) presented in abovementioned table varies from one to zero, in the reverse of the traditional coefficient of determination. For clarity, it is worth mentioning that the estimator  $b_1$  is the parameter distribution related to the labor variable. Surprisingly enough, in all three cases, the optima of the models are reached for the Tsallis q-parameter around 7/3. In a recent work [1], as shown below in Fig. 1, the same value

was found while estimating the CES production model for the aggregated economies of the 27 EU countries. Likewise, Authors [7] using different time-space scaled data have suggested the same optimal value of Tsallis q-parameter.



**Fig.1. Model error variation as a function of Tsallis q-parameter, for  $[1 < q < 2.6]$**

As shown in Table 2 and Table 3, outputs from classical econometric techniques(NLLS, GMM) display lower parameter precision than the NCEE approach. Furthermore, error distribution from the NCEE, not presented in this abstract, suggests efficient and unbiased estimators.

Table 1. Model outputs from NCEE approach						
Estimators	GERMANY		FRANCE		UNITED KINGDOM	
	Values	EC	Values	EC	values	EC
$b_0$ (constant)	2.231		2.439		2.248	
$b_1$ (parameter distribution)	0.475	0.003	0.550	0.004	0.485	0.009
$\nu$ (scale parameter)	1.00		1.00		1.00	
$\sigma$	0.005		0.005		0.005	

Tsallis q-parameter value is around 7/3 for all three countries.

Table 2. Model outputs from the GMM approach						
Estimators	GERMANY		FRANCE		UNITED KINGDOM	
	Values	p-values	Values	p-values	Values	p-values
$b_0$	3,369	<0,00001	4,126	<0,00001	2,446	0,47
$b_1$	0,537	<0,00001	0,6642	<0,00001	1,826	0,808
$\nu$	0,947	<0,00001	0,89	<0,00001	1,059	0,048
$\sigma$	109,008	0,59935	-0,002	0,975	-8,438	0,982

Table 3. Model outputs from the NLS approach						
Estimators	GERMANY		FRANCE		UNITED KINGDOM	
	Values	T-stud	Values	T-stud	Values	T-stud
$b_0$	2,023	4,201	5,222	17,862	1,71	1,521
$b_1$	0,463	31,182	0,685	2,588	0,119	4,182
$\nu$	1,007	-	0,943	-	1,018	-



$\sigma$	-1,123	-	-26,749	-	-19,347	-
R <sup>2</sup>		0,999		0,999		0,999

#### 4. CONCLUSIONS

The work briefly presents additional applications of NCEE – for the econometric modelling of instable, nonlinear models. The case study is based on the CES production models of the three selected EU countries. Outputs from some traditional competitive approaches have been presented too. One of the points to note is that only outputs produced by Tsallis formalism reflect higher similitudes among corresponding parameters of the three selected countries, characterized by relatively comparable economic attributes. Next, following comment of section Three above, the new insight of the research is that the Tsallis q-parameter steadily evolving over a convex space towards global minimum point of the model errors corresponds to q-value around 2.3. Therefore, this Tsallis q-parameter value could stand for a free scale global transition phase-point for some class of PL-driven complex systems. While potential applications of the new approach for nonlinear modelling promise to be important, further theoretical and empirical investigations are needed to better understand its properties and application scope.

#### REFERENCES

- [1] S. Bwanakare, Non-Extensive Entropy Econometrics: New Statistical Features of Constant Elasticity of Substitution-Related Models. *Entropy* 2014, 16, 2713-2728.
- [2] K.J. Arrow; H.B. Chenery; B.S. Minhas; R.M. Solow, Capital-labor substitution and economic efficiency. *Rev. Econ. Stat* 1961, 43, 225–250. - See more at: <http://www.mdpi.com/1099-4300/16/5/2713/htm#sthash.AeSMpV1M.dpuf>.
- [3] X. Gabaix, Power laws in economics and finance, September 2008, <http://www.nber.org/papers/w14299>, ber.
- [4] C. Tsallis, *Introduction to Non-extensive Statistical Mechanics: Approaching a Complex World*, Springer, Berlin, 2009.
- [5] R.S. Mendes; A. R. Plastino; C. Tsallis, The role of constraints within generalized non-extensive statistics, *Physica A: Statistical Mechanics and its Applications*, North-Holland, 1998/12/15.
- [6] S. Kullback; R. A. Leibler, On information and sufficiency. *Annals of Mathematical Statistics*, 1951, 22:79-86.
- [7] T. Oikonomou; A. Provata; U. Tirnakli, Nonextensive statistical approach to non-coding human DNA. *Physica A* 2008, 387, 2653–2659. - See more at: <http://www.mdpi.com/1099-4300/16/5/2713/htm#b25-entropy-16-02713>.

# The evolution of the disparities among the EU member states regarding FDI – the case of the former communist countries

Vasile Alecsandru Strat ([strat\\_vasile@yahoo.com](mailto:strat_vasile@yahoo.com))<sup>1</sup>

**Keywords:** Foreign direct investments, European Union, disparities analysis, concentration analysis.

## 1. INTRODUCTION

The FDI are regarded as a very important source that can fuel the growth of an economy, especially by the governments of the developing countries. Moreover, the benefits brought by FDI are very diverse and they have a significant impact in many fields. They are regarded as a very important source for new and innovative management skills, an important source for new jobs and sometimes for higher salaries.

The former communist countries from Eastern Europe (the eleven countries which are now members of the European Union: Croatia, Bulgaria, Romania, Hungary, Poland, Slovenia, Slovakia, Czech Republic, Estonia, Latvia and Lithuania) are no exception in this regard and the FDI were and still are considered by their governments as a very important source of capital. The level and the typology of the FDI attracted by a country in a period could be interpreted as a very accurate indicator of the development level of that particular economy, and sometimes of the entire socio-economic environment of that country. Following this direction, we will try to analyze and, afterwards, provide a clear description of the evolution of the attractiveness of the countries mentioned above, in this regard.

Taking into consideration all the aspects presented above, we state clear that the main goal of this research paper is to assess if convergence is being achieved in this domain related to foreign direct investments in this part of Europe (Eastern part of the European Union). Consequently, we will try to connect the main events which took place during the last two decades with the evolution of the analyzed phenomena in order to identify a pattern.

## 2. LITERATURE REVIEW AND GENERAL ASPECTS

Due to the fact that the FDI are a very important economical phenomena in the nowadays reality, they receive (and they have received) a great attention in the scholarly literature. Most of the studies dealing with FDI are concerned with the main determinants that attract these types of investments in a host country. A large variety of studies identify market size as a very important determinant of the FDI [1], [2], [3]. Broadman and Sun, in a paper published in 1997, provide evidence that infrastructure is also an important determinant of the foreign direct investments [4]. Other studies identify as important determinants: labour market [5], research and development [6], trade openness [7], [1], macroeconomic stability, corruption level and taxation policies. When dealing with the disparities related with different aspects that are identified as determinants of the foreign direct investments, most of the studies conduct the analysis at regional level [8].

---

<sup>1</sup> Affiliation - Bucharest University of Economic Studies, Department of Statistics and Econometrics

Even though there are many studies which try to rank the attractiveness and the potential of countries in attracting FDI, the rankings provided by UNCTAD have the biggest visibility. Also noteworthy are the studies which show that the benefits brought by FDI in a host country depend heavily on the type of investment and on the characteristics of that economy. Following this direction, is important to mention that there is a new current followed by scholars, who suggest that FDI are not improving the convergence between economies, but they are rather increasing the disparities.

### 3. METHODS

The magnitude of the disparities registered between these countries (at the level of the entire area), when discussing their attractiveness in the eyes of foreign investors will be evaluated with the help of the Gini coefficient.

#### 3.1. Data issues

The time series used in this research paper were gathered from multiple sources, and therefore, the results need to be regarded with caution and they need to be interpreted in the framework imposed by this limitation.

The time series for population (expressed in number of inhabitants) were downloaded from the web site of the EUROSTAT, the GDP (expressed in US \$ at constant prices 2005 and constant exchange rates) time series and the time series for FDI stocks (expressed as % from GDP) were downloaded from the web site of UNCTAD. Due to comparability reasons, (between economies of different size) along this research paper we have used indicators expressed as % from GDP or as value per capita.

#### 3.2. Methodology

In order to analyze the evolution of the indicators related with the FDI, (dynamics) indices were used. They were also used for assessing the convergence level for each economy. The evolution of the inequalities was analyzed using a time series of Gini coefficients. The analysis was performed at the level of the entire area and also at the level of three sub areas. These three sub areas were defined as follows: East-North (Pl, Ee, Lv and Lt), East-Centre (Cz, Hr and Si) and East-South (Hu, Sk, Bg and Ro).

### 4. RESULTS

We have decided to conduct the main analysis using the stocks of foreign direct investments/capita. The main reason behind this decision was to ensure a better comparability between the analysed economies, which are of very different sizes.

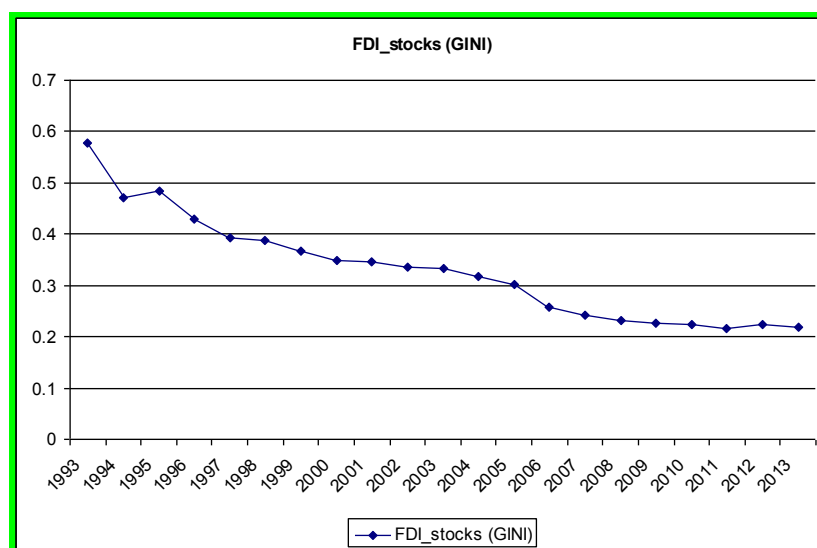
**Table 1. The stocks of FDI/capita (US \$) for 1993 & 2013; as % from EU15\*\_avg**

	St_1993	St_2013	93/2013 (%)	93/EU15*	2013/EU15*
Estonia	294.3	10591.0	2.78	9.34	64.97
Latvia	147.2	4421.2	3.33	4.67	27.12
Lithuania	82.9	3899.4	2.13	2.63	23.92
Poland	113.6	5244.4	2.17	3.61	32.17
Cz Republic	750.6	9671.5	7.76	23.82	59.33
Croatia	126.5	5854.7	2.16	4.01	35.92
Slovenia*	1609.7	5976.2	26.94	51.09	36.66
Hungary	1017.3	9482.1	10.73	32.29	58.17
Slovakia	253.9	7231.6	3.51	8.06	44.36

Bulgaria	59.8	4687.2	1.28	1.90	28.75
Romania	23.2	2670.8	0.87	0.74	16.38

As it might be expected Romania and Bulgaria possessed the lowest FDI stocks/capita in 1993. At the end of the analysed period Romania is still on the last place with only about 68% from the stocks/capita of Lithuania (10<sup>th</sup> place) and almost 16.4% from EU15\* average. Noteworthy is the fact that, at the end of the analysed period, Bulgaria managed to increase its stock, so that it surpassed Latvia and Lithuania. Another significant aspect is represented by the evolution registered by Estonia which managed to increase its stocks /capita so that it leads the hierarchy in 2013 (it reached almost 65% from EU15\* average).

During the entire period Romania registered the highest growth rate, and Slovenia registered the lowest one (these two economies were also ending and leading the hierarchy in 1993). Also important to mention is that Slovenia is the only one which presents a negative convergence index.



**Figure 1. The evolution of the Gini coefficient for the entire area**

Noteworthy is the fact that the disparities (concentration) were very important at the beginning of the studied period, when the coefficient was almost 0.6, and they have decreased continuously until 2013, when the coefficient reached values little over 0.2.

Another important finding is represented by the fact that the decreasing trend has been significantly affected by the global crisis and, therefore, after the year 2009 it shows constantly values around 0.22.

Also notable is the fact that the disparities decreased in all three subareas included in this analysis, but the rhythm is significantly different.

## 5. CONCLUSIONS

Before going further with the final remarks, we need to mention that, taking into consideration the limitations imposed by the quantitative tools employed and by the set of data used, the findings of the present study should be regarded with caution. They should be used as basis in more complex research studies and could also be used by the

policymakers as a starting point when trying to optimize or to develop new policies in order to increase the attractiveness of their economy for foreign investors.

The study revealed that, at the beginning of the period, Romania had the lowest stock, followed by Bulgaria and that Slovenia was leading the hierarchy. Even though the growth rhythm registered by Romania was the highest, due to the fact that the initial stock was very low, it is still the last at the end of the period. On the contrary Bulgaria had an impressive evolution, surpassing Latvia and Lithuania (the year 2013). Having a very small growth rhythm, Slovenia is only at the middle of the hierarchy in 2013 (negative convergence index).

Noteworthy is the fact that the disparities decreased during the entire analysed period. Also very important is the fact that the economic crisis had a significant impact on the decreasing trend and, since 2009, the value of the Gini coefficient is somehow stable around 0.22.

**Acknowledgements:** This work was supported from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007 – 2013, project number POSDRU/ 159/1.5/S/142115, project title “Performance and Excellence in Postdoctoral Research in Romanian Economics Science Domain”.

## REFERENCES

- [1] Cleeve, E. 2008. How Effective Are Fiscal Incentives to Attract FDI to Sub-Saharan Africa? *The Journal of Developing Areas*, 42, 135-153
- [2] Crozet, Matthieu & Mayer, Thierry & Mucchielli, Jean-Louis, (2004), How do firms agglomerate? A study of FDI in France, *Regional Science and Urban Economics*, Elsevier, vol. 34(1), pp 27-54, January
- [3] Schneider, F. & Frey, B. S. (1985). Economic and political determinants of foreign direct investment. *World Development*, 13, 161-175
- [4] Khadaroo, J. & Seetanah, B. (2009). The Role of Transport Infrastructure in FDI: Evidence from Africa using GMM Estimates. *Journal of Transport Economics and Policy (JTEP)*, 43, 365-365
- [5] Wheeler, D. & Mody, A. (1992). International investment location decisions. *Journal of International Economics*, 33, 57-76
- [6] Cantwell, J.A. and Iammarino, S. (2001) EU regions and multinational corporations: change, stability and strengthening of technological comparative advantages, *Industrial and Corporate Change*, 10, 1007–1037
- [7] Al-Sadig, A. (2009). The effects of corruption on FDI inflows. *The Cato Journal*, 29, 267
- [8] Taylor, J. and Bradley, S. (1997) , Unemployment in Europe: A Comparative Analysis of Regional Disparities in Germany, Italy and the UK, *International review for Social Sciences*, Volume 50, Issue 2, pp 221–245
- [9] Broadman, H. G. and Sun, X. (1997), The Distribution of Foreign Direct Investment in China, *The World Economy*, Volume 20, Issue 3, pp 339–361

# Guidelines for statistical organisations when forming Big Data partnerships

Prepared by the Task Team on Big Data Partnerships, within the Big Data project overseen by the High-Level Group for the Modernisation of Statistical Production and Services<sup>1</sup>

**Keywords:** big data, collaboration, partnership, official statistics, statistical organisation, UNECE

## 1. INTRODUCTION

The UNECE's project on The Role of Big Data in the Modernisation of Statistical Production has an objective to identify, examine and provide guidance for statistical organisations to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry.

One area of strategic importance is identified as the need to collaborate with other organisations as the reality is that no statistical organisation alone can take advantage of the opportunities, or respond to the challenges, that Big Data brings; even together the industry would struggle to develop the access to data sources, analytical capability and technology infrastructure needed to deliver Big Data strategies.

Partnership with data providers and sources is an important and often first step, but in order to optimise the use of the data, other types of partnerships might be needed. Statistical organisations might need to collaborate with each other but also to partner with organisations such as academia, scientific communities, research institutes, and technology providers, not only to develop Big Data standards, processes and methodologies, but also to gain access to organisations with analytical capability and access to the most advanced technology. Relationships with other stakeholders such as those concerned with ethical and privacy issues may also be key in order to build the trust and support required for any statistical organisation.

A task team was set up with representatives from statistical organisations across the world to examine the known issues around partnering with different types of organisation within a Big Data context. General guidelines were prepared on establishing partnerships for Big Data.

## 2. METHODS

Partnership opportunities need careful consideration across various themes to ensure that an optimal arrangement is agreed, for both statistical organisation(s) and the partner(s)

---

<sup>1</sup> The task team was composed by the following members with affiliations: Jenine Borowik, Australian Bureau of Statistics (ABS); Kenneth Iversen and Markie Muryawan, United Nations Statistical Division (UNSD); Matjaz Jug, United Nations Economic Commission for Europe (UNECE); Peter Struijs, Statistics Netherlands; Susan Williams, Office for National Statistics (ONS); and Albrecht Wirthmann, Eurostat.

involved. In order to better understand the themes to be considered, and the approaches necessary to set up a successful partnership with respect to Big Data it is necessary to first obtain information on the current experiences across a range of statistical organisations.

Two questionnaires were sent to National Statistical Offices and International Organisations. The first focused on the overall organisation and strategy for Big Data in the organisation, while the second focused on specific projects on Big Data. The questionnaire identified a number of different partnerships arrangements that contributed to the development of a report containing guidelines to consider when partnering on Big Data projects.

Feedback from the other task teams within the UNECE project was also incorporated into the report. These teams specifically provided input on the Privacy and Quality dimensions of partnering on Big Data projects but also the partnering arrangements experienced within each of the individual pilot projects in the practical Sandbox element of UNECE's overall project.

### **3. RESULTS**

The examples from the questionnaire and Sandbox demonstrated how most partnership arrangements encounter similar forms of issues related to financial/contractual arrangements, legislative, privacy and confidentiality issues, responsibilities and ownership issues and other risks to be managed. Of course, the relative importance of these issues depend on the type of partner, be it partnerships with data providers or design, technology, IT infrastructure or analysis partners. However, many of the issues are comparable, and drawing upon the experiences from National Statistical Offices and international organization, this report highlights some important aspects in each of these issues in order to facilitate future partnerships on Big Data projects.

The questionnaire indicated that some institutions have been tapping and expanding the use of administrative data owned by line ministries and agencies (e.g., taxation records, civil registry, customs records, etc.). Other data providers are from private sector (e.g., internet companies, mobile phone operators, etc.) and many of these players are multinational enterprises. Therefore, it is important to establish good relationships with both government and private sector partners.

At the national level, a number of projects are already completed, and an even larger number of projects are in the early stages of implementation or at the idea stage. Most of these projects involve at least one partner, often from the private sector or academia. However, the questionnaire has identified that most partnerships are arranged individually and sometimes ad hoc, with some partnerships more successful than others. There are therefore important lessons learned and experiences from different organisations that should be shared in order to facilitate better and more fruitful Big Data partnerships in the future.

Well-functioning partnerships with data providers are seen as the crucial aspect in the successful implementation of any Big Data project. In the Sandbox experiments the largest point of failure was repeatedly timely access to data, often due to concerns of confidentiality. In many cases a project can only exist if a working partnership can be forged with a data provider. While the technical issues of data analysis are certainly challenging, they are often secondary to establishing a reliable data source. Furthermore, since data providers often aggregate and clean data, a good working relationship ensures



clear communication of methodology and, at times, allows for much of the data cleaning to be completed by the provider.

#### **4. CONCLUSIONS**

The activity of the task team has made an initial attempt at providing a number of initial guidelines for partnership in Big Data related projects. By building on experiences and lessons learned in a number of countries, specific guidelines for financial/contractual arrangements, legislative, privacy and confidentiality issues, and responsibilities and ownership have been established. However, further work is needed to provide more detailed and operational guidelines that could perhaps serve as a framework for partnership arrangement in Big Data projects. Such a partnership framework could include templates for different partnership agreements. These templates would provide standard suggestions for the various issues that need to be defined in each partnership agreement, such as financial issues, privacy and confidentiality, and other issues.

# A Shared Computation Environment for International Cooperation on Big Data

Matjaz Jug<sup>1</sup>, Carlo Vaccari<sup>2</sup>, Antonino Virgillito<sup>1 2</sup>  
<sup>1</sup>UNECE, <sup>2</sup>Istat - Istituto Nazionale di Statistica

**Keywords:** Big Data, modernisation, collaboration, official statistics

## 1. INTRODUCTION

The HLG (High-Level Group for the Modernisation of Statistical Production and Services) Big Data Project concerns the role of Big Data in the modernisation of official statistical production. The project is tackling strategic and practical issues that are multi-national in nature, rather than those that are specific to individual organizations or national sources. While the project does involve a practical component and a consideration of methodological issues, its aim is not to focus on the technical details of analysis of Big Data, which are covered by other national and international projects, unless these are sufficiently cross-cutting to be of concern internationally. The project is composed of four task teams:

- *Partnership Task Team*: identifies and reviews partnership models with data providers, sources and organisations such as academia, scientific communities, research institutes, and technology providers, to develop Big Data standards, processes and methodologies and to gain access to organizations to the most advances technology.
- *Privacy Task Team*: reviews existing tools for risk management in view of privacy issues, focusing on Big Data characteristics and their implications for data privacy.
- *Quality Task Team*: designs a complete quality framework for Big Data, identifying indicators for different phases of Input, Process, and Output.
- *Sandbox Task Team*: designs, installs and uses a web-accessible environment where researchers coming from different institutions can explore tools and methods needed for statistical production, verifying in practice the feasibility of producing Big Data-derived statistics.

This paper describes in detail the work of the Sandbox task team.

## 2. THE SANDBOX

A web-accessible environment for the storage and analysis of large-scale datasets has been created and used as a platform for collaboration across participating institutions. The “Sandbox” environment has been created, with support from the Central Statistics office (CSO) of Ireland and the Irish Centre for High-End Computing (ICHEC). It provides a technical platform to load Big Data sets and tools, with the goal of exploring the tools and methods needed for statistical production and the feasibility of producing Big Data-derived statistics and replicating outputs across countries. The Sandbox infrastructure is a shared distributed computational environment composed of 28 machines running a Linux operating system. The nodes are physically located within the ICHEC data centre and are connected with each other through a dedicated, high-speed network. The core software platform for Big Data processing in the Sandbox is Hortonworks Data Platform (HDP). This is a Hadoop distribution comprising the distributed file systems (HDFS), the MapReduce core, and several Hadoop side-projects used for data analysis (Pig and Hive). Besides HDP, other distributed computation tools are installed on the Sandbox that exploit the Hadoop infrastructure, namely RHadoop and

Spark. In general, all such tools are high-level interfaces to MapReduce, so they allow to write programs that translate into MapReduce jobs that are executed by the platform. Finally, a choice of databases is available (both relational and NoSQL) and a visual analytics platform (Pentaho) that allows to connect to Big Data sources and easily create visualizations that can be used for analysis or dissemination.

### 3. THE SANDBOX TASK TEAM

The Sandbox Task Team is composed of 38 participants from 18 among national institutes of statistics and international organizations. The team is organized around a set of “activity teams”, focusing on topics related to different statistical domains. The topics of interest were identified by project participants during a “sprint” meeting, held in Rome in March 2014. Then, the general objective of the experimentations were set and the activity for the acquisition of data sources was started.

#### 3.1. Activity Teams

In the following the activity teams are listed, along with a short description of their objectives:

- *Consumer Price Indexes*. Worked on testing performance of Big Data tools by experimenting the computation of price indexes through the different tools available in the Sandbox. The data sources are synthetic data sets that model price data recorded by point of sales in supermarket (“scanner data”), the use of which within price statistics is currently under study in several NSIs.
- *Mobile Phones Data*. Worked on exploring the possibility of using data from telecom providers as a source for computing statistics (tourism, daily commuting etc.). The team used real data acquired from the French telecom provider Orange.
- *Smart Meters*. Experimented the computation of statistics on power consumptions starting from data collected from smart meters reading. Two data sets are available: data from Ireland and a synthetic data set from Canada.
- *Traffic loops*. Worked on computing traffic statistics using data traffic loops installed on roads in Nederland. Methodological challenges for cleaning and aggregating data were addressed. Moreover, the team will use the Sandbox environment for testing the possibility of computing production statistics over a huge dataset (6Tb).
- *Social data*. Explored the possibility of using Twitter data generated from Mexico and collected along several months for analysing sentiment and detect touristic activity.
- *Job portals*. Worked on computing statistics on job vacancies starting from job advertisements published on web portals. Tested the approach by collecting data from portals in various countries.
- *Web scraping*. Tested different approaches for collecting data from the web.

#### 3.2. Acquisition of Data Sets

7 different datasets were loaded in the Sandbox environment. Acquisition of the datasets has been a complex task from which several lessons have been learned. One obvious outcome was that data sets that are more “interesting” from a statistical point of view, that is carrying more information and not aggregated, are in general more difficult to retrieve, since they are limited by privacy constraints. The mobile phones dataset from Orange was an interesting case in this sense. Despite the fact that the data was already freely released to researchers in the context of a competition, we had to go through a

process of legal review, only because the purpose of utilization was different from the original one. Smart meters data from Ireland were released under similar conditions. In order to adhere to the Terms & Conditions of the above mentioned data sets we had to enforce privacy in the Sandbox environment, implementing access control in order to authorize access to sensitive data only users in the relevant team. Sources from web and social medias were also used in the Sandbox experiments, in particular data from Twitter and from job portals in Slovenia and Italy, as well as price data from e-commerce web sites. Although this form of data is easily available, we experienced issues with quality, in terms of coverage and representativeness. We tested different techniques and tools for scraping data. Finally, we cite the case of some data sets that could not be acquired. Satellite data from Australia, to be used for agricultural statistics, could not be released by the providers in time for the first phase of the project and is planned to be loaded in a possible extension of the project. That would be the occasion of acquiring of other interesting data sets, such as marine and air transport data.

#### **4. PRELIMINARY FINDINGS**

At the time of writing this abstract, the activity teams have not completed their work. In the following we present some preliminary output from the teams' work.

##### **4.1. Technology**

One of the key questions behind the project was related to testing the use of tools for Big Data processing for statistical purposes. This is a novel approach for NSIs, whereas these tools were only used in an industrial context. Project results show that “traditional” tools such as relational databases and statistical software, especially if well-tuned and running on servers, still provide better performance when working on a size up to the Gigabyte order. However, the distributed approach of Hadoop allows to overcome the limits of memory and computing power of single machines, when datasets grow in the order of 10th to 100th of Gbs. The scalability of Big Data tools comes at the price of complexity of installation and management. Specialized IT support is required for set up and maintenance of the infrastructure. Although all the tools are open source, some professional support can help dealing with problems that can potentially block the production activity. Moreover, these tools are generally “unstable”. During the project activity some tools were added and several updates were required even in the short time frame of the project. Researchers should be aware of this and be prepared of working in “unfriendly” environments, as well as considering frequent switches from one tool to another.

When these data sizes are involved, data acquisition itself can be a difficult task. The traffic loops activity team had to load the full traffic loops dataset, which amounts to 6Gb. This would have taken weeks to load through FTP, so a disk had to be physically shipped to ICHEC's data centre.

##### **4.2. Methodology.**

A common computation environment enables shared work on methodology, especially where the data sets have the same form in all countries, so methods can be developed and tested in the shared environment and then applied to real counterparts within each NSI. Examples of these data sets are smart meters, scanner data, web sources and social media. As examples of this approach we tested the application of a methodology for sentiment analysis developed in the Netherlands on data from Mexico. We also obtained an R program for treating smart meters data from ONS and applied it to Canadian data

with easy data manipulation. Although web sources and social media are appropriate for sharing work, the language difference can represent a problem when cleaning and classifying text data.

Other work on methodology has been done in the mobile phones team, for computing movement of people starting from call traces recorded by antennas, and in the traffic loops team, for calculating the average number of vehicles in transit for each day and for each road. More details on this will be given in the full paper.

### **4.3. Skills**

The Sandbox environment allowed to test and to build novel competence without direct economic expenses for organizations on hardware and software licences. The collaborative approach allowed to gain a technical know-how in a relatively short time and to create a community of Big Data experts involving both statistical and IT resources. Training sessions were organized on the Sandbox tools, with training material that is available for being shared, together with findings on technology that will provide practical indications for organizations that need to set up Big Data infrastructures.

### **4.4. Statistics.**

The project is showing for the first time, on a practical basis and on a broad scale, the potential and the limits of the use of Big Data as sources for computing statistics. Improvements in efficiency and quality are possible by replacing currently used data sources with novel ones (e.g. smart meter data, scanner data, job vacancies web ads). New products can be obtained from different sources such as traffic loops, mobile phones and social data. However, sources can be of low quality so they may require some serious pre-processing before being used (e.g. web sources). In general, Big Data sources can be effectively used as additional sources, benchmarks or proxies.

The possibility of relying on a shared environment for production statistics is severely limited by privacy constraints on data sets, that often pose limits to the personnel authorized with regards to data treatment and do not allow files to be moved outside the physical boundaries of a single organization. These limitations can be partly bypassed through the use of synthetic data sets. A synthetic data set can be obtained by perturbing a privacy-sensitive data set so that it loses any links with entities of the real world, maintaining sufficient resemblance with the real thing to be considered statistically meaningful. Another solution is to generate the data by modelling its behaviour through a specifically developed software. We used both approaches in the project, respectively for smart meters data and scanner data.

## **5. CONCLUSIONS AND FUTURE WORK**

The HLG Big Data Sandbox is the first example of shared international statistical Big Data Research capability. Many statistical organisations have been working on their own research projects, however the feedback from participating organisations about potential of global big data sources, shared computation environment, international partnerships and exchange of expertise and experience has been highly positive. Sandbox therefore isn't just a set of research projects testing methodology, technology, quality and other aspects of use of Big Data for Statistics. It is practical test of a new innovative model of international collaboration and shared capabilities with the aim to leverage potential of Big Data sources, enabled by technological advances, new methodologies, partnerships and skills that would be difficult to mobilise by any individual statistical organisation.

# On Model-Representativeness

Beat Hulliger (beat.hulliger@fhnw.ch)<sup>1</sup>

**Keywords:** Non-random, convenience samples, prediction, bias, variance.

## 1. INTRODUCTION

Official statistics must deliver information for defined populations like countries or regions defined by administrative limits or socio-demographic groups or branches of the economy. The quality of such information is vital for the credibility of official statistics. Since the debate on the representative method in the first part of 20<sup>th</sup> century and the seminal paper by Neyman [1] the method of official statistics, to establish statistical information based on representative surveys with probability sample designs, is acknowledged as a corner stone of the quality and credibility of official statistics.

There are two main threats to the paradigm of representative surveys in official statistics. The first is increasing and differential non-response. The second is costs. Outside official statistics representative surveys have been abandoned mainly due to costs. While quota sampling and variants of it have been used since a long time in market research, the use of data sources from the internet of persons or things promise now a new move away from representative surveys simply because of costs. An important argument against representative surveys in this move is non-response. However, the paradigm of random sampling has not been replaced by a convincing alternative up to now. The AAPOR has issued a report where different methods for non-probability samples are discussed [2]. More research and more documentation are certainly needed for all methods proposed. The basic problem is that non-probability samples are difficult to compare with probability samples.

This paper gives a discussion of a model based version of representativeness. Such a definition has been used to establish the representativeness of a branch association needed to submit a proposal for the ordinance of a special law for the branch [3]. An example illustrates model-representativeness and the effect on variance.

## 2. METHOD

Probability sampling is the basis of representative surveys. Randomisation gives a coherent framework for sampling which allows the application of statistical concepts. In particular expectation and variance are clearly defined and thus properties of estimators can be found. The framework has been extended to asymptotic theories and the role of assisting statistical models has been justified under the randomisation approach. In particular the use of models to reduce non-response bias is a standard technique today.

A more prominent role has been assigned to models by Royall [4] in the prediction approach. Finite population characteristics are still the main objective but the role of random sampling is reduced. In spite of the debate about the value of model assisted and model based approaches the basic objective, estimation of finite population sampling characteristics, is the same.

---

<sup>1</sup> University of Northwestern Switzerland (FHNW), School of Business, 4600 Olten, Switzerland

Combinations of model-based and sampling-based estimators are prominent also in Small Area Estimation.

A radical step toward modelling is quota sampling. The randomisation approach is abandoned, though randomly filled quota are mimicking stratified random sampling. Quota sampling is model based in the sense that the quota are based on variables with predicting power for the key survey characteristics [5].

The condition for model-representativeness for a population variable can be phrased as follows: Let  $Y$  be a variable of interest and let  $X$  be a covariate, possibly multivariate. We have observations from a sample on  $Y$  and on  $X$ . The purpose is to estimate a population characteristic  $\theta(F_{U,Y})$ , where  $F_{U,Y}$  denotes the population distribution of  $Y$ . Two conditions are necessary to establish inference on  $\theta(F_{U,Y})$ :

- 1) There is a predictive model for  $Y$  based on  $X$  :  $Y_i = f(X_i) + E_i$ , where  $f(\cdot)$  denotes the structural relationship and  $E$  is an error term. The predictive model holds in the sample ( $i \in S$ ) and in the population ( $i \in U$ ).
- 2) The distribution of  $X$  in the sample  $F_{S,X}$  is the same as the population distribution  $F_{U,X}$ . And the population distribution  $F_{U,X}$  is assumed to be known.

The quality of the model is crucial. If the error is large then the explanatory power is low and if the model is not appropriate instead of bias-reduction an increase in bias may result. Therefore, representativeness must be restricted to particular variables for which there is a good model. The model quality can only be examined at the sample. The condition that the model does hold in the unobserved part of the population  $U \setminus S$  cannot be checked with the data at hand.

What can we say about  $\theta(F_{U,Y})$  under these conditions? We may estimate  $f$  from the sample, i.e.  $\hat{f}$ . Based on the assumption that the model holds for the non-sampled observations, too, we can simulate predicted values  $\hat{Y}_i = \hat{f}(\tilde{X}_i)$  by random draws from the known distribution  $F_{U,X}$  for all  $i \in U$  since we assume  $F_{U,X}$  is known. Actually the last condition, that we know the distribution  $F_{U,X}$  may be often relaxed. We get an estimate of the distribution  $F_{U,Y}$ , say  $\hat{F}_{U,Y} = F_{U,\hat{Y}}$  and we can calculate  $\hat{\theta}(\hat{F}_{U,Y})$ . If in the sample instead of the predicted values  $\hat{Y}_i$  we take the original observation  $Y$  then this is well aligned with the prediction approach from Royall [4].

The second condition above, that the sample is distributed like the population, i.e.  $F_{S,X} = F_{U,X}$ , adds robustness and simplicity to the approach. The assumption  $F_{S,X} = F_{U,X}$  will get rid of the exact form of the model, i.e. we do not have to estimate  $f$ . What is needed is that there is a predictive model involving the variables in  $X$ . Under the assumption that  $F_{S,X} = F_{U,X}$  the distribution  $F_{U,Y}$  can be estimated by  $F_{S,Y}$  directly since the realisations  $\hat{Y}_i = \hat{f}(\tilde{X}_i), i \in U \setminus S$  follow the distribution of the sampled  $Y$ , apart from the model error. Thus  $\hat{\theta} = \theta(F_{S,Y})$ . If  $F_{S,X} \neq F_{U,X}$  we may calibrate the sample such that the approximation is enhanced and then we may use a weighted version of  $F_{S,Y}$  to estimate  $F_{U,Y}$ . Alternatively we may use a regression estimator.

### 3. AN ILLUSTRATIVE EXAMPLE

We use the data set MU284 from [6] for an illustration of the principle with a simple linear model. The data set contains data on 284 municipalities in Sweden. We delete the three largest Municipalities to avoid problems with outliers. We use the per capita



revenue from taxes 1985 ( $\text{pcrmt85}=\text{RMT85}/\text{P85}$ ), the real estate value in 1984 ( $\text{pcrev84}=\text{REV84}/\text{P85}$ ), the per capita number of municipal employees ( $\text{pcme84}=\text{ME84}/\text{P85}$ ) and the proportion of the social democrat seats in the municipal council ( $\text{ssp}=\text{SS82}/\text{S82}$ ). Variable  $\text{pcme84}$  is our covariate while  $\text{pcrmt85}$  and  $\text{pcrev84}$  are examples of well modelled and bad modelled target variables.

The sample is established by the 28 municipalities in MU281 which have the lowest proportion of social democrat seats ( $\text{ssp}$ ). This can be considered a convenience sample and it will obviously be biased for the proportion of social democrat seats. The estimators considered are the sample mean, a GREG estimator [7] and a pure prediction under the simple linear regression model using  $\text{pcme84}$  as explanatory variable. Table 1 shows that the initial deviation of the sample estimate from the population target is reduced for both variables  $\text{pcrmt85}$  and  $\text{pcrev84}$  when using a regression estimator. However, the bias reduction for  $\text{pcrmt85}$  is moderate due to a model that is biased for the population. Actually in the population a quadratic form would be more appropriate. The standard error of the GREG of 0.135 seems to be biased downward when compared with the standard error of the predictor of 0.165. The downward bias of the GREG standard error seems even worse for variable  $\text{pcrev84}$ . Using  $\text{pcme84}$  as auxiliary information is useless for the variable  $\text{SSP}$  as can be seen in the last column of Table 1. The bias of the GREG and Predictor are worse than of the sample mean.

Table 1: Targets and estimates for a convenience sample

Estimator	$\text{pcrmt85}$	$\text{pcrev84}$	$\text{pcme84}$	$\text{ssp}$
Population mean	7.177	118.896	52.043	0.466
Sample mean	6.59	116.619	45.828	0.278
SEM	0.180	4.554	0.959	0.007
GREG	7.561	117.451	52.043	0.261
SE(GREG)	0.135	4.970	0	0.013
Prediction	7.561	117.451	52.043	0.261
SE(Prediction)	0.163	7.418	0	0.011
adjusted R-squared	0.680	-0.038		0.100
$\text{vinf}=\text{SE}(\text{Pred})^2/\text{SEM}^2$	0.816	2.653		2.300

If the model for the population can be estimated from the sample, as for  $\text{pcrmt85}$ , then the prediction estimator is unbiased. However, the deviation of the sample distribution of the covariates (in our example  $\text{pcme84}$ ) from the population will inflate the variance and this may offset the variance gain due to the using a regression estimator. In our example the standard error of the prediction estimator is 0.163 compared with the standard error of the mean of 0.180. Now the adjusted R-squared of the regression of  $\text{pcrmt85}$  on  $\text{pcme84}$  is  $R_a^2 = 0.680$  and we would expect a much larger reduction of the standard error due to the regression. Comparing the standard error of the prediction estimator with the standard error of the mean we obtain the variance inflation

$$\text{vinf} = \left\{ 1 + \frac{n}{n-1} \frac{(\bar{x}_U - \bar{x}_S)^2}{\hat{\sigma}_x^2} \right\} (1 - R_a^2)$$

The relative difference between the population mean and the sample mean is injected into the variance inflation of the prediction estimator. Furthermore the goodness of fit of the model plays an essential role due to the second factor in the variance inflation.

#### 4. CONCLUSIONS

In view of the pressure on using data from many sources where official statistics cannot ensure proper random sampling or where the non-response is so large that it is questionable whether the sample should be used at all it is necessary to think of models and how they can support estimation. Provided that the covariates have a similar distribution in the sample as in the rest of the population it is not necessary to find the correct form of the model to make calibration or regression estimation work even for non-random samples or when the response mechanisms are unknown. However, the quality of the inference relies on the model and must be restricted to variables that are predictable by covariates. The additional uncertainty due to deviations of the sample model from the population model cannot be assessed. In any case, the usual practice to check whether the distributions of the covariates in the sample and in the population are similar is not sufficient. Model-representativeness comes mainly through the model between the target variable and the covariates and this should be the main aspect to be verified. The similarity of the distributions of the covariates in the sample and in the population yields some robustness against miss-specification of the model form but not against a poor predictive power of the covariates. If the model in the population can be estimated from the sample then the bias of the prediction estimate may vanish. However the variance of the prediction estimator may suffer considerably when the sample distribution of the covariates is far from the covariate distribution of the population. Nevertheless, model-representativeness may become the basis for inference in situations where random sampling and a high response rates cannot be ensured. More research into this approach is necessary to establish procedures to evaluate the conditions and to establish procedures to estimate variances.

#### REFERENCES

- [1] J. Neyman, «On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection,» *Journal of the Royal Statistical Society*, pp. 558-625, 1934.
- [2] R. a. B. J. M. a. B. N. A. a. B. M. a. C. M. P. a. D. J. A. a. G. K. J. a. T. R. Baker, «Summary Report of the AAPOR Task Force on Non-probability Sampling,» *Journal of Survey Statistics and Methodology*, pp. 90-143, 2013.
- [3] B. B. M. Hulliger, «Repräsentativität von Unternehmens-Gruppen als Vertreter von Branchen,» Hochschule für Wirtschaft FHNW, Olten, 2014.
- [4] R. M. Royall, «On Finite Population Sampling Theory Under Certain Linear Regression Models,» *Biometrika*, pp. 377-387, 1970.
- [5] J.-C. Deville, «A Theory of Quota Surveys,» *Survey Methodology*, 1991.
- [6] C.-E. a. S. B. a. W. J. Särndal, *Model Assisted Survey Sampling*, Springer, 1992.
- [7] T. Lumley, "*survey: analysis of complex survey samples*". *R package version 3.30.*, 2014.

# Disclosure Risk Measurement with Entropy in Sample Based Frequency Tables

Laszlo Antal ([laszlo.antal@postgrad.manchester.ac.uk](mailto:laszlo.antal@postgrad.manchester.ac.uk))<sup>1</sup>, Natalie Shlomo ([natalie.shlomo@manchester.ac.uk](mailto:natalie.shlomo@manchester.ac.uk))<sup>1</sup>, Mark Elliot ([mark.elliott@manchester.ac.uk](mailto:mark.elliott@manchester.ac.uk))<sup>1</sup>

**Keywords:** information theory, conditional entropy, attribute disclosure

## 1. INTRODUCTION

Statistical agencies measure the disclosure risk before releasing statistical outputs, for example frequency tables. This work discusses how information theoretical definitions, such as entropy and conditional entropy, can be employed to measure the disclosure risk in sample based frequency tables. A similar approach has been followed and a disclosure risk measure has been introduced in [1] for population based frequency tables. The above mentioned disclosure risk measure is modified and applied to sample based tables in this paper. The properties around which the disclosure risk measure for population based tables is built are in [1] and can be seen below.

1. If only one cell is populated in the table, then the disclosure risk is high.
2. Uniformly distributed frequencies imply low risk.
3. Small cell values (i.e. ones and twos) are more disclosive than higher values. In general, the greater the cells, the lower the disclosure risk.
4. Assume that two tables are given and there is only one cell populated in each table. The frequencies of the non-zero cells are equal. In this case we deem the table that has more cells (and therefore more zeroes) to be of higher disclosure risk.
5. We would like the disclosure risk measure to be bounded by 0 and 1.

Below we compare the disclosure risk measurement of population based tables to sample based ones.

## 2. METHODS

### 2.1. Information theoretical definitions

In order to define the disclosure risk measure below, we need to introduce the entropy and the conditional entropy. The entropy of an  $X$  random variable is defined as

$$H(X) = - \sum_{i=1}^K \Pr(X = c_i) \cdot \log \Pr(X = c_i)$$

where  $C = \{c_1, c_2, \dots, c_K\}$  is the range of  $X$ . Entropy is zero if the distribution of  $X$  is degenerate, and it is maximal if the distribution is uniform.

Assume that  $X$  and  $Y$  are two random variables with a common range  $C = \{c_1, c_2, \dots, c_K\}$ . (The domain of the variables must also be common.) The  $H(X|Y)$  conditional entropy is defined as follows.

---

<sup>1</sup>University of Manchester, UK

$$H(X|Y) = - \sum_{j=1}^K \Pr(Y = c_j) \cdot \sum_{i=1}^K \Pr(X = c_i|Y = c_j) \cdot \log \Pr(X = c_i|Y = c_j)$$

It is known that the conditional entropy cannot exceed the entropy,  $H(X|Y) \leq H(X)$ .

## 2.2. The disclosure risk measure

In this work we focus mainly on attribute disclosure. Attribute disclosure occurs if confidential information about an individual or more individuals can be retrieved from the released data. Attribute disclosure is likely to happen if the individuals are concentrated in one cell of the frequency table, i.e., the distribution of the table is degenerate, while the chance of attribute disclosure is low if the frequencies are distributed uniformly. This fact is in line with properties 1 and 2 listed above and is reflected exactly by  $1 - \frac{H(X)}{\log K}$ .

### The disclosure risk measure for population based tables

Denote the population based frequency table by  $F = (F_1, F_2, \dots, F_K)$ . It implies that the size of the frequency table is  $K$ . Assume that the distribution of  $X$  is  $(\frac{F_1}{N}, \frac{F_2}{N}, \dots, \frac{F_K}{N})$ , where  $N = \sum_{i=1}^K F_i$ . The disclosure risk for a population based frequency table is a weighted average defined as follows.

$$R_1(F, \mathbf{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \quad (1)$$

Here  $D$  denotes the set of zeroes in the table, therefore  $|D|$  provides the number of zeroes.  $\mathbf{w} = (w_1, w_2, w_3)$  is the vector of weights, while  $e$  is the base of the natural logarithm. The disclosure risk measure reflects the properties listed in section 1.

If a statistical agency deem that a population based frequency table is of high disclosure risk, then a statistical disclosure control (SDC) method must be applied to the table. Denote the perturbed frequency table by  $G = (G_1, G_2, \dots, G_K)$ . The disclosure risk of the perturbed table also must be assessed. However, the previously defined disclosure risk measure cannot be applied to the perturbed table since the perturbed table has more uncertainty. We define the disclosure risk after perturbation below.

$$R_2(F, G, \mathbf{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \quad (2)$$

Here  $E$  is the set of zeroes in the perturbed frequency table. The disclosure risk of a perturbed table is expected to be lower than that of the original table.  $R_2(F, G, \mathbf{w})$  satisfies this requirement.

### 2.2.1. The disclosure risk measure for sample based tables

While population based tables include every individual, only selected individuals contribute to sample based frequency tables. The  $R_1(F, \mathbf{w})$  and  $R_2(F, G, \mathbf{w})$  disclosure

risk measures are developed for population based tables and cannot be applied directly to sample based ones.

Sampling can be considered as a special SDC method. The smaller number of individuals in sample based tables ensures protection against attribute disclosure to a certain extent. An intruder faces more uncertainty in a sample based table than in a population based table. Zeroes in the sample based table seemingly increase the chance of attribute disclosure. However, a zero in the sample based table is not necessarily zero in the population based table. If we consider the sample based table as the result of an SDC method, then  $R_2(F, G, \mathbf{w})$  should be applied in order to evaluate the disclosure risk, where  $F$  is the population based table and  $G$  is the sample based table.

Our assumption is that only a sample based table and  $N$  are known, therefore our aim is to estimate  $H(X)$ ,  $H(X|Y)$  and  $|D|$  for the population based table. In this paper this aim is achieved by estimating population frequencies. From the estimated population frequencies the above mentioned quantities can be calculated as for ‘true’ population based tables. The table cell probabilities may be estimated from the sample by models.

### 3. RESULTS

Our preliminary results are given below. More detailed results will be available in the full paper.

The data we used is an extract from the 2001 UK census tables. We applied the disclosure risk measures to two-dimensional tables. The population of 10 selected output areas consists of 2449 individuals. Three sampling fractions are considered here, they are 0.1, 0.05 and 0.01. Both the population based and sample based tables were generated. A log-linear model, applied to the sample based table, provided the estimated cell probabilities. Denote these probabilities by  $P = (p_1, p_2, \dots, p_K)$ . The sample based table, denoted by  $f = (f_1, f_2, \dots, f_K)$ , and the population based table are assumed to follow the multinomial distribution,

$$f \sim \text{Multinom}(n; p_1, p_2, \dots, p_K),$$

$$F \sim \text{Multinom}(N; p_1, p_2, \dots, p_K).$$

It implies that the distribution of  $F - f$  is also multinomial,

$$F - f \sim \text{Multinom}(N - n; p_1, p_2, \dots, p_K) \quad (3)$$

Therefore we drew  $N - n$  ‘new individuals’ from (3) and added them to the sample based table, thereby estimating the population frequencies. (The full paper will discuss more models to estimate the population frequencies.)  $R_1(F, \mathbf{w})$  and  $R_2(F, f, \mathbf{w})$  can be calculated and compared on both the known and estimated population frequencies. For each sampling fraction the sampling was carried out 1,000 times, for each sample based table the population frequencies were estimated 1,000 times and the average disclosure risk measures were calculated. The results are shown below for the output area (10 output areas)  $\times$  religion table, which has a size of  $K = 90$ . The weights we used are  $\mathbf{w} = (0.1, 0.8, 0.1)$  for both (1) and (2). More numerical results will be available in the full paper.

		Sampling fraction		
		0.1	0.05	0.01
$R_1(F, \mathbf{w})$	From true population frequencies	0.2315	0.2315	0.2315
	From estimated population frequencies	0.2299	0.2417	0.3106
$R_2(F, f, \mathbf{w})$	From true population frequencies	0.1695	0.1533	0.0955
	From estimated population frequencies	0.1720	0.1711	0.1881

Table 1. Disclosure risk measures of output areas  $\times$  religion table from true population frequencies and from estimated population frequencies for three sampling fractions

#### 4. CONCLUSIONS

The preliminary results indicate reasonable estimates for  $R_1(F, \mathbf{w})$  and  $R_2(F, f, \mathbf{w})$  for a two-dimensional table. These disclosure risk measures were applied to population frequency tables previously. The paper demonstrates that the disclosure measures can be applied to sample based frequency tables. The estimations are less precise for lower sampling fractions but it may be attributed to the small number of individuals (and therefore more zero frequencies) in sample based tables. Investigation of more estimation methods, especially for low sampling fractions, needs to be accomplished.

In the literature disclosure risk measures have been applied to cells. In this paper  $R_1(F, \mathbf{w})$  and  $R_2(F, f, \mathbf{w})$  quantify the disclosure risk of the entire frequency table. The relatively easy computation of  $R_1(F, \mathbf{w})$  and  $R_2(F, f, \mathbf{w})$  makes the disclosure risk assessment quick and they can be estimated ‘on-the-fly’. The properties of  $R_1(F, \mathbf{w})$  and  $R_2(F, f, \mathbf{w})$  will be further investigated.

#### REFERENCES

- [1] Antal, L., Shlomo, N. and Elliot, M. (2014). Measuring Disclosure Risk with Entropy in Population Based Frequency Tables. In PSD'2014 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer), Springer LNCS 8744, pp. 62-78.
- [2] Antal, L., Shlomo, N. and Elliot, M. (2014). Measuring Disclosure Risk and Information Loss in Population Based Frequency Tables. [http://www.cmist.manchester.ac.uk/medialibrary/archive-publications/reports/2014-02-Measuring\\_Disclosure\\_Risk.pdf](http://www.cmist.manchester.ac.uk/medialibrary/archive-publications/reports/2014-02-Measuring_Disclosure_Risk.pdf)
- [3] Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory. Wiley, 2nd edition.
- [4] Oganian, A and Domingo-Ferrer, J. (2003). A Posteriori Disclosure Risk Measure for Tabular Data Based on Conditional Entropy. Statistics and Operations Research Transactions 27, pp. 175-190.
- [5] Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. Lecture notes in statistics. Springer.

# Privacy Preserving Probabilistic Record Linkage

Duncan Smith (Duncan.G.Smith@manchester.ac.uk)<sup>1</sup>

Natalie Shlomo (Natalie.Shlomo@manchester.ac.uk)<sup>2</sup>

**Keywords:** Anonymisation, String comparators, Similarity scores, EM algorithm

## 1. INTRODUCTION

Probabilistic record linkage as set out in the seminal paper [1] is an important area of research in official statistics. This is especially true as more and more administrative sources are being used to improve the quality of surveys or to replace traditional censuses. Traditionally, all datasets are held within one organization, eg. the National Statistics Institute, and record linkage is carried out on original variables, such as first name, last name, ID number, etc. without the need to anonymise these strings. Probabilistic record linkage joins all possible pairs on a set of matching variables (typically within blocks) and the pairs are classified to matches, non-matches and clerical review according to a test statistic based on the likelihood ratio under a Bayesian framework. The matching parameters for the linkage are estimated using an EM algorithm under the Binomial distribution for an agree/disagree (0,1) indicator function. String comparators are used to scale the likelihood ratios to improve the classification.

Data on individuals however can be contained in distinct databases held by different organisations and there may be a variety of data sharing agreements. A common practice is to suppress information that might directly identify an individual which inhibits the possibility of record linkage. For this reason, techniques have been developed to anonymize data in such a way that they can still be used for linkage. Exact matching can be carried out on anonymised strings and methods of classification based on ‘fuzzy’ matching have been introduced in the privacy preserving record linkage literature. A similarity score is calculated to measure the distance between two anonymised strings. One drawback is that when strings are anonymised for record linkage, it is not possible to carry out a clerical review of potential matches for the ambiguous cases and the possible linkages must be dichotomised into true matches and non- matches.

In this paper, we present a new method for classifying pairs into matches and non-matches using an EM algorithm under the Multinomial distribution where string comparators in the non-private setting or similarity scores in the private setting can be used to define classes within which matching parameters can be estimated. In this way, the string comparator/similarity score is included directly into the models for estimating matching parameters instead of through the ad-hoc method of scaling the likelihood ratios and classification into matches/non-matches is much improved. In addition, we examine methods for string anonymisation and propose a new method which provides protection against ‘attack’ scenarios on anonymised strings.

---

<sup>1</sup> The Cathie Marsh Institute of Social Research, University of Manchester

<sup>2</sup> Social Statistics, School of Social Sciences, University of Manchester



## 2. METHODS

### 2.1. Probabilistic Record Linkage

From two datasets A and B we produce the set of all possible matches:  $A \times B = \{(a, b); a \in A, b \in B\}$  typically within a block defined by an exact match on an error free variable, such as geographical area, to reduce the number of pairs that need to be investigated. We aim to classify the pairs into sets: Matches  $M = \{(a, b) | a = b, a \in A, b \in B\}$  and non-matches  $U = \{(a, b) | a \neq b, a \in A, b \in B\}$ . [1] define a decision rule based on the likelihood ratio of agreement  $m(\gamma)/u(\gamma)$  or disagreement  $[1 - m(\gamma)]/[1 - u(\gamma)]$  where  $m(\gamma)$  is the probability of agreement for the comparison vector  $\gamma$  given a match and  $u(\gamma)$  is the probability of agreement for the comparison vector  $\gamma$  given not a match. In the simplest form, the comparison vector will be a vector of 1's or 0's where 1 denotes an agreement in matching variable  $q$  ( $q=1, \dots, Q$ ) and 0 otherwise. Under conditional independence, we can treat each matching variable separately and define  $m_q = P(\gamma_q^j = 1 | (a, b)_j \in M)$  and  $u_q = P(\gamma_q^j = 1 | (a, b)_j \in U)$ . The two probabilities for each matching variable and the overall number of correct matches  $P(M) = P((a, b)_j \in M)$  is estimated using the EM algorithm under the Binomial distribution. Further details will be provided in the paper.

String comparators are used to scale the likelihood ratios. The Jaro string comparator is commonly used for official statistics and defined as:

$$\Phi_q(X_a, X_b) = 1/3 \times (\#common / str\_len1 + \#common / str\_len2 + (1 - 0.5 \times (\#half\ transposition / \#common)))$$

where str\_len1 and str\_len2 length of strings, #common is number of common letters and #half transposition where a letter can move one position left or right. The Jaro-Winkler string comparator is based on the Jaro string comparator but provides different weights depending on the position in the string. The string comparator is used to down-weight the agreement likelihood ratio so that if there is little difference in the string, the pair will not be regarded as a disagreement rather will receive a value at a proportional distance from the agreement likelihood ratio. The new likelihood ratio is defined as:

$$\Phi_q(X_a, X_b) \frac{m_q}{u_q} + (1 - \Phi_q(X_a, X_b)) \frac{1 - m_q}{1 - u_q}. \text{ Alternately, the log-likelihood can be}$$

$$\text{adjusted as follows: } \left( \frac{m_q}{u_q} \right)^{\Phi_q(X_a, X_b)} \left( \frac{1 - m_q}{1 - u_q} \right)^{(1 - \Phi_q(X_a, X_b))}.$$

### 2.2. String Anonymisation

Hash functions are used to anonymise strings. These hash functions convert strings to integers where equal strings produce equal hash values so linkage can be carried out on the equality or inequality of hash values. Strings are converted to tokens, typically bigrams, and each bigram is hashed [2]. For example, for the names John and Jon we obtain:

$$\begin{aligned} \text{'john'} &\rightarrow \{\text{'jo'}, \text{'oh'}, \text{'hn'}\} \rightarrow \{21299418, 21496024, 20971735\} \\ \text{'jon'} &\rightarrow \{\text{'jo'}, \text{'on'}\} \rightarrow \{21299418, 21889246\} \end{aligned}$$

A similarity score such as the Dice Coefficient can be used:  $D_{A,B} = \frac{2 \times (\#common)}{Total\ hashes} = \frac{2}{5}$

[3] propose the use of Bloom filters (an array of 0 and 1) which can be represented as an integer in its binary form.  $m$  bits (all initially set to 0) and  $k$  hash functions are used to map an element to one of the  $m$  array positions. To add an element, produce  $k$  hash functions to get  $k$  array positions and set these positions to 1. To query whether an element is in the set, feed it to each of the  $k$  hash functions to get  $k$  array positions and check whether there is a zero. If there's a zero, the element is not in the set, otherwise if all 1's it might be in the set since there may be false positives. The Dice coefficient can be estimated from a pair of Bloom filters. These methods may be open to attacks where intruders can learn the length of strings or identify tokens, eg. bigrams.

[4] proposed a technique called minwise hashing where many hash values are calculated for a set of tokens and the minimum hash values returned. The probability of a hash collision on the minimum hash value is the Jaccard Similarity Score defined as

$$J_{A,B} = \frac{\#common\ hashes}{total\ discrete\ hashes}.$$

Using the minwise hashing, the estimate of the Jaccard Similarity Score is the number of collisions on the minimum hash value where  $m$  is the number of hash functions:  $n \sim Bin(m, J_{A,B})$  and the estimate is  $\hat{J}_{A,B} = \frac{n}{m}$ . The variance

$$is: Var(\hat{J}_{A,B}) = \frac{J_{A,B}(1-J_{A,B})}{m}.$$

[5] propose to return only  $b$ -bits of the minimum hash value. In [6], we propose to return only 1-bit of the minimum hash value to produce a concatenated 1-bit minwise hashing. In this case, the estimate of the Jaccard Similarity

$$score is  $\hat{J}_{A,B} = 2\frac{n}{m} - 1$  with a variance of:  $Var(\hat{J}_{A,B}) = \frac{1-J_{A,B}^2}{m}.$$$

In the example in Table 1, we show the first 5 of the minwise hash values and the 1-bit minwise hash values. From table 1, for minwise hashing with 5 hash functions, the estimate of the Jaccard similarity score is 2/5 and for the concatenated 1-bit hashing 3/5, the true value is 1/4. More on string anonymisation will be covered in the paper.

**Table 1. Minwise hashes and 1-bit minwise hashes under a binary representation for  $S1=\{'jo','oh','hn'\}$  and  $S2=\{'jo','on'\}$**

	H1	H2	H3	H4	H5	...	Hm
S1	451153726	1123790273	2501120381	2030682762	965995804		
S2	797504823	1123790273	262296169	1744666338	965995804		
...							
Sn							
	H1	H2	H3	H4	H5	...	Hm
S1	0	1	1	0	0		
S2	1	1	1	0	0		
...							
Sn							

### 2.3. Multinomial EM Algorithm

We extend the EM algorithm where instead of two categories agree/disagree for each matching variable and then scaling the likelihood ratios with string comparators, we define  $k$  categories,  $k=1, \dots, K$  where each category represents a class based on an interval of string comparators (or similarity scores in the private setting). For example, 8 classes with (inclusive) upper bounds based on the string comparator/similarity score:  $[0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.999, 1]$ .

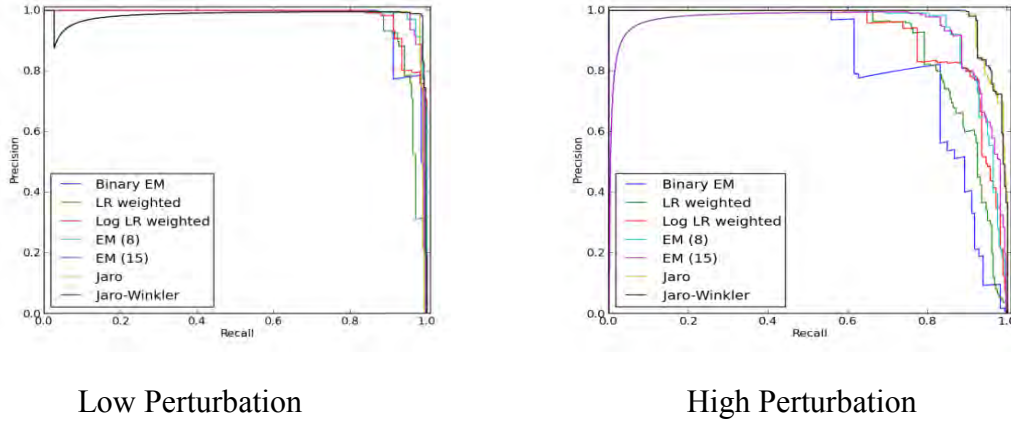
If there is an agreement for variable  $q$  in class  $k$  of the string comparator/similarity score for a pair  $j$ , then:  $\gamma_{q,k}^j = 1$ , otherwise it is zero. The EM algorithm under the multinomial distribution is now used to estimate the match parameters for each variable  $q$  in class  $k$  (not shown here). Further details about the method will be in the paper.

### 3. EXPERIMENT

From a 1995 Israel Census Sample file, we select 700 records as a test file. From this file, 400 records were selected and varying levels of perturbation applied. We show in Figure 1 the low and high levels of perturbation. No blocking was carried out. Since we know the match status, we can evaluate the linkage using precision/ recall plots where:

$$\text{Precision} = \frac{tp}{tp + fp} \quad \text{and} \quad \text{Recall} = \frac{tp}{tp + fn}, \quad \text{where } tp \text{ are the true positive, } fn \text{ the false negatives and } fp \text{ the false positives under varying levels of thresholds.}$$

The best approaches produce curves in the upper right sector of the plot. The results of the experiment are shown in Figure 1.



**Figure 1. Precision/Recall plots for low and high level of perturbation for: Binary EM algorithm (no use of similarity score); down-weighted likelihood ratio (LR); down-weighted log-likelihood ratio (Log LR); 8-bin EM algorithm with estimated Jaccard Score; 15-bin EM Algorithm with estimated Jaccard Score; 8-bin EM algorithm with Jaro String comparator; 8-bin EM algorithm with Jaro-winkler String comparator**

### 4. CONCLUSIONS

From Figure 1, all approaches perform better with low levels of perturbation. The Binary EM without similarity scores performs the worst. Down-weighting log likelihood ratios outperforms down-weighting of likelihood ratios. Multinomial EM outperforms Binary EM with no clear difference between 8 classes and 15 classes Jaccard score schemes. The non-private setting with Jaro and Jaro-Winkler string comparators provide the best performance, although these are not privacy preserving.

Given that in the private setting, there is no possibility to carry out clerical review it is important that we improve the classification of pairs into matches and non-matches. The multinomial EM Algorithm approach shows an improvement over the traditional method of the Binomial EM algorithm and down-weighting of likelihood ratios.

## ACKNOWLEDGEMENT

*The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries).*

## REFERENCES

- [1] I.P. Fellegi and A.B. Sunter, A theory for record linkage, JASA, Vol. 64 No. 238 (1969), 1183-1210.
- [2] T. Churches and P. Christensen, Some methods for blindfolded record linkage, BMC Medical Informatics and Decision Making, 4:9 (2004).
- [3] R. Schnell, T. Bachteler, and J. Reiher, Privacy-preserving record linkage using Bloom filters, BMC Medical Informatics and Decision Making, 9:41 (2009).
- [4] A. Z. Broder, On the resemblance and containment of documents, Compression and Complexity of Sequences: Proceedings IEEE, (1997), 21-29.
- [5] P. Li and A.C. König, Theory and applications of b-bit minwise hashing, Communications of the ACM, Vol. 54, No. 8, (2011) 101-109.
- [6] D. Smith and N. Shlomo, Privacy Preserving Record Linkage, Data Without Boundaries Deliverable D11.2, Report 2014-01, CMIST Working Paper (2014) [http://www.cmist.manchester.ac.uk/medialibrary/archive-publications/reports/2014-01-Data\\_without\\_Boundaries\\_Report.pdf](http://www.cmist.manchester.ac.uk/medialibrary/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf)

# Development of pseudonymised matching methods for linking multiple administrative datasets

Pete Jones, Office for National Statistics

**Keywords:** record linkage, pseudonymisation, administrative data, census, population

## 1. Introduction

This paper discusses research to develop methods to link record level administrative datasets that have been pseudonymised at source. Matching multiple administrative sources is both resource intensive and elevates risks relating to the privacy of data about people and households. This paper outlines new approaches that have been developed to accurately link datasets that have been pseudonymised with a secure hashing algorithm.

It is well known that pseudonymising data prior to linkage inhibits the use of probabilistic matching algorithms and clerical resolution. Critical to the research has been the development of techniques that can identify similarity between pseudonymised matching fields and the use of statistical techniques to accurately designate candidate pairs into matches and non-matches. A quality assurance exercise has been undertaken to test these methods in a comparison of match results between the NHS Patient Register and the 2011 Census. Results so far are highly promising showing relatively low levels of false positives (less than 1%) and false negatives (around 2%). The paper will also discuss further ideas for improving the quality still further.

## 2. Methods

### 2.1 Deterministic Matching

The matching algorithm comprises three stages. The first is deterministic linkage based on a sequence of *link-keys* that are derived on each dataset prior to importing data into the linkage environment. Link-keys are based on concatenations of selected components of identifiers (i.e. names, dates of birth and addresses), resulting in unique identifiers for the majority of records on any dataset. These identifiers are then pseudonymised with the hashing algorithm and used to match records between datasets in the linkage environment.

The algorithm runs a sequence of pseudonymised link-keys, each of which attempts to resolve particular types of inconsistency that occur across match fields. Records linked uniquely are designated as matches, with the remaining residuals moved to the second stage of the algorithm.

### 2.2 Similarity Tables

Link keys are successful in identifying true matches between records that have relatively low levels of disagreement between match fields. For more complex cases involving disagreement across multiple match fields, measures of similarity between match fields are

required. A method for achieving this has been developed through the construction of *similarity tables* which are also constructed during pre-processing and before pseudonymisation.

Two de-duplicated lists of forenames and surnames occurring across the two datasets are compiled prior to pseudonymisation process. String comparison algorithms are undertaken between all names in the list to develop a thesaurus of names that are similar, with a comparison metric indicating the level of similarity. These tables are then pseudonymised and imported into the linkage environment and used to identify candidate match pairs between the two datasets.

### **2.3 Score based matching**

From the records identified as possible matches from the similarity tables, a small sample of candidate pairs are made available for clerical matching (approx. 1000 pairs). A decision is made whether to designate each pair as a match or non-match. This clerically matched dataset now serves as training data to classify the match status of all other candidate pairs identified by the similarity tables. The current method relies on a logistic regression model, where the dependent variable is the binary outcome (yes or no) regarding match status. The predictors include the following: name similarity scores, date of birth similarity score, name commonality, sex agreement, postcode agreement and geographic distances between locations. Candidate pairs with a match likelihood  $\geq 0.5$  are designated as matches, those below as non-matches.

### **2.4 Comparison Study**

The pseudonymised matching algorithm has been tested and compared with the results of a previously undertaken linkage exercise, where links were identified using a combination of exact matching, probabilistic automatching, clerical resolution and clerical searching. In both cases a sample of NHS Patient Register records have been linked to the 2011 Census of England and Wales. The results of this matching exercise have been used as a measure of quality for the pseudonymisation algorithm on the basis that it is close to a ‘gold standard’ in identifying all of the true matches between the two datasets.

The pseudonymisation process and the methods developed are described in detail in ONS(2013).

## **3 Results**

Table 1 shows the results of the comparison exercise. The match rates achieved by the pseudonymisation method are slightly lower than those achieved by the Census matching team. Eight local authorities were used as the basis for comparison, five of which were in cities (London and Birmingham), and three were in more rural areas. Matching scenarios are generally more complex in city areas owing to high rates of population migration and greater ethnic diversity in naming conventions. The table shows that the false positive rate is only slightly higher in city areas than more rural areas, however the increase in false negatives is more evident.

**Table 1 – Match rate comparison and errors for pseudonymised matching method**

<b>Local Authority Type</b>	<b>Census match rate</b>	<b>Pseudonymised match rate</b>	<b>Pseudonymised false positive rate</b>	<b>Pseudonymised false negative rate</b>
City Local Authorities	71.3%	70.3%	0.5%	2.0%
Rural Local Authorities	91.2%	90.7%	0.3%	0.8%
Total	72.7%	71.7%	0.5%	1.9%

#### **4 Conclusions**

The results of this comparison study indicate potential for a quality pseudonymised linkage approach. The method described has been implemented to produce trial outputs on the size and characteristics of the population from 2015 onwards. It may not be possible that pseudonymisation methods will meet the very high quality levels achieved through more conventional approaches. However research continues in this area, with a provisional target of lowering the false negative rate to less than 1 per cent.

The motivation for researching a pseudonymisation approach is critical to the development of future Census taking in the UK. Exploring administrative data as the basis for improving the quality of census outputs will be a major focus in preparations for the 2021 Census and research continues to investigate the potential of an administrative data based census to replace the traditional ten yearly census in England and Wales after 2021. In undertaking this research, linkages between administrative datasets will be made on a large scale (most datasets have tens of millions of records) and consideration for the associated risks to privacy need to be taken into account. This necessitates a different and more automated approach to be taken. There may be other solutions to alleviate the privacy risks in due course, but the scale of the matching in this context prohibits anything other than a highly automated approach.

#### **REFERENCES**

- [1] Office for National Statistics, Beyond 2011: Matching Anonymous Data. Methods & Policies Report (M9). Available at: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html> (2013).

# Generating synthetic geocoding information for public release

Jingchen Hu ([jh309@stat.duke.edu](mailto:jh309@stat.duke.edu)), Jörg Drechsler ([Joerg.drechsler@iab.de](mailto:Joerg.drechsler@iab.de))

**Keywords:** confidentiality; disclosure control; geocoding; synthetic data

## 1. INTRODUCTION

In recent years more and more statistical agencies started collecting detailed geocoding information for some of their surveys or administrative databases. This information enables researchers to define their own geographical levels when investigating spatial effects. Furthermore, detailed geocoding information can be used to link data from different sources. However, these additional research opportunities come at the price of increased risks of re-identification. For this reason external researchers usually cannot get access to the detailed geocodes.

Generating synthetic data is an innovative approach for disseminating data to the public with high utility and low risks [1]. In our paper we compare three different strategies for generating synthetic geocodes: The first approach (denoted DPMPM in the results section) uses Dirichlet process mixture of products of multinomials for the synthesis, treating the geocoding information as categorical with combined information of its latitude and longitude. The second approach (denoted as CART1) is based on CART models [2], also treating the geocoding information as one categorical variable. The final approach (CART2 and CART3), originally suggested by [3], treats the information on the latitude and longitude as two separate continuous variables and generates synthetic values by using CART models for both variables. CART2 and CART3 differ in the ordering of the two variables to be synthesized.

Our evaluations are based on a subset of the Integrated Employment Biographies (IEB), a rich administrative data source at the Institute for Employment Research (IAB). We refer to [4] for a detailed description of the database. We selected all 3,537,556 complete-case observations from the state of Bavaria and for 6 variables shown in Table 1.

## 2. THE CART AND DPMPM SYNTHESIZER

As noted in [1], CART models are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. CART models can also be used to generate partially synthetic data ([2]).



Table 1: Variables included in the dataset

variable	characteristics
exact geocoding info	recorded as distance in meters from the point 52 northern latitude (Y), 10 eastern longitude (X)
sex	male/female
foreign	yes/no
skills	low/medium/high
wage	low/medium/high
distance to work	5 categories ( $\leq 1$ , 1-5, 5-10, 1-20, $>20$ km)

The DPMPM uses a Dirichlet process mixture of products of multinomial distributions, which is a Bayesian version of a latent class model for unordered categorical data. A full description of the models will be included in the final paper.

### 3. RESULTS

For our evaluations we only synthesized the geocoding information for the place of living. To run the DPMPM model on the entire dataset is very expensive, so we cluster all the complete-case observations based on their geographic locations into clusters containing 20,000 records (for computational reasons, each of the last two clusters contains roughly 19,000 records). All synthesis models are run separately on each cluster.

#### 3.1. Analytical validity

For our utility evaluations we assume that potential users of the data would be interested in relative frequencies for various cross tabulations of the variables contained in the dataset on a detailed geographical. Thus, we compute these frequencies on a specific geographical level (zip code). Table 2 presents the distribution of the absolute differences of the frequencies between the original and the synthetic data across all possible cells with a minimum cell sizes of 50 records for the different synthesizers.

The DPMPM provides slightly better results than the CART models for all cross tabulations. However, the differences are relatively small and all methods perform relatively similar in terms of analytical validity. There are hardly any differences between the CART models but it should be noted that CART2 and CART3 can produce implausible geocodes, such as places of living in the middle of a lake or in industrial areas. This cannot happen with CART1 since the geocodes are modelled as categorical and thus only geocodes that were observed in the original data could appear in the synthetic data.

Table 2: Summary of the DPMPM model and the CART model results

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	# of cells
one-way							
DPMPM	-46.26	-2.43	-0.16	-0.54	2.14	37.47	27117
CART1	-52.58	-2.56	-0.18	-0.56	2.29	40.36	27116
CART2	-55.39	-2.63	-0.21	-0.56	2.30	58.39	27113
CART3	-52.65	-2.66	-0.21	-0.56	2.28	59.51	27112
two-way							
DPMPM	-45.60	-2.23	-0.25	-0.80	1.28	35.31	102802
CART1	-52.29	-2.36	-0.28	-0.83	1.37	38.59	102802
CART2	-53.46	-2.41	-0.28	-0.82	1.39	53.90	102732
CART3	-52.33	-2.40	-0.28	-0.82	1.38	39.31	102712
three-way							
DPMPM	-35.87	-1.82	-0.30	-0.85	0.67	22.15	149528
CART1	-40.50	-1.91	-0.33	-0.88	0.72	25.28	149525
CART2	-42.22	-1.93	-0.32	-0.86	0.74	35.43	149309
CART3	-40.66	-1.93	-0.32	-0.87	0.73	36.74	149297

### 3.2. Disclosure Risk Evaluations

To evaluate disclosure risks, we compute probabilities of identification using methods developed in [5]. The details are omitted for brevity and a full description will be included in the final paper. For our evaluations we assume that the intruder knows the exact geocode, sex, and the information whether the individual is a foreigner or not and uses this information to try to identify the individuals in the database. We sample 100 records from each cluster and assume that these  $J=17,700$  records are the target records that the intruder tries to find in the data. For the geocode we assume that the intruder constructs grids of different size and considers all records that fall in the same grid as matches. We evaluate the risks for three different grids: 50x50, 500x500, and 2000x2000 square meter grids. We also compute the risk measures if the intruder would match on the exact geocodes. Details on the two risk measures presented in Table 3 will be included in the final paper. To compare the synthesizers it suffices at this point to note that larger values mean higher risks for both measures.

The risks are very small in all scenarios. The fraction of correctly identified single matches (the true rate) is far below 0.1% in all cases. The expected match risk can be interpreted as the expected number of correct guesses if the intruder would pick one record at random from the records with the highest matching probability for each of his or her target records. Given that the intruder tries to identify almost 18,000 records, expected numbers that are always less than 10 indicate very low risks of disclosure. A detailed discussion on the trade-off between utility and risk will be included in the full paper.

Table 3: Expected match risk und true match rate for various grid sizes

Grid	Measures	DPMPM	CART 1	CART 2	CART 3
exact	Exp. risk	3.83	1.06	0.88	0.88
	True rate (in %)	0.018	0	0	0
50x50	Exp. risk	4.43	2.03	1.28	1.11
	True rate (in %)	0.053	0	0	0
500x500	Exp. risk	5.62	0.42	2.71	3.90
	True rate (in %)	0.040	0	0.012	0.024
2000x2000	Exp. risk	1.00	0.7	5.52	3.56
	True rate (in %)	0	0	0.039	0.020

#### 4. CONCLUSIONS

The study indicates that all synthesizers provide almost similar trade-offs between disclosure risk and analytical validity. The DPMPM synthesizer seems to perform slightly better in terms of validity but at the price of an increased disclosure risk. Still, the risk is very low for all synthesizers implying that the synthetic data could be released for all synthesizers under the given intruder assumptions. Given that the continuous CART synthesizer performs similar to the categorical CART synthesizer in terms of validity and arguably slightly worse in terms of disclosure risk the former should be preferred in practice. Especially, since treating the geocode as categorical will ensure that no implausible geocodes are generated.

We note that the number of variables included in the dataset is very small. It is likely that the quality of the CART synthesizer would only improve if more variables were available that could be used when building the tree. Comparing the two approaches for a larger set of variables is an interesting area for future research.

#### REFERENCES

- [1] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Lecture Notes in Statistics 201, New York: Springer (2011).
- [2] J.P. Reiter, Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* **21** (2005), 441-462.
- [3] H. Wang and J.P. Reiter, Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics* **6** (2012), 229–252.
- [4] P. Jacobebbinghaus and S. Seth, Linked-Employer-Employee-Daten des IAB: LIAB-Querschnittmodell 2, 1993-2008. Tech.rep., FDZ-Datenreport, No.5 (2010).
- [5] J.P. Reiter and R. Mitra, Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1** (2009), 99-110.

# Automated methods for providing bespoke synthetic data for the UK Longitudinal Studies

Beata Nowok ([beata.nowok@ed.ac.uk](mailto:beata.nowok@ed.ac.uk))<sup>1</sup>, Gillian M. Raab<sup>1</sup> and Chris Dibben<sup>1</sup>

**Keywords:** synthetic data, confidentiality, statistical disclosure control, CART, UK Longitudinal Studies

## 1. INTRODUCTION

Synthetic data methods were designed to address the conflicting demands placed on data holders to unlock the research and policy potential of microdata while at the same time preserving the confidentiality of individuals [1]. These methods, detailed in a monograph [2], have been recently recognized by the UK National Statistical Agencies (NSAs) which plan to use bespoke synthetic data to expand the use of the UK Longitudinal Studies (LSs). The generation of useful synthetic data often involves, however, a substantial investment of research time. Automated synthesising methods are therefore essential when the process has to be conducted rapidly and on a regular basis. This paper describes the application of synthetic data to the LSs, presents method implemented in an **R** [3] package *synthpop*<sup>2</sup> for producing non-disclosive entirely synthetic data and evaluates automated synthesising approaches.

### 1.1. The UK Longitudinal Studies

The England and Wales Longitudinal Study (ONS LS) [4], the Scottish Longitudinal Study (SLS) [5] and the Northern Ireland Longitudinal Study (NILS) [6] are rich microdata sets linking samples from the national Census in each country to administrative data (births, deaths, marriages, cancer registrations and other sources) for individuals and their immediate families across several decades. The sensitive nature of the information they contain, and the legal restrictions that apply to Census data, mean that access to the microdata is restricted to approved researchers and LSs support staff, who can only view and work with the data in safe settings controlled by the NSAs. The fairly restrictive access regime has a detrimental impact on usage and limits potential impact of the three LSs.

### 1.2. Application of synthetic data

Synthetic data with no real individuals, but which mimic the original observed data and preserve the relationships between variables and transitions of individuals over time could be made available to accredited researchers to analyse on their own computers. As no user of the LSs has access to all of the variables, researchers would be provided with a synthetic version of an extract with just the data they require for the population relevant to their research. Synthetic data would be created for each extract separately and would thus be project-specific. They need to resemble the actual data as closely as possible, but would never be used in any final analyses. The users carry out exploratory analyses and test models on the synthetic data, but they or LSs support staff use the code developed on the synthetic data to run their final analyses on the original data. This approach recognises the limitations of synthetic data produced by these methods. A similar

---

<sup>1</sup> Administrative Data Research Centre - Scotland (ADRC-S), University of Edinburgh

<sup>2</sup> see <http://cran.r-project.org/package=synthpop>

approach is currently being used for synthetic products made available by the U.S. Census Bureau<sup>3</sup>.

## 2. METHOD AND SOFTWARE

The *synthpop* package for **R** has been developed as part of the ESRC funded SYLLS project (Synthetic Data Estimation for UK Longitudinal Studies) to facilitate production of synthetic versions of LS data extracts requested by users. Via the function *syn()* synthetic data are produced using a single command and to run default synthesis only the data to be synthesised have to be provided as a function argument, e.g. *syn(data)*. The package offers a variety of options to customize synthesis [7-8] which can be used to influence the disclosure risk and the utility of the synthesised data. The essential features of the synthesis procedure remain, however, unchanged. Variables are synthesised one-by-one using sequential regression modelling. It means that conditional distributions, from which synthetic values are drawn, are defined for each variable separately. Note that the fitted regression models are conditioned on the original variables that are earlier in the synthesis sequence. Similar conditional specification approaches are used in most implementations of synthetic data generation. They are preferred to joint modelling not only because of the ease of implementation but also because of their flexibility to apply methods that take into account structural features of the data such as logical constraints or missing data patterns. To reproduce any such restrictions in the synthesised data they have to be specified in optional parameters of the *syn()* function.

With practicality and flexibility in mind, classification and regression trees (CART) are used as the default conditional models for synthesis but various parametric alternatives are also available. CART [9] are an algorithmic modelling approach that can be applied to any type of data. The basic idea is to recursively split a data set into groups with increasingly homogeneous outcome. The splits are specified as yes-no questions referring to the predictor space. The values in each final group approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. The synthetic values are generated by sampling from an appropriate group. CART models were suggested for generation of synthetic data by Reiter [10] and then evaluated as performing well by Drechsler and Reiter [11]. The key advantage of CART models is the ability to capture, in an automatic manner, non-linear relationships and interaction effects in the data that can be difficult to model using parametric approach.

## 3. EXAMPLE

The characteristics and quality of synthetic data depend on the models used to generate them. Specifying appropriate models that capture all essential features of the original data is therefore crucial but it requires expertise in both the data to be synthesised and statistical methods. Moreover, it can be cumbersome (if at all possible) and pose a major obstacle for data custodians with limited resources. With this in mind, we use the *synthpop* package and its *syn()* function to synthesise data with default settings for CART and parametric models and we evaluate and compare the quality of the results. In general, we aim to reproduce the logical structure of the data, univariate distributions and multivariate relations among the variables so that an analysis based on the synthetic data

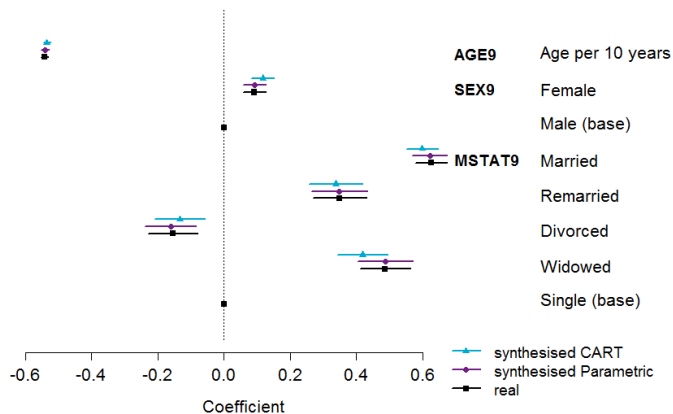
---

<sup>3</sup> see <http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html> and <https://www.census.gov/ces/dataproducts/synlbd/>

leads to the similar statistical inference as an analysis based on the observed data. We focus here on the final aspect and compare model estimates with and without interaction.

We have extracted data on age, sex, marital status and long-term illness from the SLS database for the 1991 Census and the acronyms AGE9, SEX9, MSTAT9 and ILL9 are used to describe them. Multiple synthetic data sets were produced for both non-parametric (CART) and parametric (Parametric) synthesis. In the latter case default polychotomous or logistic regression was used depending on variable type. In terms of exploratory analyses (details not shown here) the CART method gave more satisfactory results than parametric methods, e.g. it was much better at retaining a constraint that was obeyed by the original observed data (marital status ‘single’ for under 16s) when it was not forced by the customized parameters of the *syn()* function. Note that before regression analysis data syntheses were rerun with this constraint imposed.

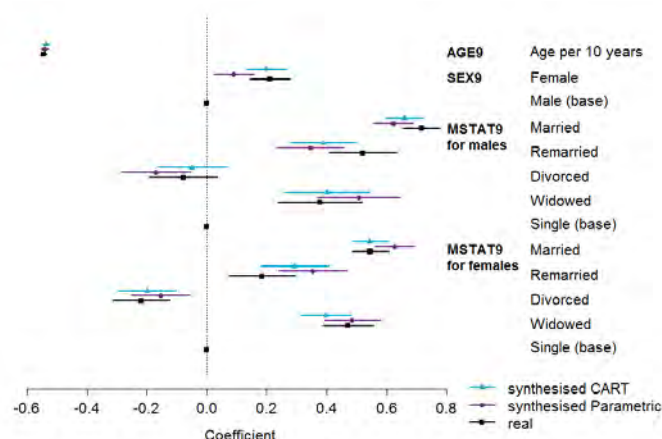
Optimally an analysis based on the synthesised data should lead to the same statistical inferences as an analysis based on the observed data. For illustration we investigate here results of a logistic regression to absence of long-term illness in 1991 as a model from age, marital status and sex, a model that was part of the parametric syntheses. Figure 1 compares the estimates from the real data with the averages from the 50 syntheses from Parametric and CART models respectively. The estimates from the Parametric models are very close to the real estimates but the estimates from the CART syntheses are also similar to the real estimates and the same conclusions are drawn in all three cases. We can see that freedom from long-term illness decreases sharply with age and is higher for females than males. Adjusting for age and sex, those married, remarried or widowed are more likely to be free from long-term illness than those who are single, whereas the opposite is true for the divorced.



**Figure 1. Coefficients of fit to ILL9=“No” from AGE9, SEX9 and MSTAT9 for real and synthetic data.**

Despite very good performance of Parametric methods in the above example where a model of interest was part of the parametric syntheses the methods possess some major disadvantages. They preclude an analyst, with access only to the synthetic data, from checking departures from an assumed model, such as lack of linearity or the absence of interactions. This is illustrated here by fitting a further model which includes a sex by marital status interaction. Results are shown in Figure 2. For the observed data there is evidence of an interaction. The association of being married with lack of illness is stronger for men than for women. The CART syntheses do a reasonable job of reproducing this, whereas the Parametric syntheses show no evidence of this interaction since they are generated from an interaction-free model. As illustrated, with the CART

method it is possible to capture various features of the underlying original data without any tuning of the synthesis. The CART method would be therefore preferred over the parametric methods.



**Figure 2. Coefficients of fit to ILL9="No" from AGE9, SEX9 and MSTAT9\*SEX9 interaction for real and synthetic data.**

#### 4. DISCUSSION

Synthetic data offer a way to expand the use of confidential microdata such as the UK LSs. The *synthpop* package for **R** with its default CART method has been developed to facilitate generation of such data. It is not, however, a final product and feedback from package and synthetic data users is absolutely invaluable for further improvements and development of best practices. Next to facing the challenge of producing synthetic data, we also need to address the concerns of data custodians responsible for protecting confidentiality. Finally, weaknesses and limitations of synthetic data have to be clearly communicated.

#### REFERENCES

- [1] D.B. Rubin, Discussion: Statistical disclosure limitation, *Journal of Official Statistics* 9(2) (1993), 461-468.
- [2] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control*, Vol. 201, *Lecture Notes in Statistics*, New York: Springer (2011).
- [3] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria (2014).
- [4] L. Hattersley and R. Cresser, *The Longitudinal Study, 1971-1991: History, organisation and quality of data*, LS Series no.7, The Stationery Office, London (1995).
- [5] P. Boyle, P. Feijten, Z. Feng, L. Hattersley, Z. Huang, J. Nolan and G.M. Raab, Cohort profile: The Scottish Longitudinal Study (SLS), *International Journal of Epidemiology* 38(2) (2009), 385-392.
- [6] D. O'Reilly, M. Rosato, G. Catney, F. Johnston and M. Brolly, Cohort description: The Northern Ireland Longitudinal Study (NILS), *International Journal of Epidemiology* 41(3) (2012), 634-641.
- [7] B. Nowok, G.M. Raab and C. Dibben, *synthpop: Bespoke creation of synthetic data in R*, *submitted* (2014), <http://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf>.
- [8] G.M. Raab, B. Nowok and C. Dibben, Simplifying synthesis with the *synthpop* package for R, Paper presented at the Privacy in Statistical Databases 2014 Conference (2014).
- [9] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth (1984).
- [10] J.P. Reiter, Using CART to generate partially synthetic, public use microdata, *Journal of Official Statistics* 21 (2005), 441-462.
- [11] J. Drechsler and J.P. Reiter, An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Computational Statistics and Data Analysis* 55 (12) (2011), 3232-3243.

# **A MORE VISUAL DISSEMINATION FOR ATTRACTING NEW USERS**

**PH. BAUTIER<sup>1</sup>, B. LE GOFF<sup>2</sup>**

**Keywords:** visual dissemination, user behaviour, user needs, user feedback

## **1. Introduction**

Today, each national statistical office is confronted with the same challenge: how to build a dissemination and communication strategy in a world where users have easy access to a deluge of data and information from various origins and where IT tools and design standards change so quickly that users behaviour and their expectations are continuously modified?

The first issue is clearly to know what users want: we know our different types of users (decision makers, media, researchers, businesses, students, public at large...) but we have to identify how they get our data, what they do with our data, how they react to our outputs and which sort of new services they would like us to propose. In our changing world, this information cannot be obtained only through an annual user survey, but requires continuous and "real time" feedback from our users.

The second issue is to develop a range of dissemination products or services which allows to reply to the different needs identified. The objective of this paper is to present Eurostat's experience in proposing new and more visual dissemination tools which aim at making our statistics more understandable to less experienced users and also attracting new users of European statistics.

## **2. Finding our way in the labyrinth of user needs**

Since a few years, Eurostat has been developing a number of different and complementary tools which give an interesting and up-to date representation of our user needs. Each of them helps to assemble a more global picture of what modern users expect from suppliers of statistical data.

### ***2.1. Measuring satisfaction***

To get an overview of the general level of satisfaction of users, Eurostat conducts an annual on-line user satisfaction survey. This classical method still provides valuable information and feedback on the most consulted statistical domains, the purpose and the frequency of the consultation, as well as an assessment of the quality of our data, publications, and dissemination practices. It also gives a crucial information on the level of confidence that users have in our data. In the 2014 satisfaction survey, trust remained overwhelmingly positive with 95% of the respondents stating they greatly trust European statistics or tend to trust them.

---

<sup>1</sup> Eurostat, Commission européenne, philippe.bautier@ec.europa.eu

<sup>2</sup> Eurostat, Commission européenne, bernard.le-goff@ec.europa.eu



## ***2.2. Detecting user behaviour***

Website log files provide a wealth of information which is exploited through a detailed and extensive web analytics effort. Each month a 30 page monitoring report on Eurostat electronic dissemination is published on the intranet, which contains a long list of information such as the number of consultations for each page, the navigation and origin of the consultation (Eurostat website, Google, apps,...) or the average time spent on each visualisation tool.

## ***2.3. Getting feedback in real time***

Successful dissemination cannot be measured by means of web analytics and usage figures alone, but it needs to take into account new ways of information. For instance, the monitoring of social media brings further insight into who is using our information, how they use it, what they say and think about it and how Eurostat is perceived on the internet in general. To measure the impact of its dissemination, Eurostat uses a tool to analyse its e-reputation in real time. The tool provides a better knowledge of our users and of our daily impact in the media, social networks or blogs and gives a quantitative but also qualitative feedback on our work.

## ***2.4. Communicating with users***

Apart from measuring usage, Eurostat also communicates with users via a permanent user support network, ad-hoc focus groups and benchmarking exercises. For ten years, Eurostat has managed a system of national user support centres offering assistance in nearly all EU languages. The valuable feedback collected via this permanent structure enables Eurostat to identify concrete user requirements and helps us to improve the quality of our services.

Eurostat organised ad-hoc internal and external focus groups to allow an exchange of views on the current website's strengths and weaknesses.

The benchmarking exercises measure the overall quality of the Eurostat website against current best practices and in particular against the websites of other statistical institutes and/or international organisations.

## **3. Translating the needs for attracting new visitors**

One of the most frequent remarks made by different groups of users when we consulted them in the preparatory phase of the new website, was that Eurostat's website was judged as too complex for non-specialists, which is the case for most of European citizens. In order to attract a larger public of less experienced users, we decided to strengthen the user-orientation of our website and to propose a set of new visualisation tools, infographics and apps that are meant to be informative and easy to use.

### ***3.1 A more attractive website, less statistical jargon and a more powerful search***

The launch of the new Eurostat website mid-December 2014 has been perceived as a good opportunity to better reply to user needs (see <http://www.ec.europa.eu/eurostat>). The layout and the design of the web site have undergone a major overhaul to make it more appealing and attractive, with for example, a more colourful design, the possibility to insert photos or videos and a daily management of the editorial content of the homepage to make it more lively.

A new search engine has also been developed which provides, on the basis of keywords, the most relevant datasets and articles/publications available, in a similar way to how Google

works. To facilitate the search, bridges have been created to enlarge the search terms written in common vocabulary (such as profits, apartment or family for example) to the associated statistical terminology (gross operating surplus, dwelling or household).

### ***3.2 Simple infographics and visualisation tools***

Data visualisation tools are another possibility to help users to better understand our statistics. Their aim is to communicate clear information or a story through graphs, maps or charts. In recent years, several tools have been implemented by Eurostat, such as country profiles, inflation dashboard, statistical atlas, regional statistics illustrated and widgets. However, the use of these tools sometimes requires the user to have already a good understanding of statistics.

For that reason, Eurostat decided to complement its offer by presenting regular infographics on the homepage of its new website, in order to arouse the interest and provide assistance to less experienced users.

#### ***"Economic Trends"***

A new infographic is associated with the publication of a selection of euro-indicator news releases, where non-specialists can get a better understanding of the most recent economic trends in the EU, the euro-area and the Member States:  
<http://ec.europa.eu/eurostat/cache/infographs/economy/desktop/index.html>

#### ***"Young Europeans"***

"Young Europeans" is a new tool released in connection with a new Eurostat publication on youth. It provides the possibility to compare the way of living of a young people aged 15-29 with those of any other young Europeans of the same age and sex. This tool is also intended for parents, decision-makers, politicians or teachers who want to know more about the young generation in Europe.

"Young Europeans" is an interactive tool constructed around a number of questions about the life of young Europeans on 4 different themes: family, work, free time and studies, and internet. Before starting, users have to define their profile: gender, country and age.



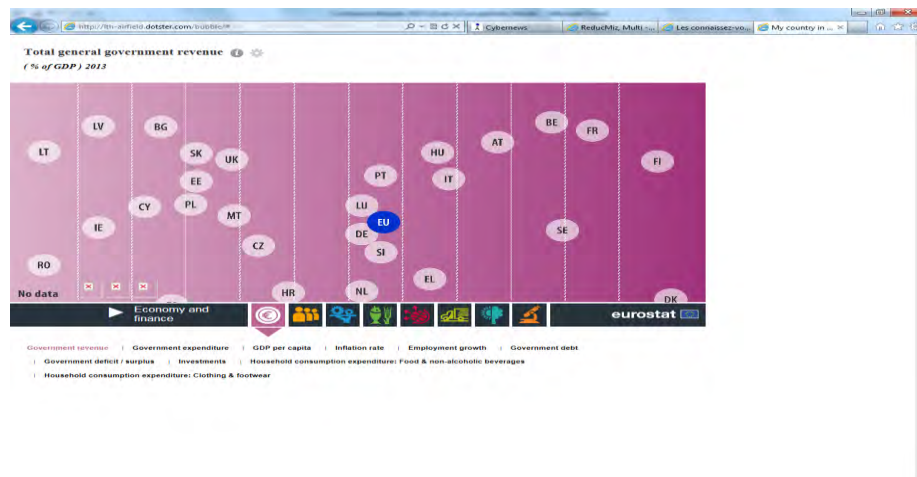
## ***"Quality of life"***

Linked to the release of a Eurostat publication on quality of life, this infograph shows both objective and subjective indicators covering 9 quality of life themes. It proposes a combination of photos and graphics to display the information in an attractive and innovative way. A new easy recognizable logo for quality of life statistics has also been created.



## ***"My country in a bubble"***

This simple visualization tool allows users to see in one image the situation in Europe for more than 140 statistical indicators covering all economic, social and environmental domains. This tool is not done to provide some precise numbers to users but to allow them to immediately perceive the situation of a given country compared with other EU countries and to encourage them to discover more.



## **4. Conclusion**

All these actions are part of Eurostat's efforts to better respond to user needs and to continuously adapt its visualization dissemination with the latest IT development. Today, statistical institutes are confronted with the same challenge. In a period where human and budgetary resources are limited, this challenge can only be faced if a reinforced cooperation among ESS members is put in place in particular in sharing best practices and tools.

# Targeted designs in surveys and longitudinal studies

Annamaria Bianchi (annamaria.bianchi@unibg.it)<sup>1</sup>, Silvia Biffignandi (silvia.biffignandi@unibg.it)<sup>1</sup>

**Keywords:** non-response, representativity, response rate

## 1. INTRODUCTION

The concept of representativity is rather complex and refers to the aim that sample data gain external validity in relationship to the target population they are meant to represent. The definition of representativity is even more complex in the longitudinal context, where one wants to make inference both with respect to the initial population (longitudinal estimates) and the population at different time points (cross-sectional estimates). The definition depends here on the target population for a specific inference.

To improve the representativity of surveys, recently, researchers have explored the idea of treating sample subgroups differently and intervening in the data collection process in order to achieve a “better” response set. Different terms are used to identify these methods, such as targeted design, responsive design, adaptive design, tailored design (Grooves and Heeringa, 2006; Wagner, 2008). They refer to slightly different versions of the method.

This technique is important both for cross-sectional and longitudinal studies, because it allows to optimize data quality, costs, and time during data collection and it could be implemented in real time using data from administrative records and also variables generated during the collection process (like server side paradata).

Experiments have been performed by National Statistical Institutes (NSIs), mainly in the cross-sectional context (Luiten and Schouten, 2013; Lundquist and Särndal, 2013; Shlomo et al., 2013). This is currently a topic of research at NSIs. The method proved to be useful. However, it is not standard practice yet.

In longitudinal studies, targeted designs could take advantage of the wealth of information on panel members, collected during recruitment and over time. This information could be exploited to improve the data collection process. Such strategies are not used in longitudinal surveys on a standard basis as well. A few experiments have been performed (e.g., Lynn, 2014).

In this paper, we analyse at first main requirements for the application of targeted survey designs in general and in the specific context of longitudinal studies, giving a theoretical systematization of what can be done and the necessary information. Afterword, we provide an overview of research going on in the official statistics, mostly in the cross-section context. Finally, we present some experiments we made in the panel context and discuss potentialities and problems of the approach.

## 2. METHODS

The application of targeted designs includes several components:

1. Definition of subgroups to be targeted, depending on covariates and /or available paradata;

---

<sup>1</sup> DSAEMQ, University of Bergamo, via dei Caniana 2, 24127 Bergamo, Italy

2. Identification of the survey strategies, that is, the survey design features to be applied to different units;
3. Allocation of sample units to subgroups.

The application of such designs take advantage of the use of indicators of the quality of the response (such as response rates, subgroup response rates, balance indicators (Särndal, 2011), and R-indicators (Schouten et al., 2009; 2011)) and cost functions.

Depending on a number of features, several forms of this technique may be identified in the literature. We discuss them, highlighting pros and cons for the application in cross-sectional and longitudinal studies and providing some new evidence. We also discuss different measures of the quality of response, with special reference to the longitudinal context and taking into account that response rates are in some sense more important for longitudinal surveys than for other surveys, as the utility of a longitudinal survey relies on the continued participation of a high proportion of the sample members (Lynn, 2014). Indeed, in order to remain representative of their target population and to support longitudinal analysis, it is important for longitudinal studies to maintain high response rates over time.

### 3. APPLICATION

The application of targeted designs is still at the beginning and it is not standard practice. We briefly revise applications of this technique, also at NSIs, and then focus on applications in the framework of longitudinal studies.

Longitudinal surveys provide researchers with a large number of auxiliary variables that could be used to identify alternative subgroups with different responding behaviours, at which different strategies may be targeted. Further, longitudinal studies offer more features for manipulation than cross-sectional studies, e.g. between-wave interventions. Hence, the application of targeted designs in the longitudinal context is envisaged and is of interest for Official Statistics since in this context one can gain deeper insights thanks to the large availability of information on panel members. Conclusions in this context may be useful in supporting ideas, experiments and decisions in Official Statistics.

Up to now only a few examples and experiments have been performed. The authors are aware of the studies by Lynn (2014) on Understanding Society, Calderwood et al. (2012) on the UK Millennium Cohort Study, and Bianchi and Biffignandi (2014) on the Paadel panel.

Examples of some in-depth analyzes are reviewed with reference to the different panels. The results are promising. Table 1 shows, as an example, some of the results that will be described in the study. A new application is also presented, which demonstrates the usefulness of the method. The application makes reference to the Understanding Society Innovation Panel (Uhrig, 2011). Understanding Society is a longitudinal household panel in the UK, which is so much related to issues of interest for Official Statistics since it provides an accepted and trusted set of National Statistics which helps people understand and monitor well-being.

**Table 1. Comparison of the estimates obtained applying different types of responsive design for the variable firm legal status in the Paadel panel**

Data set	Average Relative Difference
Full data set	0.259
Responsive design 1	0.200
Responsive design 2	0.182

#### **4. CONCLUSIONS**

Targeted survey designs seem to be very promising both in the context of cross-sectional and longitudinal studies. They allow to monitor representativeness and optimize fieldwork efforts with the saving of resources. In the longitudinal context, they can take advantage of the richness of information that these datasets provide.

Not much guidance is given in the literature on the application of such methods. More experimentations are needed to increase knowledge in this field of application.

This paper, after reviewing both theory and practice, provides a contribution in the direction of setting up guidances. A new experiment in the longitudinal context is presented as well.

#### **REFERENCES**

- [1] R.M. Groves and S.G. Heeringa, Responsive design for household surveys: tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society: Series A* 169 (2006), 439-457.
- [2] J. Wagner, Adaptive survey design to reduce non-response bias. *PhD Thesis*, University of Michigan (2008).
- [3] A. Luiten and B. Schouten, Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction, *J.R. Statist. Soc. A* 176 (2013), 169-189.
- [4] P. Lundquist, and C.E. Särndal, Aspects of responsive design with applications to the Swedish Living Conditions Survey, *J. Off. Stat.* 29 (2013), 557-582.
- [5] P. Lynn, Targeted response inducement strategies on longitudinal surveys, in *Improving Survey Methods: Lessons from Recent Research* (ed.s U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis), Routledge/ Psychology Press, 2014.
- [6] C.E. Särndal, The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection, in *Estimation*, *J. Off. Stat.* 27 (2011), 1-21.
- [7] N. Shlomo, B. Schouten and V. de Heij, Designing Adaptive Survey Designs with R-Indicators, paper presented at the *New Techniques and Technologies for Statistics Conference 2013*, Brussels.
- [8] B. Schouten, F. Cobben and J. Bethlehem, Indicators for the representativeness of survey response. *Surv. Methodol.* 35 (2009), 101-113.

- [9] B. Schouten, N. Shlomo and C. Skinner, Indicators for Monitoring and Improving Representativeness of Response, *J. Off. Stat.* 27 (2011), 231-253.
- [10] L. Calderwood, A. Cleary, G. Flore and R.D. Wiggins, Using response propensity models to inform fieldwork practice on the 5<sup>th</sup> wave of the Millenium Cohort Study, paper presented at the International Panel Survey Methods Workshop Melbourne, Australia, July 2012.
- [11] A. Bianchi and S. Biffignandi, Responsive design for economic data in mixed-mode panels, in *Contribution to Sampling Statistics*, eds. Mecatti, F., Conti P.L., and Ranalli, M.G., Springer, 2014, 85-102.
- [12] S.C.N. Uhrig, Using experiments to guide decision making in Understanding Society: Introducing the Innovation Panel, chapter 13 in S. L. McFall & C. Garrington (ed.s), *Understanding Society: Early Findings from the First Wave of the UK's Household Longitudinal Study*, Colchester: University of Essex, 2011. At: <http://research.understandingsociety.org.uk/findings/early-findings>

# Improving the response rates in business surveys

## The case of LCS 2012

Ciro Baldi ([baldi@istat.it](mailto:baldi@istat.it))<sup>1</sup>, Marilena A. Ciarallo<sup>1</sup>, Stefano De Santis<sup>1</sup>, Rossana Renzi<sup>1</sup>, Graziella Spera<sup>1</sup>

**Keywords:** Response rates, Business survey, Contact strategy, Data collection

### 1. INTRODUCTION

The Labour Cost Survey (LCS) is a 4-yearly business survey, based on Regulation (EC) N. 503/1999 that aimed at producing harmonized estimates on a number of variables concerning employment, the number of hours worked and paid and a very detailed structure of labour costs of enterprises and institutions both in public and in private sector.

All the structural business surveys are plagued in Italy by quite low response rates, frequently below 50% and among them, LCS was no exception. Moreover, in the general context of budget cuts, retired staff between 2008 and 2012 has not been completely substituted and the number of human resources dedicated to the survey has been reduced.

To maintain the data quality in the 2012 edition and avoid a further depletion of response rates (which, in addition to worsen the sampling error, carries the risk of higher non response bias) it has been implemented a work reorganization and introduced several innovations especially in the data collection. The final result of these innovations has been a considerable increase in the response rate.

### 2. METHODS

The innovations introduced to increase the response rates pertain three areas: the questionnaire redesign, the contact strategy and the data collection strategy.

#### 2.1. The questionnaire redesign

The LCS regulation requires a considerable number of items and sub-items on quantitative variables at the enterprise level. Moreover, in addition to the variables needed to respond the Regulation, the past editions of the survey questionnaire carried questions needed to gain insights on how wages were differentiated by occupation. The result was one of the most burdensome questionnaire in the field of economic statistics. In recent times a new source of administrative data, the individual level social security declarations, have been made available to Istat. Although not adequate to satisfy the EU Regulation on LCS this availability has triggered the possibility of modifying the survey questionnaire, eliminating part of the questions. Moreover the remaining questions have been restructured making explicit reference to the concepts and the measures to which the enterprises are more used to, that is those used to perform the contained in the payroll system, or needed to satisfy the obligations on social security. See [1] for more details on this issue.

As some studies have shown (see e.g. [2]) the simplification of the questionnaire has likely contributed to increase the response rate by easing the burden of the enterprises.

---

<sup>1</sup> Italian National Institute of Statistics (ISTAT)



## 2.2. The contact strategy

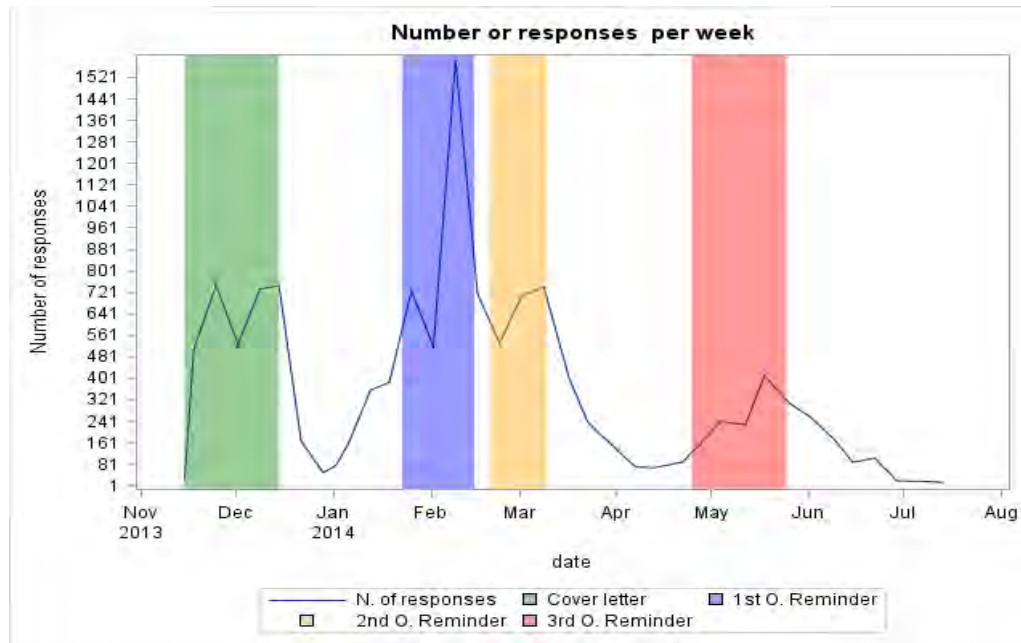
The contact strategy has been completely changed and innovated using three main tools. The first tool has been the use of Certified Emails (here after PEC) system. A recent change in legislation has imposed that the official communications between public administrations and enterprises should be carried out only through electronic systems and not any longer through postal service. In this context the PEC technology allows to send and receive official communications with safe protocols and certainty of delivery, with the same legal value of letters sent by recorded delivery and receipt return. Virtually all the official communications towards the enterprises (the cover letter aimed at informing the enterprises of the survey and three subsequent reminders) have been sent through this system. The second pillar of the contact strategy has been the use of a dedicated contact service to perform part of the front office operations. The main task performed was to receive the incoming calls from the enterprises in order to answer common answers on deadlines, obligations, and helping the enterprises to solve the most common issues regarding registration on the web site or to understand the errors of the consistency checks built in the electronic questionnaire.... In the past editions of the survey these tasks were performed internally with a large portion of the staff being occupied in these activities. With the reduction of the resources dedicated to the survey the old strategy was no longer feasible so that a portion of the budget savings has been used to contract out these services. The second task performed by the contractor were outgoing calls, after the cover letter, aimed at gaining survey cooperation, and to identify the person, within the enterprise, capable of responding the survey and her direct contact references (telephone, mail, etc..). These persons were subsequently contacted through mass-emailing (ordinary email, not PEC) to solicit the sending of the questionnaire in more informal reminders. Thanks to the activities performed by the external agency, the internal staff work (the third pillar) could concentrate in answering the emails (those with more official communications sent through PEC and the rest through ordinary emails) and answering the telephone calls on more technical issues or on particular cases. In paragraph 3 some measures of the pervasiveness of these contact operations are reported.

## 2.3. The data collection strategy

To increase the response rate the data collection strategy has used several modes, to reach enterprises with different preferences/necessities. The main mode is a *web questionnaire* with built in checks. In the first stage of the survey, together with it, a *standard file* was offered to enterprise groups to provide data for the all units of the group at once. In the ending stages of data collection a *CATI* was performed trying to collect data through direct phone interviews in order to get the responses of hard-to-convince enterprises. Furthermore, it was offered to small enterprises the possibility to fill an *offline electronic questionnaire* (developed with a pdf Form and Javascript technology), sent by PEC and identical to the web one. Since multi-mode data collection methods, while effective in reducing nonresponses, may involve measurement differences between modes, the data collected either through the standard file or through the offline questionnaire after being checked for formal errors and loaded into the data base were further screened by the reference persons of the respondents. In the case of standard file the reference person for each group was asked to correct/integrate the data on the web form, while in the case of PDF forms, the CATI service recalled the enterprises to solve inconsistencies and errors. In the past editions, while the major channel was a structured datasheet file, there was a large portion of the questionnaires sent by fax which needed to be printed out and sent to an external data entry service. The method was either time and money expensive or prone to more errors.

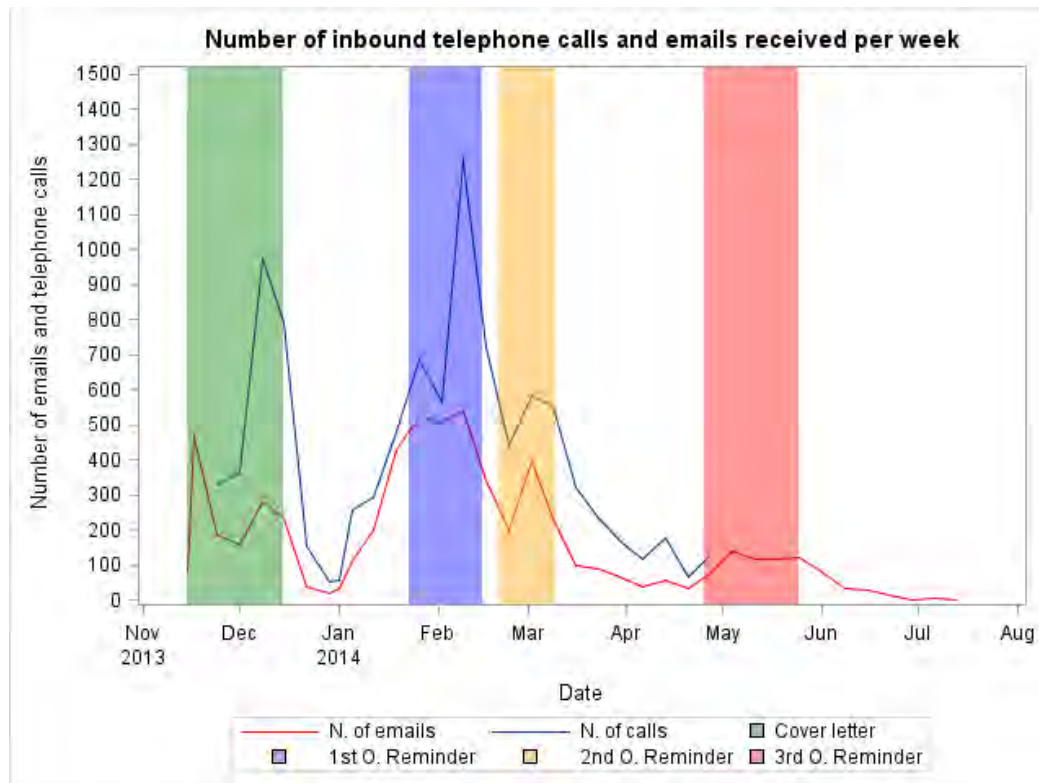
### 3. RESULTS

The strategy of several reminders is apparent in figure 1. The shaded area goes from the sending of the cover letter/reminder to the fixed deadline. One can see how the number of responses increased steadily in proximity of the deadlines. A second feature notable from the figure is the necessity of soliciting the enterprises a number of times in order to get the answer. The maximum number of questionnaires was collected with the first reminder, but also the second and the third have had a considerable effect. The very low cost of the PEC system together with the relative easiness of the procedure, compared to the costs and complexity of the postal sending used in the past, has allowed a greater degree of flexibility in the course of the data collection in deciding the number and the timing of the reminders.



**Figure 1. Number of responses per week**

The impact of the second pillar of the data collection, that is a dedicated contact center, is suggested in figure 2. The total number of incoming telephone calls received by the contact center has been just below 10000, in the period in which the service was active, with peaks (up to 1000 per week and almost 400 per day) in the days soon after the sending of a reminder or close to the deadlines. Given resource constraints, in the past the internal staff was not able to manage such a workload of calls with the results that possibly a large quantity of calls went unprocessed and the likely outcome of lowering the response rate. The outsourcing of the telephone activities has in turn allowed to concentrate the reduced number of internal personnel on the processing of the email. The chart shows that the internal staff had to manage peaks of up to 500 emails (ordinary and PEC) per week.



**Figure 2. Number of incoming calls and emails received per week**

The total number of emails received was over 6000 of which about 70% has required an answer (the residual was represented by notices from enterprises to Istat not requiring answering) (table 1).

**Table 1. Number of emails managed by internal staff**

	Received	Answered	Total
Ordinary	4792	3779	8571
PEC	1290	627	1917
Total	6082	4406	10488

#### 4. CONCLUSIONS

The paper has illustrated the complex strategy set up to improve the response rate in the LCS 2012, notwithstanding budget cuts and staff reduction. The role of contact strategy, with the big innovations of PEC mailing that has allowed a more flexible and intensive use of the reminders and an organization ready to respond to large quantities of incoming requests from the enterprises, and a pervasive multi-mode collection strategy are the drivers of the steep increase in survey response rate to 69%.

#### REFERENCES

- [1] C. Baldi, M.A. Ciarallo, S. De Santis, S. Pacini, The converging pattern between Business statistics and Administrative data. Towards an “industrialized” statistical production process, Paper presented at the Conference Q2014, Wien June 2014 (2014).
- [2] G. D. White Jr. and A. Luo Business, Survey Response Rates – Can They Be Improved? ASA Section on Survey Research Methods (2005)

# Imputation under edit restrictions and known totals

Ton de Waal ([t.dewaal@cbs.nl](mailto:t.dewaal@cbs.nl))<sup>1</sup>, Wieger Coutinho<sup>2</sup> and Natalie Shlomo<sup>3</sup>

**Keywords:** Imputation, edit restrictions, calibration

## 1. INTRODUCTION

Missing data form a well-known problem that has to be faced by basically everyone who collects survey data. Missing data can arise due to two kinds of non-response mechanisms: unit non-response and item-nonresponse. Unit non-response occurs, for instance, when units that are selected for data collection cannot be contacted or refuse to respond altogether. Unit non-response is usually corrected by weighting the responding units. Item non-response occurs when data on only some of the items in a record, i.e. the data of an individual respondent, are missing. The most common solution to handle item non-response is imputation, where missing values are estimated and filled in. In this paper we focus on item non-response for numerical data, and whenever we refer to missing data in this paper we will mean missing data due to item non-response.

In many cases data have to satisfy constraints in the form of edit restrictions, or edits for short. For numerical data such edits are given by

$$a_{1k}x_{i1} + \dots + a_{pk}x_{ip} + b_k = 0 \quad (1a)$$

or

$$a_{1k}x_{i1} + \dots + a_{pk}x_{ip} + b_k \geq 0. \quad (1b)$$

Here the  $x_{ij}$  denote the value of variable  $j$  ( $j = 1, \dots, p$ ) in record  $i$  ( $i = 1, \dots, r$ ), and the  $a_{jk}$  and the  $b_k$  are certain constants, which define edit  $k$  ( $k = 1, \dots, K$ ). Records that do not satisfy edits are inconsistent, and are hence considered incorrect. Numerical data sometimes also have to sum up to known totals, for instance because these totals are known or have already been estimated from other sources. This leads to sum constraints of the form  $\sum_{i=1}^R w_i x_{ij} = X_j^{pop}$ , with  $X_j^{pop}$  the known total and  $w_i$  the survey weight of record  $i$ .

In [1] imputation methods have been developed that ensure that edits are satisfied and at the same time known totals are preserved. A drawback of these methods is that they are quite complex and hard to implement. In this paper we propose a simpler method that achieves the same goals. A novel element of our method is that unequal survey weights can be taken into account. The rest of this paper is organized as follows. Section 2 sketches our new imputation method. Section 3 describes some results. Section 4 concludes with a brief discussion.

## 2. METHODS

Our proposed imputation method is based on a hot deck approach. When hot deck imputation is used, for each record containing missing values, the so-called recipient record, one uses the values of one or more other records where these values are observed, the so-called donor record(s), to impute the missing values. Usually, hot deck imputation is applied multivariately, i.e. several missing values in a recipient record are imputed simultaneously, using the same donor record. For our situation, where we have edits and

---

<sup>1</sup> Statistics Netherlands and Tilburg University

<sup>2</sup> Loket Aangepast-Lezen

<sup>3</sup> University of Manchester

known totals, that approach is often not feasible as it is unlikely that a single donor record will lead to satisfied edits for the recipient record and to preserved totals. We therefore apply sequential univariate hot deck imputation, where for each missing value in a record in principle a different donor record may be selected. The variables with missing values are imputed sequentially. The univariate hot deck imputation method is used to order the potential imputation values for a certain missing field. Only values of donor records that can result in a record that satisfies all edits and preserves the total of the variable under consideration may be used for imputation. How we ensure that edits and sum constraints can be satisfied is explained below.

## 2.1. Using a Sequential Approach

In order to be able to use a sequential approach, we apply Fourier-Motzkin (FM) elimination (see [2]). FM elimination projects a set of linear constraints involving  $q$  variables onto a set of linear constraints involving  $q - 1$  variables. The essence of FM elimination is that two constraints, say  $L(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq}) \leq x_{ij}$  and  $x_{ij} \leq U(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq})$ , where  $x_{ij}$  is the variable to be eliminated and  $L(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq})$  and  $U(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq})$  are linear expressions, lead to a constraint  $L(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq}) \leq U(x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq})$ .

The main property of FM elimination is that the original set of constraints involving  $q$  variables can be satisfied if and only if the corresponding projected set of constraints involving  $q-1$  variables can be satisfied. By repeated application of FM elimination we can derive an admissible interval for one of the values to be imputed. The main property of FM elimination guarantees that if we impute a value within this admissible interval, the other variables can be imputed such that all constraints are satisfied.

We illustrate our method by means of an example where we have  $r$  records with four variables  $T$  (turnover),  $P$  (profit),  $C$  (costs), and  $N$  (number of employees in fulltime equivalents). The edits for record  $i$  ( $i = 1, \dots, r$ ) are given by

$$T_i - C_i - P_i = 0 \quad (2)$$

$$P_i \leq 0.5T_i \quad (3)$$

$$-0.1T_i \leq P_i \quad (4)$$

$$T_i \leq 550N_i \quad (5)$$

$$T_i \geq 0 \quad (6)$$

$$N_i \geq 0 \quad (7)$$

$$C_i \geq 0 \quad (8)$$

We want to impute the variables in the following order:  $N$ ,  $T$ ,  $C$  and  $P$ . We assume that variable  $N$  has already been imputed in all records in which its value was missing, and we are now ready to impute variable  $T$ . Suppose that in a certain record  $i_0$  we have  $N = 5$  (either observed or imputed), and the values of  $T$ ,  $P$  and  $C$  are missing. We fill in the value for  $N$  into the edits and obtain (2), (3), (4), (6), (8) and

$$T_{i_0} \leq 2750. \quad (9)$$

In order to eliminate variable  $P$ , we use equation (2) to express  $P$  in terms of  $T$  and  $C$ . After elimination, we obtain constraints (6), (8), (9),

$$T_{i_0} - C_{i_0} \leq 0.5T_{i_0}, \quad (\text{equivalently: } 0.5T_{i_0} \leq C_{i_0}) \quad (10)$$

and

$$-0.1T_{i_0} \leq T_{i_0} - C_{i_0} \quad (\text{equivalently: } C_{i_0} \leq 1.1T_{i_0}) \quad (11)$$

To eliminate variable  $C$  from (6) and (8) to (11), we first copy the constraints not involving  $C$ , i.e. (6) and (9). Eliminating  $C$  from (8), (10) and (11) leads to constraints that are equivalent to (6). The admissible interval for  $T$  for record  $i_0$  is hence given by

$$0 \leq T_{i_0} \leq 2750.$$

In a similar way we can derive admissible intervals for  $T_i$  for all records  $i$  ( $i = 1, \dots, r$ ) for which the value of  $T_i$  is missing. After we have done this, we impute  $T_i$  in these records (see Sections 2.2 and 2.3). After variable  $T$  has been imputed in all records in which its value was missing, we can derive admissible intervals for variable  $C$ , and later variable  $P$ , in a similar manner. The main property of FM elimination guarantees that the original edits will be satisfied, if we select donor values lying in these admissible intervals.

## 2.2. Nearest-neighbour hot deck imputation

If we want to impute a certain variable  $x_j$  in a record  $i_0$ , we calculate for each other record  $i$  for which the value of  $x_j$  is not missing the distance to record  $i_0$ . Before we calculate this distance, we first scale the values by subtracting the median observed value and dividing by the interquartile distance. We then calculate the distance between records  $i$  and  $i_0$  by means of the Euclidean metric. Based on these distances we make an ordered list of potential donor values for variable  $x_j$  in record  $i_0$ .

## 2.3. The imputation algorithm

We first examine the case where all survey weights are equal. When we want to impute a missing value for the variable  $x_j$  under consideration in a certain record  $i_0$  we basically apply the following procedure.

0. Set  $t := 1$ .
1. Select the  $t$ -th value on the list of potential donor values for variable  $x_j$  in record  $i_0$ .
2. We check whether this value lies in the admissible interval for  $x_{i_0j}$ . If so, we continue with Step 3. Otherwise, we set  $t := t + 1$  and return to Step 1.
3. We check whether the potential donor value would enable us to preserve the total for variable  $x_j$ . If so, we use this potential donor value to impute the missing value. Otherwise, we set  $t := t + 1$  and return to Step 1.

We can efficiently combine the checks in Steps 2 and 3. The check in Step 2 is simply whether  $l_{i_0j} \leq x_{i_0j}^* \leq u_{i_0j}$ , where  $x_{i_0j}^*$  is the potential donor value drawn in Step 1,  $l_{i_0j}$  the lower bound of the admissible interval for variable  $x_j$  in record  $i_0$  and  $u_{i_0j}$  the corresponding upper bound (see Section 2.1). The check in Step 3 amounts to checking whether

$$\sum_{i > i_0} l_{ij} \leq X_{j,imp} - \sum_{i < i_0} \hat{x}_{ij} - x_{i_0j}^* \leq \sum_{i > i_0} u_{ij},$$

where  $M(j)$  is the set of records with missing values for variable  $x_j$ ,  $\hat{x}_{ij}$  ( $i \in M(j), i < i_0$ ) are the already imputed values, and  $X_{j,imp}$  is the total to be imputed for variable  $x_j$ . This total to be imputed equals the total  $X_j^{pop}$  minus the sum of the observed values for variable  $x_j$ . The checks in Steps 2 and 3 can be combined into a single check:

$$\max \left( X_{j,imp} - \sum_{\substack{i \in M(j) \\ i < i_0}} \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} u_{ij}, l_{i_0j} \right) \leq x_{i_0j}^* \leq \min \left( X_{j,imp} - \sum_{\substack{i \in M(j) \\ i < i_0}} \hat{x}_{ij} - \sum_{\substack{i \in M(j) \\ i > i_0}} l_{ij}, u_{i_0j} \right)$$

We can easily extend this check to the case of unequal sampling weights  $w_i$  for each record  $i$ .

### 3. RESULTS

For our evaluation study we have used two data sets. For each of these data sets we have a version with missing data and a corresponding complete version available. We have imputed the versions with missing data and compared that, for a number of evaluation measures, to the complete versions. The results of 3 evaluation measures for one of the evaluation data sets having 8 variables  $R_1$  to  $R_8$  are given in Table 1, which is just an excerpt of our evaluation results. MI-SAS in Table 1 refers to the multiple imputation routine in SAS; CHDI refers to the calibrated hot deck imputation method we propose in this paper. The results are in terms of the percent relative difference to the true values.

Table 1. Evaluation results (percent relative difference)

Mean								
	$R_1$	$R_2$	$R_3$	$R_4$	$R_5$	$R_6$	$R_7$	$R_8$
MI-SAS	0.00	-2.81	0.33	1.24	0.15	-0.90	-0.03	-2.84
CHDI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Standard Deviation								
	$R_1$	$R_2$	$R_3$	$R_4$	$R_5$	$R_6$	$R_7$	$R_8$
MI-SAS	0.00	-4.07	-0.11	-3.28	-0.14	-0.01	0.00	-0.10
CHDI	0.00	-1.61	0.02	-0.43	-2.53	-0.02	0.00	0.00
Median								
	$R_1$	$R_2$	$R_3$	$R_4$	$R_5$	$R_6$	$R_7$	$R_8$
MI-SAS	0.00	-1.86	0.00	0.03	0.00	0.41	-0.13	0.00
CHDI	-0.02	1.8	-0.04	0.52	1.41	4.31	0.28	0.00

As we calibrate to totals in our CHDI method, means are automatically preserved. In this respect CHDI naturally performs better than MI-SAS where the imputations are not calibrated to totals. Much more interesting is that for most variables CHDI also preserves the standard deviation better than MI-SAS. Apparently, satisfying edits rules and calibrating to totals improves imputation in this respect. Unfortunately, CHDI is not better than MI-SAS in all aspects as the median is better preserved by MI-SAS than by CHDI.

### 4. CONCLUSIONS

In this paper we have proposed an imputation method that can simultaneously satisfy specified edit rules and preserve known. The proposed method is much simpler than previous methods known from the literature. The first results of the new method are encouraging.

### REFERENCES

- [1] Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. *Annals of Applied Statistics* 7, pp. 1983-2006.
- [2] De Waal, T., J. Pannekoek and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New York.

# Progress in Sharing Statistical Data and Metadata using International Standards

## The implementation of SDMX in Istat beyond the dissemination and reporting

Francesco Rizzo ([rizzo@istat.it](mailto:rizzo@istat.it))<sup>1</sup>

**Keywords:** SDMX, data sharing, data reporting, data dissemination, SDMX Reference Infrastructure, SDMX-RI, RDF, Google/DSPL

### 1. INTRODUCTION

Since 2010, Istat started a multiannual programme, named Stat2015, aligned with the ESS 2020 vision, whose main aim is to modernize its information system and the way to produce statistics for responding efficiently to the new challenges. An important prerequisite for the realization of Stat2015 is the standardization and industrialization of the statistical processes through the adoption and implementation of statistical and technical standards, such as SDMX.

In this context, Istat has been developing and putting in production a set of cross-cutting building blocks (SDMX Istat Framework) allowing to move from the simple use of the standard for reporting data and metadata to International Organizations towards a more strategic perspective, namely for streamlining internal business processes through the harmonization of content, the management of metadata, and the dissemination of data and metadata in several formats and for several platforms.

The main aim of this paper is to illustrate the SDMX Istat Framework and how it responds to the different business needs identified within the Stat2015 multi-annual program.

#### 1.1. Background

Istat has been working on SDMX since 2004. At the beginning, in order to get experience, Istat participated to pilot projects launched by Eurostat within the European Statistical System: SDMX Open Data Interchange (SODI), Demography Rapid Questionnaire, EuroGroups Register, Census Hub. During those pilots, Istat developed a set of software SDMX 2.0 compliant.

From 2009 to 2012, Istat has participated to ESSnets [4][5] on SDMX (phase 1 and 2), contributing to the development of software and guidelines shared with ESS members. Furthermore, Istat actively participated in the design of the SDMX Reference Infrastructure developed by Eurostat.

Starting from 2010, SDMX became part of the strategic multi-annual Istat program, named Stat2015. In this context, Istat has defined an SDMX implementation strategy that can be summarized in Figure 1:

---

<sup>1</sup> ISTAT, Italian National Institute of Statistics.



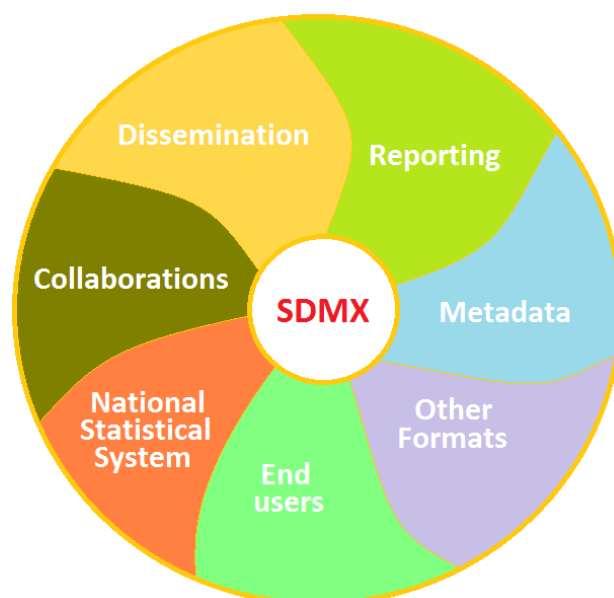


Figure 1. The seven dimensions of the SDMX Istat strategy

**Collaborations:** it is a fundamental part of the strategy, allowing a reduction of development costs through the sharing of SDMX software and know-how with other organizations. Currently, Istat is part of several communities that, besides other topics, are also focusing on SDMX aspects (OECD SIS – Collaboration Community, Statistical Network, Eurostat SDMX-RI User Group). Furthermore, Istat is running a bilateral collaboration with INEGI-Mexico and is part of the SDMX Technical and Statistical Working Groups.

**Dissemination and Reporting:** the SDMX Single Exit Point (SEP) [3] is already in production. The SEP is used for reporting to International Organization (e.g. Census data and National accounts data to Eurostat; Short Term Economic Statistical data to OECD, SDDS Plus data to IMF), and to disseminate data/metadata from the dissemination data warehouse (I.Stat) in machine-to-machine modality.

**Metadata:** the Istat Unified Metadata System is aimed at managing structural and reference metadata in an integrated way. Structural metadata are handled in an SDMX registry and synchronized with the dissemination/reporting metadata repository, while the reference metadata system (SIDI/Siqua) has been extended in order to be compliant with the Euro SDMX Metadata Structure (ESMS) and the ESS Standard for Quality Report Structure (ESQRS).

**Other formats:** in order to speed up data interoperability, the SDMX SEP web service has been extended to generate other formats, starting from the SDMX-IM: Resource Description Framework (RDF), Google dataset publishing language (DSPL), CSV and JSON.

**End users:** the interaction with the SDMX SEP is facilitated by a series of guidelines, tools and web GUI, accessible from the Istat website [3]. In particular the web GUI connected to the metadata repository allows browsing structural metadata needed to interpret correctly all of the SDMX datasets that can be extracted from the SEP, while the SDMX MS-Excel plug-in can be used to extract SDMX datasets from the SEP.

**National Statistical System:** Istat has been implementing a “distributed data warehouse” accessible through a hub web application. End users can browse the hub to define a

dataset of interest via SDMX structural metadata and retrieve the data directly from the data providers' databases. The whole architecture will be based on an SDMX hub architecture where the central application communicates with remote web services through SDMX messages.

## **2. METHODS**

One of the main constraints in developing the SDMX Istat Framework has been the optimization of the development costs and searching the suitable funds. Istat decided two strategic lines:

- harness the opportunity of Eurostat grants;
- reuse and extend the SDMX Reference Infrastructure (SDMX-RI) [2] developed by Eurostat.

The SDMX-RI “ultimate version” is based on the SDMX Common APIs, an open source project fostered by the SDMX Secretariat and implemented (SdmxSource) by Eurostat and Metadata Technology Ltd.

### **2.1. Financing the development**

In December 2013, Istat signed with Eurostat a grant agreement on “Horizontal and vertical integration: implementing technical and statistical standards in ESS”, with a duration of 24 months (extended till August 2015).

The grant foresees 24 deliverables split into 4 working packages:

WP1: coordination;

WP2: enhancement of the SDMX Reference Infrastructure, and its integration within the Istat information system;

WP3: enhancement of the Istat metadata management system;

WP4: contribution to the development of the SDMX standard and implementation of capacity building action.

7 units from 2 different Directorates are involved in the grant, for a total of 778 man/days. Furthermore, the grant is used to support over 1000 man/days from a subcontractor.

### **2.2. The overall SDMX architecture in Istat**

Within Istat, an SDMX architecture based on SDMX-RI and complemented by the SDMX Istat Framework building blocks is already in place.

The main aim of this architecture is to allow end-user applications to browse and query data stored in the dissemination data warehouse I.Stat. To this purpose, the suitable SDMX structural metadata (Data Structure Definitions, Code lists, Concept schemes, Category schemes, etc.) have been developed and mapped against the data stored in I.stat.

The SEP web service gives access to end-user applications in machine-to-machine modality, using SOAP and/or REST protocols, and sending and receiving SDMX messages. From the SEP web service, it is possible to extract data in other formats (RDF, Google/DSPL, JSON).

The same architecture is used for reporting to International Organizations in both push and pull mode. For example, datasets from National Accounts, Job Vacancy Survey and Labour Cost Indicator are extracted by domain managers and sent to Eurostat through the eDamis channel (Eurostat's Single Entry Point). At the contrary, census data and

economic indicators are collected in pull mode by Eurostat (Census Hub), OECD (STES) and IMF (SDDS Plus).

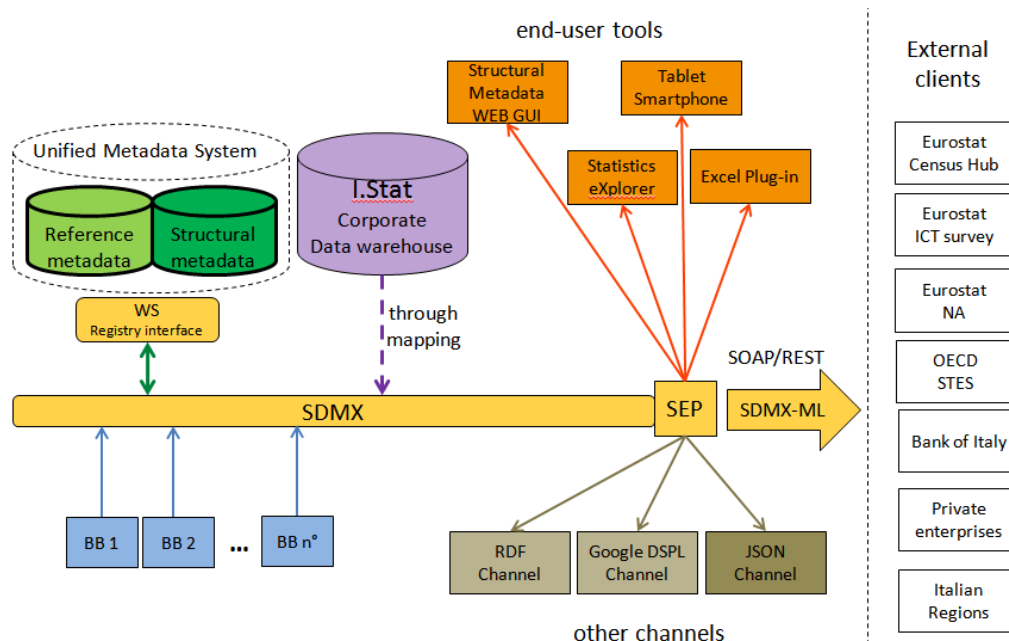
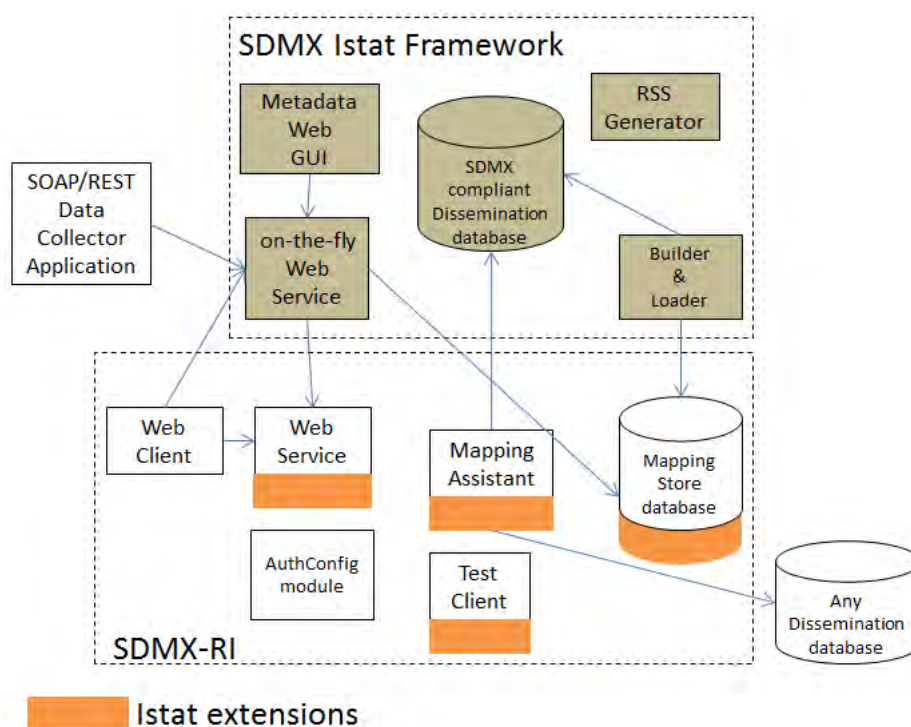


Figure 2. SDMX overall architecture in Istat

### 2.3. The SDMX Istat Framework building blocks

The SDMX Istat Framework is a set of pick-and-choose building blocks allowing a statistical office to handle data and metadata that has to be shared with other organizations or end users. It acts as a series of add-ons of the SDMX-RI in order to cover most of the aspects identified within the Istat SDMX strategy.

The following component diagram depicts the view of the SDMX Istat Framework:



The SDMX Istat Framework building blocks are the ones with brown colour, while the orange blocks identify extensions of the SDMX-RI components developed by Istat;

**Builder** - it allows creating an SDMX compliant database.

**Loader** - it allows loading into the SDMX compliant database csv or SDMX files.

**On-the-fly Web Service** - it is responsible for exposing the data and structural metadata using a Web Service interface that follows the guidelines of the SDMX v2.0 standard for Web Services and the Web Service Guidelines for SDMX v2.1 standard [6]. Data can be extracted directly from the SDMX compliant database created through the Builder or from any dissemination database mapped using the Mapping Assistant. Furthermore a Data Collector Application can negotiate other format beyond SDMX-ML such as RDF, Google/DSPL and JSON.

**Metadata Web GUI** – it provides a Web graphical user interface that helps domain managers in designing SDMX artefacts, and end-users to browse and query the structural metadata repository (extended Mapping Store database).

**RSS Generator** – it is responsible for generating a feed entry on the event of new data arriving from the Loader. The generated feed is able to trigger Data Collector Applications providing them the suitable SDMX queries.

### 3. CONCLUSIONS

International standards, such as SDMX, can help the modernization processes within the statistical Organizations, but efficient tools should be available in order to speed the implementations. The SDMX Reference Infrastructure could be considered a very good starting point that can be easily integrated with existing information systems and extended using the SDMX Common APIs: in this context the SDMX Istat Framework represents a real use case.

### REFERENCES

- [1] SDMX: <http://sdmx.org/>
- [2] SDMX Reference Infrastructure:  
[https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/SDMX\\_Reference\\_Infrastructure\\_SDMX-RI](https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/SDMX_Reference_Infrastructure_SDMX-RI)
- [3] SDMX Single Exit Point: <http://www.istat.it/it/strumenti/web-service>
- [4] ESSnet on SDMX phase I: <http://www.cros-portal.eu/content/sdmx-finished>
- [5] ESSnet on SDMX phase II: <http://www.cros-portal.eu/content/sdmx-ii-finished>
- [6] SDMX Web Service Guidelines: [http://sdmx.org/?page\\_id=10](http://sdmx.org/?page_id=10)

# A new international standard for data validation and processing

Marco Pellegrino ([marco.pellegrino@ec.europa.eu](mailto:marco.pellegrino@ec.europa.eu))<sup>1</sup>

**Keywords:** Data validation, transformation, open standards, SDMX, GSIM

## 1. INTRODUCTION

In the European Statistical System (ESS), Data validation is a critical issue which faces a number of problems due to the specificities of the production process. Data collection, processing and compilation are done by Member States (MS), while more processing steps and final dissemination at European level are performed by Eurostat. The role of Eurostat is critical for spotting any error in the figures transmitted, but in many cases Eurostat and MS perform similar validation checks, while in other cases it may happen that some quality checks are performed neither by MS nor Eurostat.

The process suffers from a series of inefficiencies: lack of coordination, but also lack of documentation, lack of formalisation of validation procedures and rules, low harmonisation of software solutions. Even when they are properly documented, validation rules are described using specific languages developed by individual process managers independently from each other, instead of using a common harmonised syntax. This situation raises several issues related to quality assessment (data completeness, accuracy, timeliness and punctuality,...) while hampering the integration of validation solutions and, as a consequence, the perspectives of a large-scale reduction of IT development and maintenance costs at European level.

These issues require the elaboration of a comprehensive solution, which calls for a portfolio of actions. At European level, a comprehensive project on data validation has been launched in the framework of the ESS Vision 2020 [1]. The scope of this document is to present an international activity, which is coordinated with the ESS action mentioned above, and is specifically addressing the issue of the lack of a standard syntax for expressing validation and editing rules.

## 2. THE NEW VALIDATION AND TRANSFORMATION LANGUAGE (VTL)

Building on the SDMX [2] ISO standard for data and metadata exchange, a task-force was formed in 2013 with the purpose of elaborating a formal and standard framework for the description of logical algorithms to validate statistical data and calculate derived data.

SDMX already has, in its information model, a module for "transformations and expressions", although a specific language did not exist. To make this framework operational, a standard "language" for defining validation and transformation rules (set of operators, their syntax and semantics) is needed, together with appropriate IT formats for exchanging rules and related metadata, and web services to store and retrieve them.

The intention is to provide a language which is usable by statisticians to express logical validation rules and transformations on data, whether described as dimensional tables or as unit-level data. The assumption is that this logical formalization of validation and

---

<sup>1</sup> Eurostat, Unit B1: Methodology and corporate architecture – Standards Team

transformation rules would be converted into specific programming languages for execution (SAS, R, Java, SQL, etc.) but would provide a “technology-neutral” expression at business level of the processing taking place, against which various implementations can be mapped. Experience with existing examples suggests that this goal is achievable.

An important point that emerged is that, besides SDMX which was the starting point, other information standards (such as GSIM [3] and DDI [4]) were also interested in such a language. However, each standard operates on its model objects and produces objects within the same model: to cope with this, the VTL language has been built upon a very basic information model, taking the common parts of GSIM, SDMX and DDI. This way, existing technical standards (SDMX, DDI, others) may adopt VTL by mapping their information models against the VTL one. Therefore, although a work-product of SDMX, the VTL language will be usable with other standards as well.

In the VTL model, both unit and dimensional data are considered as mathematical functions having independent and dependent variables and are treated in the same way. For each Unit (e.g. a person) or Group of Units of a Population (e.g. groups of persons of a certain age and civil status) identified by means of the values of the independent variables (e.g. either the “person id” or the age and the civil status), the mathematical function provides for the values of dependent variables, which are the properties to be known (e.g. revenue, expenses ...). This way, the manipulation of any kind of data (unit and dimensional) is brought back to the manipulation of very simple and well-known objects, which can be easily understood and managed by users.

### **3. THE HIGH-LEVEL CHARACTERISTICS OF THE VALIDATION AND TRANSFORMATION LANGUAGE (VTL)**

The task-force identified the main characteristics that the language should follow:

- **User orientation**
  - ✓ designed for the users, who should be able to define transformation and validation expressions autonomously, without IT skills and IT people intermediation;
  - ✓ intuitive and friendly (users should define and understand expressions as easily as possible);
  - ✓ oriented towards statistics, which is the main user skill; it should be possible to include operators specifically needed in the statistical process (for example operators for data validation, editing and imputation, time-series processing, ...).
- **Integrated approach**
  - ✓ independent of the statistical domain of the data to be processed;
  - ✓ mapped unambiguously to the proper information model (IM); in other words, it should be able to operate on IM artefacts and to produce other IM artefacts (this is a basic property of any robust language).
  - ✓ suitable for various typologies of data of a statistical environment (for example dimensional data, survey data, registers data or transactions, micro and macro, quantitative and qualitative, ...), as much as they are supported by the IM;
  - ✓ independent of the steps of the statistical process (GSBPM [5]) and usable in any one of them.
- **IT implementation independence**
  - ✓ not oriented to a specific IT implementation but allowing many different implementations (this property is particularly important for a standard language, which should not be tied to a specific IT solution and should allow different institutions to rely on different IT environments);

- ✓ able to support the possible use of various IT tools in an integrated IT solution (for example, different calculation tools in different steps of statistical data processing);
- ✓ make users unaware of the IT solution as much as possible;
- ✓ avoid impacts on users as much as possible if the IT solution changes (for example following the adoption of another IT tool).
- **Active role for processing**
  - ✓ described formally as for its grammar, to be easily parsed and processed (i.e. in Backus-Naur Form);
  - ✓ able to drive the software that perform calculations, so automatically convertible in the languages of the IT tools used for calculations (once the language is defined, it might be useful to support its conversion in some widely used IT languages, for example SQL, R, XML languages ...);
  - ✓ able to generate validation results that can be unambiguously interpreted by software and as much as possible easily interpretable by statisticians.
- **Extensibility and customizability**
  - ✓ to introduce new VTL operators according to evolution of the business needs (e.g. the operators for the validation first and those for the compilation after);
  - ✓ able to include operators derived from other languages / tools (e.g. “SQL like” operators, operators for time series processing ...);
  - ✓ able to customize the operators for specific needs, for example of specific organizations / specific processing (note that this requirement is typically not fulfilled by the IT languages that have a fixed list of operators).
- **Proper governance**
  - ✓ It implies the creation of appropriate governance rules to control the evolution of the language;
  - ✓ In addition to the standard, there is the need for allowing customized parts of the language under the private governance of single institutions, which may integrate the language for their own purposes; therefore coordinated governance rules between the standard part and the customized parts should be introduced.
- **Language effectiveness**

The effectiveness is connected to some aspect of the language features, for example:

  - ✓ Historicity: possible changes of the artefacts or its sets with reference to the change of time;
  - ✓ Persistency control: possibility of specifying the persistency of the intermediate results;
  - ✓ Expressions chaining: possibility of expressions having also other expressions as an input parameter;
  - ✓ Strictly defined (or clearly stated when undefined) behaviour for missing data, multi measures, data attributes.

#### 4. RESULTS

In February 2015, after a public review period in autumn 2014, the VTL 1.0 package has been published on the web at <http://www.sdmx.org>. The set includes:

- a) General part, highlighting the main characteristics of VTL, its core assumptions and the information model the language is based on;
- b) Full library of operators ordered by category, including examples;
- c) BNF notation (Backus-Naur Form) which is the technical notation used as a test bed for all the examples throughout the document.



The operators included in this 1.0 version of VTL are summarized in the diagram below.



## 5. CONCLUSIONS

VTL 1.0 contains a high-level definition of the general characteristics of the language and a list of operators for validation and transformation, as well as a simple information model on which the VTL can operate. VTL is usable by statisticians to express logical validation rules and transformations on data, whether described as dimensional tables or as unit-level data. The assumption is that this logical validation and transformation rules provide a “technology-neutral” expression at business level of the processing taking place, against which various implementations can be mapped.

The specifications for exchanging VTL validation rules in SDMX messages, for storing rules and for requesting validation rules from web services will be provided in a specific update to the SDMX Technical Standards on which the task-force is working on. This first implementation exercise will allow a further fine-tuning and bug-fixing of the first version, leading to a VTL 1.1 within one year. At Eurostat, VTL will be primarily implemented through SDMX. Implementation tests are already foreseen with some pilot domains.



## 6. REFERENCES

- [1] European Statistical System Committee (ESSC), ESS Vision 2020 (2014), <http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255>
- [2] Statistical Data and Metadata eXchange (SDMX), <http://www.sdmx.org>
- [3] UN/ECE, Generic Statistical Information Model 1.1 (GSIM), [http://www1.unece.org/stat/platform/download/attachments/97356610/GSIM%20Specification%201\\_\\_1.docx?version=3&modificationDate=1388474361592&api=v2](http://www1.unece.org/stat/platform/download/attachments/97356610/GSIM%20Specification%201__1.docx?version=3&modificationDate=1388474361592&api=v2)
- [4] Data Documentation Initiative (DDI), <http://www.ddialliance.org>
- [5] UN/ECE, Generic Statistical Business Process Model 5.0, <http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0>

# A metadata-driven process for handling statistical data end-to-end

Denis Grofils ([denis.grofils@ec.europa.eu](mailto:denis.grofils@ec.europa.eu))<sup>1</sup>

**Keywords:** Statistical process management, modernisation of official statistics, standards-based industrialisation, metadata-driven processes

## 1. INTRODUCTION

Some fundamental changes are affecting the environment in which producers of official statistics are operating. Today, it seems widely accepted that statistical organisations will need to evolve continuously to remain relevant and sustainable. Strategic initiatives at the highest level (in particular the ESS Vision 2020 [1] and the HLG Strategic Vision [2]) consider that modernisation of official statistics will consist of a "standards-based industrialisation" of statistical production, among others meaning that:

- Statistical processes should be made more efficient and robust by intensifying the sharing of knowledge, experiences and methodologies but also by sharing data, services and resources where appropriate;
- Collaboration should be based on agreed standards and common elements of technological and statistical infrastructure which means that also processes and information models are agreed and shared on a wider scale than in the past to achieve modernisation objectives.

Statistical process management aims at managing, automating and improving processing in national and international statistical organisations. Statistical organisations essentially perform similar functions with differences depending on their contexts (e.g. languages, code lists or processing methods). In the same way, various statistical domains covered by an organisation (e.g. employment, prices, wages) are supported by specific production processes but the activities carried out as part of various processes have generally much in common. In many cases, differences between organisations and domains can be captured as metadata and be fed into processes as operating data so that a common process can serve despite the differences. This type of processes is generally referred to as "metadata-driven processes". This approach enables the evolution from a "stovepipe production model" to an "integrated production model" as envisioned by current modernisation strategies.

With an analogy to model-driven software engineering [3] the benefits of a metadata-driven approach to statistical production can be classified as following:

- Improvement in short-term productivity: increase in the value of production components by increasing the number of functionalities they deliver. The more functionalities can be provided by one component, the higher the productivity.
- Improvement in long-term productivity: reduction of the rate at which production components become obsolete. The longer a component remains operational, the greater the return derived from the effort of creating it. This can be achieved by reducing sensitivity to some types of changes in the following ways:

---

<sup>1</sup> Eurostat, Unit B1: Methodology and corporate architecture

- Knowledge externalisation: reduction of dependence to individual personnel members by extracting knowledge and representing it in structured and accessible ways.
- Agility and adaptability to changing requirements: new features and capabilities can be supplied with limited impact on existing parts in terms of maintenance efforts and disruption to existing systems. Responses to changing user needs and opportunities provided by the environment can be addressed using a configurable modular production infrastructure minimizing human intervention on the infrastructure.
- Technological independence towards tools used to create and execute processes: this is achieved by decoupling components from their development tools and by storing metadata artefacts in formats that can be used by other tools.

The metadata-driven approach relies strongly on principles of standardisation and interoperability, reuse and domain-independent standard processes. It should be noted that interoperability and standardisation may be considered at different levels (e.g. legal, organisational, semantic, technical) as underlined by the European Interoperability Framework (EIF) [4].

## 2. METHODS

Official statistics modernisation initiatives typically consider Enterprise Architecture (EA) as a framework. The EA framework aims basically at enforcing that IT developments are aligned with business strategies and advocates strong separation of concern and strict decoupling between four considered architectural layers: business architecture, information architecture, application architecture and technology architecture. In the context of official statistics the following definitions can be provided for the first 2 layers [5]:

- Business Architecture (BA) *covers all the activities undertaken by a statistical organization, including those undertaken to conceptualize, design, build and maintain information and application assets used in the production of statistical outputs. BA drives the Information, Application and Technology architectures for a statistical organization.*
- Information Architecture (IA) *classifies the information and knowledge assets gathered, produced and used within the Business Architecture. It also describes the information standards and frameworks that underpin the statistical information. IA facilitates discoverability and accessibility, leading to greater reuse and sharing.*

In this approach, standardisation is a key-enabler for achieving modernisation ambitions of increasing the level of sharing and re-use. Indeed, a sufficient level of standardisation is necessary on each EA layer to permit integration of building blocks on a service based configurable production platform.

On the top level, BA is supported by a proper Business Process Management (BPM), which can be defined as a management discipline targeting the improvement of corporate performance by managing and optimising the organisation's business processes. Interoperable business processes need to be structured according to common standards. This is achieved at the level of the global statistical community thanks to a major business architecture standard for official statistics: the Generic Statistical Business Process Model (GSBPM) [6]. Business processes modelled according to common BA standards will allow for organisational interoperability (in EIF terms). In a metadata-

driven approach, business capabilities required by production processes will be supported by services designed according to principles of a Service-Oriented Architecture (SOA). SOA may be defined as *“a paradigm for organising and utilising distributed capabilities that may be under the control of different ownership domains. It provides a uniform means to offer, discover, interact with and use capabilities to produce desired effects consistent with measurable preconditions and expectations”* [7]. In this context, a SOA service is defined as *“a discrete unit of business functionality that is made available through a service contract”* and designed according to a set of principles such as loose coupling, abstraction, reusability, autonomy and statelessness.

A distinction between generic and specific functions can be made based on the specificity of the business-logic that is encapsulated. Nevertheless, some functions may contain an important amount of business logic while having the potential to be used in various contexts. The criterion proposed to distinguish between specific and generic functions is the level of dependency towards individual statistical domains. In our context, a metadata-driven process is defined as a process that relies on activities that do not embed domain-specific business logic but rather access it through a corporate metadata registry representing it according to corporate metadata standards. Domain-specific content parameterizing metadata-driven processes and metadata outputs resulting of process executions comprise concepts and definitions, data and metadata structures, information related to data exchange, descriptions of business process models, business process implementations, descriptions of data processing, process metrics, and reference metadata of all nature.

Abstraction layers play an important role in the context of metadata-driven processes:

- Abstraction of business entities manipulated by services: data are manipulated through structural representations referenced centrally by a metadata registry. Different levels of modelling are considered: conceptual-level reference models such as the Generic Statistical Information Model (GSIM) [8] which provides an overarching pivot-model to which implementation-level standards can be mapped (e.g. SDMX, DDI, XBRL, RDF, ...).
- Abstraction of operations performed by services on business entities: operations are manipulated through representations independently of their implementation. Processes and available services are referenced centrally by a metadata registry. Information models will cover different levels of abstraction of processes, essentially process design, implementation, execution and monitoring. Representations linked to processes are based on standard information models as well (e.g. BPMN, XPDL, BPEL, VTL<sup>2</sup>, ...).

Abstraction of operations and data allow enforcing several of the SOA principles and ultimately designing and executing business processes independently of underlying technical implementations. Implementation of EA and SOA in the context of official statistics is supported to some extent by the specification of a Common Statistical Production Architecture (CSPA) [9] developed in recent years at the level of the global statistical community.

### 3. RESULTS

Metadata-driven processes allow sharing IT and methodological components across different domains and different organisations. The idea is that all the detailed technical

---

<sup>2</sup> Validation and Transformation Language, the new standard language developed by a task-force of the SDMX Technical Working Group: see [www.sdmx.org](http://www.sdmx.org).

work to implement the process can be done once and an abstracted process made available to various users. Using the process in a particular context (i.e. organisation or domain) then requires only all the required input metadata to be available. This not only achieves the goal of making the process reusable in a variety of contexts but also captures, in the metadata, all of the specific knowledge needed to run the process in the local context - knowledge that otherwise lives in minds of individuals and is hard to capture and document. This also represents a progress against the goals of standardised processes, repeatability, and error and quality management.

Operating a statistical organisation with metadata-driven processes offers several advantages among which those described below:

- Highly configurable process flows: in most cases, no software has to be updated to create or update a production flow only metadata need to be updated.
  - This results in a much higher flexibility of the whole statistical production process.
  - This results in an empowerment of statisticians limiting IT-related tasks and allowing a stronger focus on most value-adding activities such as statistical design, configuration and monitoring of process flows, interpretation and explanation of results, etc.
- High level re-use of software: generic configurable functions are supported by cross-domain shared statistical services.
  - Resulting in an increased efficiency and reduced costs by avoiding multiple developments of virtually the same software in different production lines or different organisations.
  - Resulting in an increased harmonization and interoperability through the use of standard software building blocks.
  - Resulting in an improved quality of the data through the use of widely accepted and validated software building blocks and improved comparability among data coming from different countries.
- High level re-use of metadata: common metadata elements are shared across statistical domains.
  - Resulting in an increased efficiency and reduced costs by avoiding multiple developments of redundant and potentially inconsistent metadata elements in the area of processes, exchanges, structures and concepts in different production lines or different organisations.
  - Resulting in an improved quality of the data through the use of shared and widely accepted metadata elements in the area of concepts, process models, processing instructions, etc.
  - Resulting in improved possibilities of evaluation and monitoring of the whole statistical production process through exhaustive, standardised and centrally accessible process metrics.

This being said, it should be clear that the evolution towards a metadata-driven approach to statistical production has several implications in terms of infrastructure and business changes required and represents several challenges:

- Metadata standards: the deployment of metadata-driven statistical production processes requires the availability of corporate models and standards for the representation of metadata related to data and processes. The elaboration of a robust

solution in this area represents a necessary investment to enable intensified sharing of processes, methods and IT components across statistical domains and organisations.

- Corporate metadata registry: metadata-driven processes rely on a metadata registry that is the interface from which metadata inputs are acquired and to which metadata outputs are communicated. The metadata registry is the central component of a metadata-driven production approach. The capacity to deploy metadata-driven statistical production processes depends on the availability of an adequate metadata registry implementing the corporate metadata models and standards.
- Process manager: the availability of a process manager and the integration of such a component with the corporate metadata registry is a key element enabling the deployment of metadata-driven statistical production processes.
- Re-engineering of production processes: processes need to be designed to be metadata-driven and to use corporate metadata standards and infrastructure. Although it is possible to deploy one or just a few metadata-driven processes in an organisation, the metadata-driven approach delivers its maximum value when all of the business processes are metadata-driven, as this allows the highest degree of factorisation of metadata and services. The deployment of a metadata-driven production approach would thus require ultimately that all business processes are possibly re-engineered through a comprehensive corporate Business Process Re-engineering (BPR) programme.

#### 4. CONCLUSIONS

ESS Vision 2020 modernisation objectives of improving efficiency of the overall statistical production process can be met by the generalization of metadata-driven processes. Productivity gains can be expected from this approach in the short-term through intensified sharing and re-use of production assets (IT and methodological) across statistical domains and organisations. Productivity gains in the longer term will be achieved by improving transparency, maintainability, adaptability and technological independence of production components. Such an approach is also expected to deliver benefits in the areas of processes standardisation, repeatability, and quality management.

The metadata-driven production paradigm has strong implications in terms of metadata management in statistical organisations. Current metadata management infrastructure will generally not be adapted to requirements of metadata-driven production based on a SOA [10]. This new paradigm will require a more holistic coordination of metadata management. Next-generation metadata registries will be required to provide sufficiently exhaustive and structured information to centralise and drive the complete statistical production process with a performance that allows users to meet their own Service Level Agreements (SLAs). This is expected to enable horizontal synergies (across domains and organisations) and vertical synergies (across EA layers). Such an evolution is expected to be facilitated by the adoption of explicit metadata management strategies and the development of proper information architectures by statistical organisations.

The organisational impact of implementing this new paradigm for statistical production should not be neglected and the elaboration of stepwise implementation relying on a proper change management will certainly be a key enabler for a successful evolution towards an integrated metadata-driven statistical production approach in the long-term. In this respect, experiences of organisations that have already started investing in this direction will be extremely valuable (notably INE in Spain [11] and ABS [12] in Australia).

## REFERENCES

- [1] European Statistical System Committee (ESSC), The ESS Vision 2020 (2014), <http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255>
- [2] High-Level Group for the Modernisation of Statistical Production and Services (HLG), Strategic vision of the HLG (2012), <http://www1.unece.org/stat/platform/display/hlgbas/Strategic+vision+of+the+HLG>
- [3] C. Atkinson and T. Kühne, Model-Driven Development: A Metamodeling Foundation (2003), <http://mddpapers.googlecode.com/svn-history/r25/trunk/ASE09/doc/mda-foundation.pdf>
- [4] European Commission (DIGIT), European Interoperability Framework for Pan-European eGovernment Services (2004), <http://ec.europa.eu/idabc/servlets/Docd552.pdf?id=19529>
- [5] CSPA 1.1 Glossary (2015), <http://www1.unece.org/stat/platform/display/CSPA/Annex+5.+Glossary>
- [6] UN/ECE, The Generic Statistical Business Process Model 5.0 (2013), [http://www1.unece.org/stat/platform/download/attachments/97356247/GSBPM%205\\_0.docx?version=1&modificationDate=1387861584474&api=v2](http://www1.unece.org/stat/platform/download/attachments/97356247/GSBPM%205_0.docx?version=1&modificationDate=1387861584474&api=v2)
- [7] Organisation for the Advancement of Structured Information Standards (OASIS), Reference Model for Service Oriented Architecture 1.0 (2006), <https://www.oasis-open.org/committees/download.php/19679/soa-rm-cs.pdf>
- [8] UN/ECE, The Generic Statistical Information Model 1.1 (2013), [http://www1.unece.org/stat/platform/download/attachments/97356610/GSIM%20Specification%201\\_1.docx?version=3&modificationDate=1388474361592&api=v2](http://www1.unece.org/stat/platform/download/attachments/97356610/GSIM%20Specification%201_1.docx?version=3&modificationDate=1388474361592&api=v2)
- [9] UN/ECE, The Common Statistical Production Architecture 1.1 (2015), [http://www1.unece.org/stat/platform/download/attachments/108103974/CSPA%201\\_1.docx?version=2&modificationDate=1421739512027&api=v2](http://www1.unece.org/stat/platform/download/attachments/108103974/CSPA%201_1.docx?version=2&modificationDate=1421739512027&api=v2)
- [10] Gartner, What Is a Registry/Repository, and Who Should Consider One? (2007), [http://www.gartner.com/it/content/754400/754413/qa\\_what\\_is\\_a\\_registry.pdf](http://www.gartner.com/it/content/754400/754413/qa_what_is_a_registry.pdf)
- [11] P. Revilla et al., Implementing a corporate-wide metadata driven production process at INE Spain (2012), [http://www.q2012.gr/articlefiles/sessions/17.1\\_Revilla\\_INE%20Spain%20Topic%20II%20Session%2017.pdf](http://www.q2012.gr/articlefiles/sessions/17.1_Revilla_INE%20Spain%20Topic%20II%20Session%2017.pdf)
- [12] A. Rivera et al., Metadata driven business process in the Australian Bureau of Statistics (2013) <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP21.pdf>

# Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies

Giulio Barcaroli<sup>1</sup> (barcarol@istat.it), Monica Scannapieco<sup>1</sup> (scannapi@istat.it), Marco Scarnò<sup>2</sup> (m.scarnò@cineca.it), Donato Summa<sup>1</sup> (donato.summa@istat.it),

**Keywords:** Big data, Web Scraping, Web data

## 1. INTRODUCTION

The Internet can be considered as a data source (belonging to the vast category of Big Data), that may be harnessed in substitution, or in combination with, data collected by means of the traditional instruments of a statistical survey. The survey on “ICT in enterprises”, carried out by all EU Statistical Institutes, is a natural candidate to experiment this approach, as the questionnaire contains a number of questions, related to the characteristics of the websites owned or used by the enterprises, whose answers can be deduced directly by the content of these websites (for instance the presence of web sales functionalities). An experiment is being conducted whose aim is twofold: (i) from a technological point of view, to verify the capability to access the websites indicated by enterprises participating to the sampling survey, and collect all the relevant information, (ii) from a methodological point of view, to use the information collected from the Internet in order to predict the characteristics of the websites not only for surveyed enterprises, but for the whole population of reference, in order to produce estimates with a higher level of accuracy. The first phase of the experiment has been based on the use of the survey data, that is a sample of 19,000 respondent enterprises, that indicated a total of 8,600 websites. Websites have been “scraped” and collected texts have been used as train and test sets in order to verify the validity of the applied machine learning techniques [1]. In the second phase, the whole reference population (192,000 enterprises) is involved, together with the about 90,000 websites owned or used by them. The web scraping task, that was already crucial in the first phase, becomes critical from the point of view of both efficiency and effectiveness, considering the increased number of websites. For this reason, a number of different solutions are being investigated, based on (i) the use of the Apache suite Nutch/Solr, (ii) the use of the tool HTTrack, (iii) the development of new functionalities for web scraping in the package ADaMSoft making use of JSOUP. In this paper, these alternative solutions are evaluated by comparing obtained results in terms of both efficiency and compliance to Official Statistics (OS) requirements.

## 2. WEB SCRAPING SYSTEMS

In this Section, we first introduce the Web scraping concept and we position our work with respect to previous ones (Section 2.1). Then, in the subsequent sections we provide an overview of the scraping systems under evaluation (Sections 2.2, 2.3 and 2.4)

---

<sup>1</sup> Istat

<sup>2</sup> CINECA



## 2.1. Web Scraping: State of the Art in Official Statistics

Web scraping is the process of automatically collecting information from the World Wide Web, based on tools (called scrapers, internet robots, crawlers, spiders etc.) that navigate, extract the content of websites and store scraped data in local data bases for subsequent elaboration purposes. Previous work on the usage of Internet as a data source for Official Statistics by means of Web scraping was carried out by Statistics Netherlands in recent years [2]. In particular, a first domain of experimentation was related to *air tickets*: the prices of air tickets were collected daily by Internet robots, developed by Statistics Netherlands supported by two external companies, and the results were stored for several months. The experiment showed that there was a common trend between the ticket prices collected by robots and existing manual collection [3]. Two additional domains of experimentation were *Dutch property market* and *Clothes prices*, the first exhibiting more regularity in the sites structure, the latter more challenging with respect to automatic classification due to lack of a standard naming of the items, and variability in the sites organization. Similarly, in Italy a scraping activity was performed to get on *consumer electronics prices* and *airfares* [4].

## 2.2. The Apache Stack: Nutch/Solr/Lucene

The Apache suite used for crawling, content extraction, indexing and searching results is composed by Nutch and Solr. Nutch (available at <https://nutch.apache.org/>) is a highly extensible and scalable open source web crawler; it facilitates parsing, indexing, creating a search engine, customizing search according to needs, scalability, robustness, and scoring filter for custom implementations. Built on top of Apache Lucene and based on Apache Hadoop, Nutch can be deployed on a single machine as well as on a cluster, if large scale web crawling is required.

Apache Solr (available at <https://lucene.apache.org/solr/>) is an open source enterprise search platform that is built on top of Apache Lucene. It can be used for searching any type of data; in this context, however, it is specifically used to search web pages. Its major features include full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document handling. Providing distributed search and index replication, Solr is highly scalable.

Both Nutch and Solr have an extensive plugin architecture useful when advanced customization is required. Although this web scraping approach requires an initial effort in terms of technological expertise, in the long run it can lead to a substantial return on investment as it can be used on many other contexts to access Big Data sources. As an example, it can be used as platform to access and analyse web resources like blogs or social media, to perform semantic extraction and analysis tasks.

## 2.3. HTTrack

HTTrack (available at <http://www.httrack.com/>) is a free and open source software tool that permits to “mirror” locally a web site, by downloading each page that composes its structure. In technical terms it is a web crawler and an offline browser that can be run on several operating systems (Microsoft Windows, Mac OS X, Linux, FreeBSD and Android). HTTrack’s strength points are ease of use, fine parameters configurability. It can be run via graphical user interface or in batch mode via command line.

## 2.4. JSOUP

JSOUP (<http://jsoup.org>) permits to parse and extract the structure of a HTML document. It has been integrated in a specific step of the ADaMSoft system (<http://adamsoft.sourceforge.net>), this latter selected as already including facilities that allow to handle huge data sets and textual information. In this approach we collected the content of the HTML pages and the information on their structure (TAGS), because these can contain discriminant terms that can help us in identifying the nature of the website; for example a button associated to an image called "paypal.jpg" could be a clear sign of web sales functionality.

## 3. A COMPARATIVE ANALYSIS AND RECOMMENDATIONS

The three systems described in the previous sections have been compared with respect to efficiency features (summarized in table 1). In addition, it is possible to verify how they are appropriate to address requirements that are specific to their usage in Official Statistics production processes (summarized in table 2).

Tool	# websites reached	Average number of webpages per site	Time spent	Type of Storage	Storage dimensions
Nutch	7020 / 8550=82,1%	15,2	32,5 hours	Binary files on HDFS	2,3 GB (data) 5,6 GB (index)
HTTrack	7710 / 8550=90,2%	43,5	6,7 days	HTML files on file system	16, 1 GB
JSOUP	7835/8550=91,6%	68	11 hours	HTML ADaMSoft compressed binary files	500MB

**Table 1: Efficiency features of the three systems**

Looking at Table 1:

- HTTrack and JSOUP are comparable with respect to web site reachability, while the number of websites reached by Nutch is considerably lower.
- JSOUP outperforms on the actual download of pages at the reached sites with respect to both the other systems.
- Time performances are again in favour of JSOUP. However, it is important to notice that we did not spend too much time to optimize the performance of HTTrack as it was the system that we experimented later in time, i.e. there is a margin to optimize such performances in the next steps.
- Space performances. Given that JSOUP was integrated with the ADaMSoft system, it was possible to compress the files resulting from the scraping and hence to save disk space.

Tool	Access to specific element of HTML Pages	Download site content as whole for semantic extraction and discovery	Document Querying	Scalability to Big Data Size
Nutch	Difficult	Easy	Easy	Easy
HTTrack	Easy	Easy	Difficult	Difficult
JSOUP	Easy	Easy	Difficult	Difficult

**Table 2: Effectiveness of the systems with respect to requirements**

As shown in Table 2:

- The column related to the access to specific elements of HTML pages does show that HTTrack and JSOUP are more appropriate than Nutch. In order to understand the implication of such a feature, it is important to notice that the scraping task can be design-based, i.e. it is possible to design in advance the access (i) to elements of specific website structures (e.g. the table of a specific site related to the price list of some products) or (ii) to general website elements (e.g. the label of a shopping cart image possibly present). Nutch does not allow (easily) this kind design-time access, being instead more appropriate to carry out exploratory runtime analysis (as remarked also by the following features).
- The column Download Site Content does show that the three systems perform well, i.e. they address the requirement by permitting an easy implementation of it.
- The column Document Querying shows that Nutch is the best, and this is indeed motivated by the native integration of Nutch with Solr/Lucene platform.
- Finally, to address the scalability need, the native integration of Nutch with MapReduce/Hadoop infrastructure makes Nutch again the best choice. However we do observe that the fact that JSOUP has been integrated within ADaMSOft also permits to store on a secondary structured or semi-structured storage in order to scale up (though we have not yet tested this functionality).

#### 4. CONCLUSIONS

A first remark is that a scraping task can be carried out for different purposes in OS production, and the choice of one tool for all the purposes may not always be possible.

For the specific scraping task required by the “ICT Usage in Enterprises” survey, the usage of JSOUP/ADaMSOft appears to be the most appropriate. In the second step of the project, when we will have to scale up to about 90,000 websites, we will test how such a system performs with respect to the scalability issue.

Finally, we highlight that the scraping task evaluated in this paper with three different systems is a sort of “generalized” scraping task: it indeed assumes a data collection without any specific assumption on the structure of the websites. In this sense it goes a step further with respect to all the previous experiences

## REFERENCES

- [1] Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarnò M., Summa D. - Internet as Data Source in the Istat Survey on ICT. *Accepted for publication in Austrian Journal of Statistics* (2014)
- [2] Ten Bosh O, Windmeijer D (2014). On the Use of Internet Robots for Official Statistics. In MSIS-2014.  
URL:[http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic\\_3\\_NL.pdf](http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf).
- [3] Hoekstra R, ten Bosh O, Harteveld F. Automated data collection from web sources for official statistics: First experiences." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, 28(3-4), 2012.
- [4] Giannini R., Lo Conte R., Mosca S., Polidoro F., Rossetti F. Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, Q2014 European Conference on quality in official statistics Vienna, 2-5 June 2014.

# Automatic data collection on the Internet (web scraping)

Ingolf Boettcher ([ingolf.boettcher@statistik.gv.at](mailto:ingolf.boettcher@statistik.gv.at))<sup>1</sup>

**Keywords:** web scraping, Price Statistics, Internet as data source, data collection methods

## 1. INTRODUCTION

The paper and presentation on *automatic price collection on the internet* inform the statistical community about advantages and pitfalls of the usage of web scraping technologies for official statistics.

The paper describes in detail technological, data security and legal requirements of web crawlers.

Currently, Statistical Institutes staff members manually collect already a significant amount of data on the internet. The growing importance of online trading requires even more price collection from the internet. Budgetary constraints, however, call for a more efficient deployment of existing human resources to master the additional work load. **Automatic price collection** can, at the same time, support this objective and achieve higher quality price statistics. Nevertheless, legal (permission to crawl on private websites), technological (increased exposure of IT-system to potentially dangerous internet contents), human (need for IT-training), budgetary (implementing and maintenance costs) issues need to be taken into account when deciding on using this new data collection method.

## 2. METHODS

The usage of automatic price collection on the internet as a new data collection method is a two-year project of Statistics Austria's Consumer Price Statistics department. The two-year project is supported by a Eurostat Grant and part of the activities to modernise price statistics [1].

The project is structured in several tasks which is described by the conference paper in detail:

### - Selection of software

The automatic price collection software has been selected according to several criteria.

- The software must provide a high level of *usability*.
- It has to be software that can be easily understood by non-IT price statistics staff members.
- The software should provide a surface that enables users with basic IT knowledge to change the price collection procedure (e.g. in case of website changes).
- The software must provide a well maintained documentation and should be adoptable the internal IT system.

---

<sup>1</sup> Statistics Austria – Consumer Price Index

- Before any implementation, IT-specialists assure that the software is safe to operate and that it comes along with appropriate licensing, testability and supportability.
- A risk analysis assesses the potential legal and data security problems.

#### *-Legal Analysis*

The legal department assesses the national legal framework concerning the jurisdiction on the extraction of online product information for statistical purposes. As a result, the legal requirements are taken note of and a stringent ‘rules of conduct’ for the automatic price collection have to be written and published which transparently describes the methods used to perform web scraping.

#### *- Implementation and maintenance of software and supporting IT infrastructure*

The IT department acquires and installs the selected software. Maintenance procedures to update and test the software regularly and to provide support needs to be set up and documented. Automatic web scraping has been identified as a potential leak for Statistics Austria’s IT-System. In order to avoid viruses, hackers etc. to infiltrate the internal IT-system the web scraper operates within a standalone system on a separate server. Employees access the software and the scraped data by using a remote server from their PC. Furthermore, IT develops and maintains an infrastructure (SQL Database) to store the extracted data.

#### *- Selection of Product Groups and Online Retailers*

In the beginning, Product Groups and Online Retailers are selected according to currently valid manual price collection procedure. This approach facilitates the comparison of the results from automation. In a later step, product groups and retailers not yet in the price index sample will be targeted.

#### *-Development of automatic price collection processes using the selected software*

Price statistics staff use the web scraping software and create automation scripts to continuously download price data from eligible online retailers. This step includes checking the compatibility of the specific extraction methods applied on the selected data-sources (online retailers). *Quantitative* as well as *imitating approaches* are considered. The Quantitative approach aims at continuously harvesting all the available price data from selected websites. The imitative approach collects automatically the data according to criteria, which are currently already applied in the manual price collection. Internet data sources are connected directly to output files (e.g. live databases and reports), the extracted data is analysed and cleaned for price index compilation. In the end, an automatic price collection system will produce data that can be directly used for the production of elementary aggregate price indices. Quality control and price collection supervision as well as changes to the automation scripts are done by price statistics department staff. IT infrastructure and software maintenance (updates) are supplied by IT.

#### *-Development of quality assurance methods*

Part of the quality assurance is the comparison of automatically collected price data with manually collected prices. Later, predefined research routines and consistency checks will be deployed. An optimal method would be the deployment of a second web crawler software whose results could be automatically compared with the results of the first web crawler.

#### *-Usage of automatic price collection for various price statistics*

In order to maximise the output of the investment into automatic price collection, the actions of the project will aim at the inclusion of as many price statistics as possible. Thus, all price statistics projects will cooperate on the development, in particular HICP and PPP, but also other price statistics such as the Price Index on Producer Durables.

### 3. RESULTS

Currently, the pilot project performs all project tasks using the web crawling software *import.to*. The main advantage of the tested software is that no advanced programming skills are needed to perform changes to the web crawling programs in case of website changes.

The success of automatic data collection depends on the ability of the deployed web crawler to simultaneously improve the data quality while reducing the overall data collection costs. Table 1 provides first details on the ability of automatic price collection to achieve these goals :

**Table 1. Comparison – Manual vs. Automatic Price Collection method - Flights**

Method	Product Group	# of Prices	Work load	Comment
Manual	Flights	Ca. 200	16h per month	Ca. 5 min per price
Web crawler	Flights	Ca. 4000	4h+X	X= irregular maintenance work

The monthly working hours spent to collect the prices needed to compile the index for prices on passenger flights can be substantially reduced from 16 to 2 hours. In fact, the actual manual price collection has been completely replaced and the quality of the price index will be higher due to an increased number of measured price quotes. The two hours needed for the automatic price collection method cover various tasks, such as data importing, data cleaning and data checking. In the course of the project, the work load factor X, the irregular maintenance work needed to run the web crawler, has to be assessed and quantified. Maintenance is required when website architecture is changed. There is evidence that the resources needed to perform the irregular maintenance work depends on the individual website and heavily affects the total work load. Thus, a critical cost effectiveness analysis is needed when applying automatic price collection methods.

### 4. CONCLUSIONS

The web crawling technology provides for an opportunity to improve statistical data quality and to reduce the overall workload for data collection. Using automatic price collection methods enables statisticians to react better to the increasing amount of data sources on the internet. Any implementation of the method requires thorough planning in various fields. Legal and data security aspects need to be dealt with in the beginning. Necessary IT resources and IT training required to maintain the automatic data collection system have to be estimated in the course of a pilot project and should not be underestimated.

### REFERENCES

- [1] R. Barcellan, Multipurpose Price Statistics, Ottawa Group (2013), 9. [http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/8bdac0e73d96c891ca257bb00002fdb4/\\$FILE/OG%202013%20Barcellan%20-%20Multipurpose%20Price%20Statistics.pdf](http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/8bdac0e73d96c891ca257bb00002fdb4/$FILE/OG%202013%20Barcellan%20-%20Multipurpose%20Price%20Statistics.pdf)

# Modelling sample data from smart-type meter electricity usage

Susan Williams (susan.williams@ons.gsi.gov.uk)<sup>1</sup>

**Keywords:** smart meters, households, big data, official statistics

## 1. INTRODUCTION

A smart meter is an electronic device that records and stores consumption information of either electric, gas or water at frequent intervals. These data can be transmitted wirelessly to a central system for monitoring and billing purposes.

The European Commission's Energy Efficiency Directive<sup>2</sup> is a common framework of measures for the promotion of energy efficiency within the EU. It supports the EU's 2020 headline target on 20 per cent energy efficiency, and its provision for the roll-out of smart meters requires member states to ensure that at least 80 per cent of consumers have intelligent electricity metering systems by 2020.

The separate countries of the UK collectively have one of the most ambitious roll-out policies within the EU: to put electricity and gas smart meters in every home by 2020 with rollout starting in 2016.

Smart meter electricity energy usage data is attractive to statistical organisations as, subject to data access, it provides detailed information on household energy consumption and at high levels of frequency.

Preliminary research in the UK's Office for National Statistics has focused on the potential of smart-type meter electricity usage data, made available for research through energy trials, to model days where a household may be unoccupied. As a retrospective analysis using actual smart meter data, this information may have a use in the quality assurance of census data. An extension of the research is to investigate whether long-term unoccupied households may be identified, of great use within validating estimates of such properties or to aid with processes such as survey and census fieldwork.

## 2. METHODS

Data was sourced from consumer behaviour trials of smart-type meters conducted in Ireland and held in the Irish Social Science Data Archive<sup>3</sup>. These data include 30 minute frequency electricity energy usage readings on over four thousand households over 18 months in 2009-2010. These data contain 108 million observations.

The team tested various methods for predicting whole days when a household may be unoccupied. As there is no information on whether a property is actually occupied or not, the accuracy of the methods to detect unoccupied households was conducted by examining energy profiles over time by eye. If electricity consumption is fairly flat across 24 hours then the assumption is that the property is unoccupied on that day.

---

<sup>1</sup> Office for National Statistics

<sup>2</sup> [http://ec.europa.eu/energy/efficiency/eed/eed\\_en.htm](http://ec.europa.eu/energy/efficiency/eed/eed_en.htm)

<sup>3</sup> <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/>

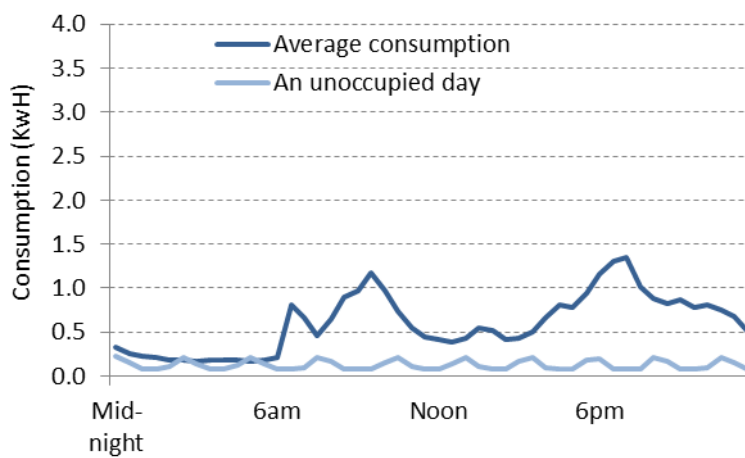


Correspondingly, if there is evidence of unsystematic variation in energy usage, particularly associated with either morning and/or evening peaks, then the property is assumed to be occupied. Validation of the methods in this way is therefore subjective; one person may think that the household is occupied based on the consumption data, another person may not.

### 3. RESULTS

#### 3.1. What electricity consumption pattern does an unoccupied household have?

The graph below shows the mean daily electricity consumption pattern during the trial period, as well as the consumption pattern for what is thought to be an unoccupied day for one sampled household.



**Figure 1: Half hourly consumption for a selected household**

The mean daily consumption over the trial period for this household is typical in that both a morning peak and an evening peak can be observed with a dip in the middle when household occupants may be temporarily absent from the home. The unoccupied day has a regular cyclical pattern of electricity consumption, typical of appliances driven by automated controllers such as a fridge or freezer. This is the pattern that has been most often observed in what have been classed as unoccupied households during this research.

#### 3.2. Overview of results

The methods developed used various attributes of energy expenditure on a given day such as the daily average, night time average and variance as well as looking at differences in these values to the 'usual' energy expenditure in a household. Although some of the methods appear to classify unoccupied days reasonably well on their own and with little misclassification, it is conjectured that using methods in combination might further improve classification.

The research is being extended to look at combining methods as well as exploring options to identify longer term vacant households, although the trial data is limited here as it does not contain many such households. With further research to refine the most promising methods and to identify the most influential variables it is proposed that logistic regression might also be used.

#### **4. CONCLUSIONS**

This abstract outlines the research undertaken into the potential of using the data from trials of smart-type electricity meters to help improve official statistics.

The dataset used contains half hourly electricity consumption of over four thousand households over an 18 month period. It was used to develop initial methods to determine the likelihood of whole days when households appear to be unoccupied.

The data on which these methods were tested is large and required processing using big data technologies. Manipulation of similar data representing all households in England will require significant knowledge of a range of big data technologies.

# From the construction to the usage of statistics beyond GDP

Marina Signore ([signore@istat.it](mailto:signore@istat.it))<sup>1</sup>, Donatella Fazio ([dofazio@istat.it](mailto:dofazio@istat.it))<sup>2</sup>

**Keywords:** progress indicators, future research needs, policy use

## 1. INTRODUCTION

The GDP and beyond debate has reached some stable points on the construction of indicators for well-being, societal progress and sustainability. Research, pushed by the increasing discussion over the last fifteen years, has broadly converged in terms of methodology and techniques for an integrated framework of measures able to represent a “real tool” to explain the phenomena beyond GDP. The European and global research has grown thanks to the various initiatives carried out since the beginning of the new millennium by the multiple actors involved in the production and utilisation of statistics. The European Statistical System (ESS), together with OECD, UNECE and other relevant partners from academia and the research world, has been deeply involved. With the Sponsorship Group on ‘Measuring Progress, Well-being and Sustainable Development’ [1], the ESS has called the National Statistical Institutes (NSIs) to outline a strategy to develop statistical information to meet the Stiglitz’s Commission recommendations [2] and then with more recent initiatives is pushing for their implementation focusing also on the understanding and usage of the new measurements.

Among the various European initiatives recently carried out, the EU FP7 project e-Frame *European Framework for Measuring Progress* represents an important effort carried out by NSIs jointly with relevant stakeholders involved in the new measurements beyond GDP.

## 2. E-FRAME PROJECT

The e-Frame Project ([www.eframeproject.eu](http://www.eframeproject.eu)), funded by the European Commission (EC), DG Research and Innovation, (duration 2012-2014), relied on a 19-partner consortium, chaired by the Italian National Institute of Statistics (Istat) and Statistics Netherlands (CBS). The consortium had a pan-European dimension and involved four main European National Statistical Institutes (Italy, the Netherlands, France and United Kingdom), two civil society organisations, seven academia and five research centres as well as the OECD giving an international perspective. To suit the EU request, e-Frame was structured into a work plan able to face the issues related to methodological and theoretical aspects of the measurement of societal progress and wellbeing. Several initiatives were envisaged in order to spread knowledge, with particular attention to the European level, on large parts of the issues related to the measurement of wellbeing and progress. Sustainable development, subjective wellbeing, social capital, human capital and labour market, intangible assets, new national account architecture, environmental indicators and the welfare effects of globalisation are some of the themes focused by the nine thematic workshops and the two conferences (one initial and the other final) held by the project. Moving from the stocktaking on existing indicators and measurements, e-Frame envisaged cross-cutting activities with the aim to provide guidelines for the use of the indicators by policy makers and to define a roadmap for addressing future research needs at the European level. With the ultimate ambition to state a European position on the measurement of well-being and progress able to interact at a global level, the

---

<sup>1</sup> Istat, The Italian National Statistical Institute

<sup>2</sup> Istat, The Italian National Statistical Institute

European Network on Measuring Progress (e-FrameNET), with a dedicated section on Wikiprogress platform hosted by OECD, was established as an off-shot of the project.

### 3. RESULTS

Fostering the European position and contributing to setting the future EU agenda are the two streams in which the outputs of e-Frame project were embedded. Against it, a *Roadmap for future research needs* for the development, the understanding and the usage of the indicators as well as a *Map of policy use of Progress indicators* are the main policy outcomes of the 30-month activity of the project.

#### 3.1. A Roadmap for future research needs

The Roadmap [3] is a cross-cutting deliverable that has benefited from the results of all the activities carried out by the project. To this effect the Roadmap has collected: feedbacks from the stocktaking and dissemination activities (workshops and conferences) envisaged by the work plan and others; contacts with the project partners for their suggestions and comments; suggestions from e-Frame Advisory Board; contacts with and feedbacks from similar work underway at national and international level through the e-FrameNET - European Network on Measuring Progress; comments and suggestions from the EC. The Roadmap was conceived to address relevant gaps and research needs to be put at the centre of future research agenda at a European level by the European Commission and by the European Statistical Systems in the area of GDP & Beyond. To permit an easier reading of the Roadmap, the research needs have been grouped into four main streams even if it is necessary to consider the limit of cataloguing these issues that are by nature and scope interrelated and partly overlapping.

The first stream reports the research needs on *Measurement issues in official statistics*. The ESS, comprising Eurostat and the EU NSIs, has worked hard to adopt the recommendations of the Sponsorship Group on ‘Measuring Progress, Well-being and Sustainable Development’, but a lot has still to be done in order to develop better statistics and concise indicators. This brings to recognise the needs for widening and strengthening the official statistical production and targeting at a better harmonisation and standardisation of concepts and indicators. Fostering the bottom-up approach initiatives of stakeholder’s consultations and the dialogue with the society at large supported by the Web 2.0 tools would also contribute to put the program into concrete. Despite the great amount of work already undertaken, there exist information needs that are not fully met by official statistics yet: continue implementing subjective indicators; report indicators at different levels (local, national, global); disaggregate at the right dimension (e.g. target groups); harmonize concepts, standards and definitions and provide metadata; improve the timeliness of data; improve indicators of sustainability; increase micro data availability; analyse quality implications for well-being measurement; train at University level on Official Statistics on measuring progress.

The second group is related to the *Exploitation of non-official sources*. The demand for measuring progress, well-being and sustainable development in a multi-dimensional way needs “more” information and data. It brings to the necessity to exploit the integration and complementarity of official data with non-official data. Non-official sources can cover product areas and sectors excluded from official sources, filling important data gaps: bridge top-down and bottom-up approaches for the construction of the statistics; foster the usage of non-official data; exploit crowd sourced locally generated data; integrate and complement official data with non-official data; develop technologies for the use of big data and open data; evaluate the role for non-official data in a cost-benefit perspective; validate the usage of non-official data in a “new” quality framework; label the non-official data to be clearly distinguished from the official ones.

The following group deals with the *Communication issues*. Communication and dissemination strategies are crucial to enforce the awareness of the importance of “better statistics” on well-being and societal progress for all the stakeholders. In particular an effort has to be driven towards the business world not used yet to read figures different from GDP. Communication and dissemination strategies can enable and encourage public debate and provide decision makers with a wealth of information to develop more informed policies. To this effects data visualisation, including new social media platforms combined with traditional media, need to be enhanced and brought up-to-date. An indicative list of the research needs for Communication is the following: facilitate the communication to policy makers; find a wording for policy makers; develop Web2.0 tools to improve the understanding of progress statistics; exploit the digital initiatives carried out by communities for statistics beyond GDP; foster the culture on the measurement of well-being; educate the opinion leaders and journalists; inform and train the business world and educate the market to read new measurements as a chance to catch for their business; enrich statistics with storytelling.

The last group refers to *A policy use of progress indicators*. There are inter-linkages between the various progress indicators. The different domains constitute a “spider net” while policy makers are used to working in a specific domain ignoring the critical role of other sectors. Work should be done to look forward to a wide integrated vision identifying the main drivers of well-being and to assess the overall impact of alternative policy options on people’s life. The main needs listed on this side are: develop models capable to describe the trade-off between different dimensions and simulate the various effects; develop an integrated framework for a policy use of progress indicators; evaluate the “Institutional” Sustainability for a well-being oriented policy; develop risk indicators to measure the social effect of not doing; construct econometric models of simulation to measure the effect of the policies on well-being; study how the use of well-being indicators positively influences the (good) policies; develop a beyond GDP narrative.

### **3.2. A Map of policy use of progress indicators**

The Map [4] aims at providing policy makers with ideas and tools that can guide and support their policies. It stands at the forefront of the debate: despite a rich literature, progress indicators are not yet part of the political action, except for few exceptions. Thus, the Map has the ambition to contribute to the current debate and to further promote policy use of progress indicators by reporting most recent advancements in measuring well-being and sustainable development, presenting successful experiences of policy use of indicators, discussing existing gaps, proposing improvements and suggesting recommendations for use. From a content viewpoint, it moves along the border between methodological issues and policy use. As the different chapters show, from e-Frame work it emerged that: i) for some subject areas the relevant measures are not fully developed yet and some research is still needed to produce the required indicators; ii) for many other subject areas measures are currently produced by official statistics but they are often ignored by policy making and iii) finally, there exist good practices of policy making which already moved “beyond GDP” that are reported as well. The Map collects contributions from all e-Frame activities that are divided into three main sections: i) measuring well-being and societal progress; ii) methodologies and tools for measuring well-being and societal progress; iii) integrated policy frameworks. The initial chapter “Policy use of progress indicators” provides a general overview of the state-of-the-art on the actual use of beyond GDP indicators.

#### 4. CONCLUSIONS

The debate generated by e-Frame project and its main policy outcomes, the *Roadmap for future research needs* and the *Map of policy use of Progress indicators*, have identified the stable points reached on GDP & Beyond so far, as well as the open issues to be addressed in future research agendas. Acknowledging the complexity and the multidimensional essence of well-being, they represent policy oriented tools supporting an effective usage of progress indicators by stakeholders and policy makers. On the communication side it is necessary to keep on working hard to increase the awareness of the importance of thinking in terms of an integrated framework in order to drive policies for optimizing the “whole” in balance with the maximisation of the single domain. e-Frame has undoubtedly contributed to the European and global debate pushing for the empowerment of the interaction among stakeholders and citizens for a change of mentality and behaviour related to the new measurements GDP & Beyond.

The experience of the EU funded project has proven to be successful in putting together the relevant expertise and the different stakeholders’ viewpoints. Once more, cooperation among different institutions and organizations has allowed to gain efficiency and reach more stable results. As suggested by the Roadmap, further development projects can be launched in order to fill in some existing gaps. As available resources are becoming more and more limited, it would be really important to prioritise further research investment with regard to the expected outcomes and European agendas’ targets.

#### REFERENCES

- [1] ESS (2011). Final Report on the Sponsorship Group on Measuring Progress, Well-being and Sustainable Development. Available at : [http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp\\_ess/0\\_DOCS/estat/SpG\\_Final\\_report\\_Progress\\_wellbeing\\_and\\_sustainable\\_deve.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SpG_Final_report_Progress_wellbeing_and_sustainable_deve.pdf)
- [2] Stiglitz et al. (2009). Final report by the Commission on the Measurement of Economic Performance and Social Progress. Available at: [http://www.stiglitz-sen-fitoussi.fr/documents/rapport\\_anglais.pdf](http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf).
- [3] M. Signore, D. Fazio, M.G. Calza (2014). The roadmap for future research needs, Deliverable 11.2 European Framework for Measuring Progress project, A collaborative project funded from the European Union’s Seventh Framework Programme under grant agreement no 290520, Istat. Available at <http://www.eframeproject.eu/fileadmin/Deliverables/Deliverable11.2.pdf>
- [4] T. Rondinella, M. Signore, D. Fazio, M.G. Calza, A. Righi (2014). Map on policy use of progress indicators, Deliverable 11.1 European Framework for Measuring Progress project, A collaborative project funded from the European Union’s Seventh Framework Programme under grant agreement no 290520, Istat. Available at: <http://www.eframeproject.eu/fileadmin/Deliverables/Deliverable11.1.pdf>

# The potential of Web2.0 communities for statistics

Donatella Fazio ([dofazio@istat.it](mailto:dofazio@istat.it))<sup>1</sup>, Katherine Scrivens ([Katherine.SCRIVENS@oecd.org](mailto:Katherine.SCRIVENS@oecd.org))<sup>2</sup>, Maria Grazia Calza([calza@istat.it](mailto:calza@istat.it))<sup>3</sup>

**Keywords:** new sources of data, crowd sourcing, IT challenges, bottom-up

## 1. INTRODUCTION

The Internet revolution has opened huge new opportunities for the construction and reshape of better statistics. The *liquid data* available in the Net - Open Data, Big Data and crowd sourced data - are being explored with the aim to exploit these new sources of data to integrate and complement the official statistics. The usage of new sources of data brings the National Statistical Institutes (NSIs) and the various stakeholders to face challenges moving from traditional to new ways of data collection and production, considering the opportunities given by Web2.0 tools, such as social networks and on line platforms. It is undeniable that the usage of these new sources drive relevant implications for issues such as validation, metadata, methods and techniques, including IT tools to allow combining user-generated data with the data produced by official statistics. The era of “Statistics 2.0” [1] arises the trade-off between having more and real-time information and the quality of the information produced. The exploitation of new sources of data implies a big investment in producers and users’ skills to combine information coming from different sources and brings to consider crucial issues such as privacy aspects, acceptance of data re-use and the management and protection of the data. The European Statistical System (ESS), jointly with relevant stakeholders, together with producers and users of statistics, is increasingly involved in projects and research activities to set up methodologies and procedures to add information and knowledge in a systematic integrated way exploiting new sources of information and data.

Among the multiple initiatives recently carried out at European level, it stands the on-going project Web-COSI *Web Communities for Statistics for Social Innovation*, led by the Italian National Institute of Statistics (Istat) - funded by the EU FP7 ICT Work Programme 2013 - to explore the potential of Web2.0 communities focusing on statistics beyond GDP.

## 2. WEB-COSI PROJECT

Web-COSI was designed to respond to the call launched by the Collective Awareness Platforms for Sustainability and Social Innovation (CAPS), based on the idea that collaborate through crowd sourced platforms can produce solutions for a wide range of social needs [2]. Web-COSI ([www.webcosi.eu](http://www.webcosi.eu)), is a co-ordination action (2014-2015) with the general objective to foster the engagement of citizens and society at large in the area of new measures of societal progress and well-being using the opportunities given by Web 2.0. Specific objective is to implement tools for collecting, producing and visualizing information and data towards a better integration of official and non-official statistics. The release of a Wiki of progress statistics, at mid-term of the project, is envisaged with the aim to foster the use of locally generated data to bridge top-down and bottom-up approaches. Web-COSI capitalizes on the last 15 years’ experience characterized by two big revolutions for the world of research and society. First, the

---

<sup>1</sup> Istat, The Italian National Statistical Institute

<sup>2</sup> OECD, The Organisation for Economic Co-operation and Development

<sup>3</sup> Istat, The Italian National Statistical Institute



“GDP & Beyond” debate that has dominated the scene of statistical and economic research. The ESS [3], together with OECD [4] and other relevant actors, has been deeply involved to outline a strategy to develop statistical information to meet the Stiglitz’s Commission recommendations [5]. Second, the Internet explosion that has radically changed the way in which information is produced and shared. Interactivity is contributing to change the roles of producers and users of data, increasing awareness and bringing to consider a bottom-up approach for the construction of statistical information. Web-COSI is based on a consortium that sees the collaboration among two relevant Institutions (Istat and OECD), a civil society organisations ([www.lunaria.org](http://www.lunaria.org)) and a social entrepreneurs’ community ([www.i-genius.org](http://www.i-genius.org)), representing society at large. The consortium is well-balanced and multidisciplinary creating synergies for the integration of the different approaches.

Specifically, Web-COSI work plan aims at: a) mapping and distilling best practices of existing digital initiatives for communities’ involvement - a specific survey is envisaged to take stock of Web2.0 initiatives carried out or planned by NSIs; b) create a critical mass through: target campaigns, data visualization competitions, setup of a European Wikiprogress University Programme; c) facilitate the communities’ access to statistics empowering the collection of civil society grass root locally generated data with the development of a Wiki of progress statistics. Moreover, various open events are organised to involve the greatest number of audience: 5 workshops, 4 focus groups and a final conference.

### **3. RESULTS SO FAR**

To map and distil the best practices of existing digital initiatives for communities’ involvement Web-COSI has carried out in 2014: i) two on-line discussions organized by OECD– “*Engaging citizens in well-being and progress statistics*” and “*Making data more accessible for society at large*”; one webinar managed by Lunaria - “*Civil society engagement in well-being statistics: good practices from Italy*”; iii) a workshop on *Using Technology to Engage Citizens with Well-being Statistics in the Perspectives from Civil Society* held at OECD. These activities have involved a large number of participants from different sectors of society, including NSIs, government, research organizations, social enterprise and civil society, generating an impressive debate on the potential of Web2.0 communities. The discussions helped to identify an initial map of the different types of initiatives set up, using collective platforms, to engage citizens with well-being and societal progress statistics. The different types of initiatives can be grouped as follows: (i) public consultation; (ii) communication; (iii) citizen-generated data, (iv) open data.

#### **3.1. Public consultation**

Consulting with the public is now widely seen as an essential step in elaboration of indicators of well-being and progress [6]. While public consultation is the type of citizen engagement most closely associated with the new measures of well-being and progress, it is an area where the potential of interactive technology is still to be fully realised. For many in the discussion, especially those working on smaller-scale community projects, citizen engagement through consultation is more suited to face-to-face events such as focus groups or community meetings, while online methods were seen as more appropriate for the collection and communication of the data. However, face-to-face events are necessarily limited in terms of representativeness, which is an especially important problem for well-being measurement projects at the national level. The UK’s ‘Measuring National Well-being’ programme, managed by the Office for National Statistics, Italy’s ‘Equitable and Sustainable Well-being’ (BES) project, led by Istat and ‘Measures of Australia’s Progress’ (MAP), run by the Australian Bureau of Statistics, all



used a mix of offline events and surveys, with online consultation tools such as online surveys and social media to reach as wide an audience as possible.

### **3.2. Communication**

Finding innovative ways to communicate the underlying meaning of data by telling a story around the data (or by enabling users to play with the data and find their own stories) is a powerful way of making statistics more accessible to a broader audience. This can be done by the data producers themselves (such as government or statistical agencies) or by intermediaries such as data journalists, civil society organisations or anyone with an interest in finding the best way to communicate the key messages of datasets. Stories can be told in the traditional way, through narrative text, or they can be conveyed in a more visual manner - through infographics and charts that organise the data in such a way that the meaning is immediately apparent. Data visualisations can be very appealing, but their importance goes beyond aesthetics: they provide a unique means of highlighting new patterns in statistics and looking at the world in a different way. The mapping exercise highlighted many innovative examples of visualisation, particularly from civil society and international organisations. NSIs also recognise the need to make their data more accessible through visualisation and most of them have made some provision of interactive data content on their websites, be it in the form of tables, charts, maps or dashboards. However, the quality and amount of data made available in this way is extremely variable [7].

### **3.3. Citizen-generated data**

Digital technologies allow members of the public to participating themselves as data producers and the prevalence of accessible yet sophisticated mapping technology through mobile platforms provides a means to crowd-source data from members of the public at minimal cost. Geographic Information Systems (GIS) allow for users to provide data in the form of Tweets, reports, photos, comments, or other types of Volunteered Geographic Information (VGI), that allow for the monitoring of outcomes related to well-being in close to real time. There are a number of different ways that platforms for citizen-generated data can function, including public reporting (e.g. of problems in their local area), monitoring of social media (to gauge public reaction to events and policies), and through the use of citizen scientists (e.g. individuals with specialised skills, hobbies, or interests can be recruited to act as data sensors to help populate scientific research databases). Given the limitations in coverage and timeliness of official statistics in many developing countries, encouraging the development and use of new forms of data collection has been seen as a core element of the data revolution needed to monitor progress towards the Sustainable Development Goals [8]. However, while citizen-generated data have a lot of potential for providing useful information and filling data gaps, they also have significant limitations when compared to official statistics related to self-selecting samples, reliability of data and comparability between areas.

### **3.4. Open data**

For data to be truly open, not only must it be freely available online, but it should also be presented in a format that maximises its potential for re-use, with semantically tagged information, open formats, and fully downloadable information, the latter through APIs (Application Programming Interfaces), machine-readable data structures and rich metadata. For many organisations, to whom open data is a new concept, this is likely to be a gradual process, requiring significant resources. The role of advocacy organisations

will be important in order to educate government and civil society of the need to engage more with citizens through open data.

#### 4. CONCLUSIONS

So far Web-COSI activities have pointed out that Web2.0 technologies are exploited – at different extent - by NSIs, government, research organizations, social enterprise and civil society to foster the interaction between data producers and data users of statistics. The discussion has identified that the usage of crowd sourced data to complement and integrate official statistics, is an opportunity to evaluate in multiple terms: having data close to real time; narrowing the distance between what official statistics say and what people perceive; using new data not included in the official surveys; optimising the costs. The usage of crowd sourced data has to be considered at different levels giving them different weights: local; national; international; global. The discussion pointed out the necessity of organizing the progress statistics (generated by official and non-official data) in an integrated framework to represent a “real tool” to make statistics accessible and understandable.

Web-COSI will continue its work to conclude its activities by the end of 2015. Next steps envisage: i) the conduction of a specific survey addressed to NSIs on Web2.0 initiatives carried out and/or planned to empower statistics using new sourced of data; ii) the organization of target citizens campaigns, data visualisation competitions, youth initiatives, and the set-up of a European Wikiprogress University Programme; iii) the organisation of 3 workshops, 4 focus groups of social entrepreneurs and a final conference. At last but not at least, the development of a new data sharing portal - Wiki of progress stat on [www.wikiprogress.org](http://www.wikiprogress.org) - designed to be a key reference for progress and wellbeing data and statistical resources such as reports, visualisations and interactive tools, able to allow external data providers to upload their own data.

Web-COSI experience is demonstrating that the integration of traditional official statistics with new sources of data is an inexorable process which requires new skills, culture and a radical change of mind set. In Web 2.0 era as the power of online communities grows ever stronger institutions of diverse type and scope cannot ignore their centrality for the “definition” of better statistics, for better policies, for a better quality of life.

#### REFERENCES

- [1] Giovannini E. (2010). “Statistica 2.0: the next level”, Introductory speech at the 10th National Conference of Statistics, Rome. Available at: [http://www3.istat.it/dati/catalogo/20120621\\_00/atti\\_decima\\_conferenza\\_nazionale\\_statistica.pdf](http://www3.istat.it/dati/catalogo/20120621_00/atti_decima_conferenza_nazionale_statistica.pdf)
- [2] Sestini F. (2012). Collective Awareness Platforms: Engines for Sustainability and Ethics. IEEE Technol. Soc. Mag. 31(4) pp. 54-62 Available at: <http://caps2020.eu/wp-content/uploads/2013/11/CollectiveAwarenessPlatformsEngineforSustainabilityandEthics-1.pdf>
- [3] ESS (2011). Final Report on the Sponsorship Group on Measuring Progress, Well-being and Sustainable Development. Available at : [http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp\\_ess/0\\_DOCS/estat/SpG\\_Final\\_report\\_Progress\\_wellbeing\\_and\\_sustainable\\_deve.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SpG_Final_report_Progress_wellbeing_and_sustainable_deve.pdf)
- [4] OECD (2012). Better Life Initiative: Measuring Well-Being and Progress. Available at: <http://www.oecd.org/statistics/betterlifeinitiativemeasuringwell-beingandprogress.htm>

- [5] Stiglitz et al. (2009). Final report by the Commission on the Measurement of Economic Performance and Social Progress. Available at: [http://www.stiglitz-sen-fitoussi.fr/documents/rapport\\_anglais.pdf](http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf).
- [6] Hall, J. and L. Rickard (2013). People, Progress and Participation: How Initiatives Measuring Social Progress Yield Benefits Beyond Metrics, Global Choices 1: 2013, [http://www.bertelsmann-stiftung.de/cps/rde/xbcr/SID-CFCC0351-B1427148/bst\\_engl/xcms\\_bst\\_dms\\_37947\\_37948\\_2.pdf](http://www.bertelsmann-stiftung.de/cps/rde/xbcr/SID-CFCC0351-B1427148/bst_engl/xcms_bst_dms_37947_37948_2.pdf)
- [7] Leib, S. and Z. Hartland (2013). ICT Delivering Tools: Catalogue of user tools and discussion relating to implementation, e-Frame Deliverable 8.1, [www.eframeproject.eu](http://www.eframeproject.eu)
- [8] The Engine Room (2014). CIVICUS Support to People-powered Accountability and the Data Revolution: A scoping Study by the Engine Room, <https://www.theengineroom.org/wp-content/uploads/CIVICUS.Scoping.Study.WEB.pdf>

# Recent Advances in the Measurement of Intangible Assets.

Mary O'Mahony ([mary.omahony@kcl.ac.uk](mailto:mary.omahony@kcl.ac.uk)), King's College London<sup>1</sup>, Carol Corrado<sup>2</sup>, Jonathan Haskel<sup>3</sup> and Cecilia Joan Lasinio<sup>4</sup>

**Keywords:** Intangible Investments, Productivity, Growth

## 1. INTRODUCTION

There is an extensive literature that highlights the importance of investment in intangible assets for understanding the drivers of economic growth. Key to this is the idea that organisational changes and other forms of intangible investment such as workforce training are necessary to gain significant productivity benefits from adopting new technologies (e.g. Bertschek and Kaiser, 2004; Black and Lynch, 2001; Bresnahan, Brynjolfsson and Hitt 2002). Given this, a measurement exercise was instigated to estimate the impact of intangible investments as a source of growth. This involved researchers across a wide range of institutions and countries, with much of the research effort funded by European Commission framework grants. The main projects were COINVEST, INNODRIVE, IAREG and INDICSER for business sector intangibles with a currently running project, SPINTAN, addressing the measurement of intangibles in the public sector. In addition a voluntary initiative of the research community was the INTAN-INVEST project which constructed a harmonised dataset on intangible capital investments merging datasets from the INNODRIVE and COINVEST projects with the underlying data from the Conference Board for the US (Corrado et al., 2009).<sup>5</sup>

## 2. METHODS

Estimating intangible investments required an identification of types of assets to include as intangible investments and methods to capitalise these investments. The pioneering work on measuring expenditures on intangible assets by businesses was Corrado, Hulten and Sichel (2005). These authors (CHS) identified a number of types of expenditure which they argued should be treated as investments rather than as intermediate expenditures. Three main categories of assets were identified by CHS (2005): computerised information, innovative property and economic competencies. Computerised information basically coincides with computer software. Innovative property refers to the innovative activity built on a scientific base of knowledge as well as to innovation and new product/process R&D more broadly defined. Economic competencies include spending on strategic planning, worker training, redesigning or reconfiguring existing products in existing markets, investment to retain or gain market share and investment in brand names.

In addition to constructing nominal investment series, the research had to decide on appropriate deflators to convert to volume measures and on the form and rates of

---

<sup>1</sup> King's College London

<sup>2</sup> The Conference Board

<sup>3</sup> Imperial College London

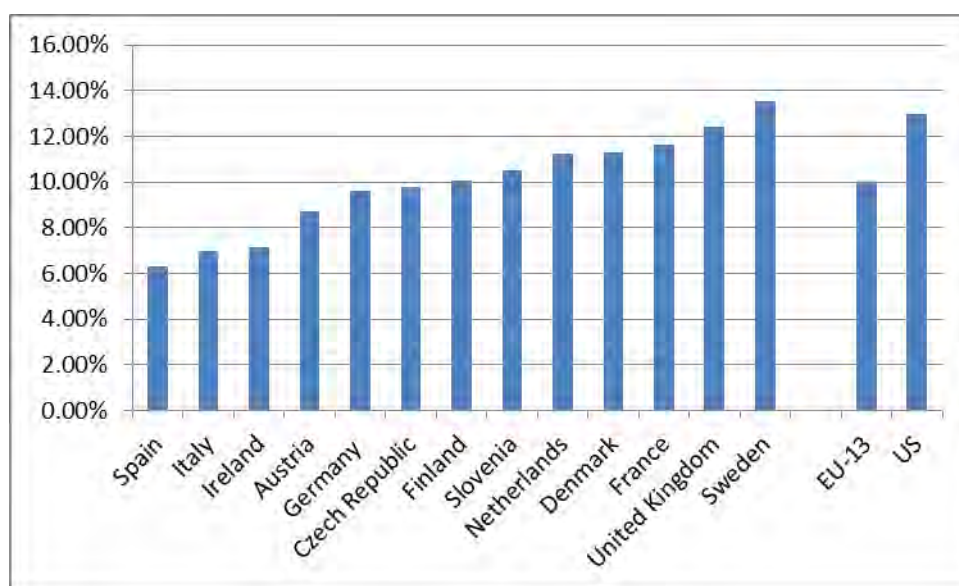
<sup>4</sup> LUISS and Istat

<sup>5</sup> Information on these projects can be found on the following: [www.coinvest.org.uk](http://www.coinvest.org.uk); [www.innodrive.org](http://www.innodrive.org); [www.iareg.org](http://www.iareg.org); [www.indicser.com](http://www.indicser.com); [www.spintan.net](http://www.spintan.net); and [www.intan-invest.net](http://www.intan-invest.net).

depreciation to capitalise these assets. GDP deflators were generally employed due to lack of information on asset-specific deflators. The geometric deprecation rates were generally assumed to be very high, as many of these assets are short lived.

This research effort combined produced annual data series from the mid-1990s to recent years for a large number of EU countries, and the US. Parallel academic exercises also constructed estimates for countries such as Canada, Japan and Korea. The researchers involved both benefited from and fed into initiatives in national statistical offices to include some intangible assets into the national accounts, most important being software and scientific R&D. However the academic research effort was much broader in scope, including many assets not currently included in the national accounts.

### 3. RESULTS



**Figure 1. Business Intangible Investment as a percent of GDP (average 1998-2005). Source: INNODRIVE (2011).**

The key result is that investment in intangible capital by businesses is sizeable and is able to explain a significant share of labour productivity growth. Results from the COINVEST project find that the US and UK invested around 13 percentage points of GDP in business intangible capital. However as Figure 1 illustrates, although the average EU investment is less than the US (9.9% vs. 13%), investments are still sizeable and range from 6% in Spain to 13.5% in Sweden. France, Denmark and the Netherlands also invest heavily in Intangible assets. In contrast Spain and Italy invest significantly lower shares in intangible capital compared to the US. Interestingly, this low level of investment in intangible capital is equally driven by lower investments in software, innovative property and economic competencies.

The high shares of intangibles shown in Figure 1 imply that business sector investment in intangible capital within the EU increases total investment significantly, indicating that the real level of investment in the EU is significantly higher than traditionally measured. Some EU countries are already at the threshold of investing similar amounts of intangible capital as tangible capital investments, e.g. France and Sweden. In contrast, countries from the Mediterranean and transition countries tend to invest significantly higher shares into tangible capital. Also the EU has to still catch-up significantly in order to reach the same ratio of intangibles to tangible investments as in the US. The implication of these results for sources of growth is that once accounting for business intangibles, capital

rather than total factor productivity becomes the dominant source of growth (Roth and Thum, 2013).

#### 4. CONCLUSIONS

The inclusion of intangible assets in national accounts has a significant impact on the levels of GDP. As highlighted by the various research findings the process of incorporating business intangibles into the asset boundary will have significant policy implications due to the fact that investments in intangible capital shows significant variation across countries. After accounting for business intangibles, with Italy and Spain being endowed with significantly lower levels of business intangibles than France and Germany, the already large macroeconomic disparities within the euro area will become even more distinct. It thus seems to be imperative to increase the level of investments in business intangible capital in countries facing low investment, such as Italy and Spain.

Empirical evidence from the EC funded projects highlights that growth of intangible capital services is able to explain the largest share of labour productivity growth within a European country sample. Research also indicates that intangible assets are important in facilitating innovation and the adoption of new technologies. Refining the measurement of these assets is crucial for evaluating comparative growth performance. This requires strong collaboration between the statistical and research communities, more regular updating of datasets produced as a by-product of academic research and a concerted effort to produce timely data for policy purposes.

#### REFERENCES

- [1] Bertschek, I., and Kaiser, U. (2004), 'Productivity effects of organisational change: Microeconometric evidence', *Management Science*, 50(3): pp 394-404.
- [2] Black, S., E., and Lynch, L., M. (2001), 'How to Compete: The Impact of Workplace Practices and Information Technology on Productivity', *The Review of Economics and Statistics*, 83(3): pp 434-445.
- [3] Bresnahan, T. F., Brynjolfsson, E., and Hitt, L.M. (2002), 'Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence', *Quarterly Journal of Economics*, 117 (1): pp 339 – 376.
- [4] Corrado, C., C. Hulten, and D. Sichel (2009). 'Intangible Capital and U.S. Economic Growth', *Review of Income and Wealth* 55: 661–85.
- [5] Corrado, C., C. Hulten, and D. Sichel (2005). Measuring Capital and Technology: An Expanded Framework. In: C. Corrado, J. Haltiwanger, and D. Sichel (eds) *Measuring Capital in the New Economy*, Chicago, IL: University of Chicago Press, 11-46.
- [6] INNODRIVE (2011). INNODRIVE Intangibles Database, (<http://www.innodrive.org/>).
- [7] Roth, F. and A. Thum (2013). Intangible capital and labor productivity growth—Panel evidence for the EU from 1998-2005, *Review of Income and Wealth* 59: 486-508.

# CORE: a concrete implementation of the CSPA architecture

Mauro Bruno (mbruno@istat.it)<sup>1</sup>, Rolando Duma (duma@istat.it)<sup>1</sup>, Monica Scannapieco (scannapi@istat.it)<sup>1</sup>, Marco Silipo (silipo@istat.it)<sup>1</sup> and Giulia Vaste (vaste@istat.it)<sup>1</sup>

**Keywords:** CSPA, Statistical Services, SOA

## 1. INTRODUCTION

The modernization of statistical organizations can be achieved through the adoption of an Enterprise Architecture, which aims to identify business needs, improves collaboration across an organization and ensures that the technology is aligned to the strategic vision [1].

In this context the development of frameworks for sharing information, tools and methodologies among statistical organizations is of utmost importance. CSPA (Common Statistical Production Architecture) provides “a framework, including principles, processes and guidelines, to help reduce the cost of developing and maintaining processes and systems and improving the responsiveness of the development cycle. Sharing and reuse of process components will become easier - not only within organizations, but across the industry as a whole.” [2]. CSPA aims to become a reference architecture based on: i) existing standard models, such as GSBPM and GSIM, as the necessary shared industry vocabulary [3, 4]; ii) the “plug and play” approach in designing, implementing, sharing and reusing statistical software solutions, largely based on the “service-oriented architecture” model (SOA).

In this paper, we present a concrete implementation of the CSPA architecture: CORE (Common Reference Environment) [1]. CORE is a platform for integration and automation of statistical services and processes, re-engineered according to CSPA principles. The need to move from a stovepipe model to a new business model based on a SOA, makes CORE a strategic pillar in the Institute innovation process, that lays its foundation on the Enterprise Architecture [6]. Therefore we expect CORE to become the reference tool in various business departments.

In the following section the steps required to build a statistical process in CORE will be described: i) abstract definition of the involved services; ii) link of each service to its actual implementation; iii) design of the process as a collection of interacting services; iv) configuration of input parameters; v) execution of the statistical process. In order to show how CORE works and how it can improve efficiency, knowledge sharing and statistical process automation, a case study (Residential care facilities) will be presented.

## 2. THE CORE PROJECT

CORE is a web application that implements the “plug and play” CSPA architecture. In order to build a statistical process in CORE, the following steps have to be performed (as shown at the top of Figure 1):

- 1) *Define*: in this phase it is possible to define a CSPA compliant statistical service. The system provides two separate functionalities which allow defining both a Statistical Service Definition [2], a conceptual representation of a statistical

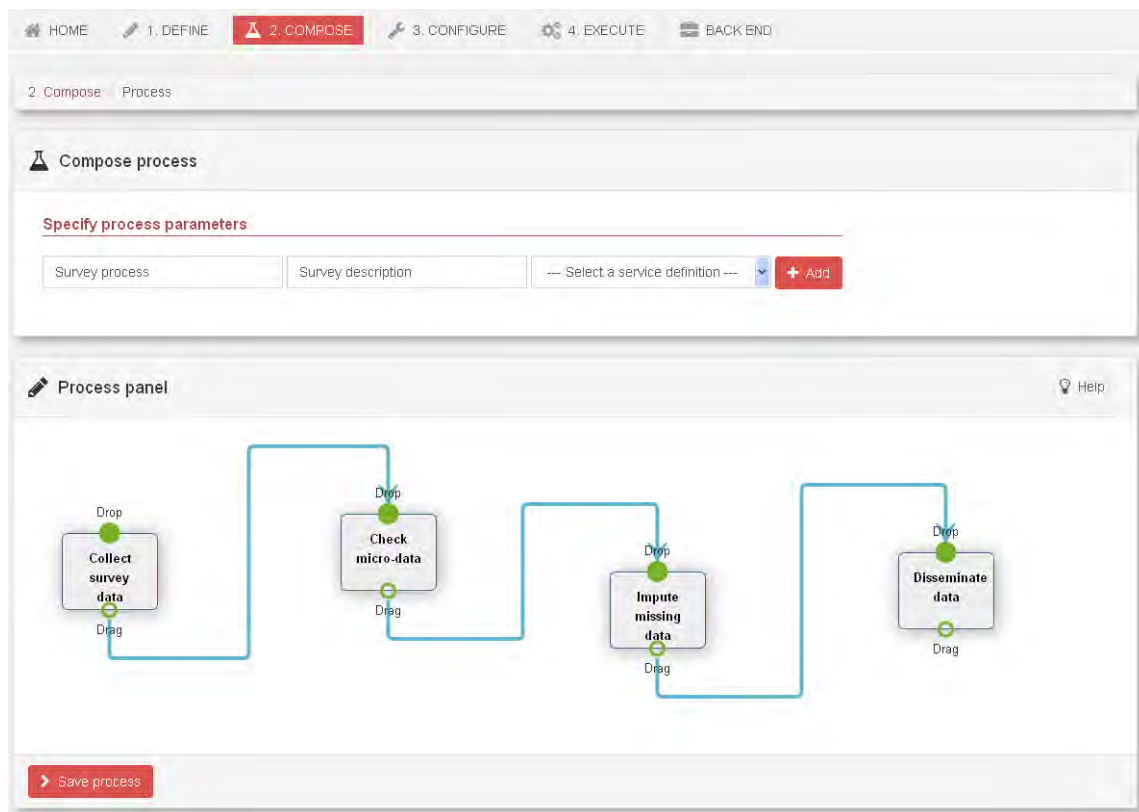
---

<sup>1</sup> Italian National Institute of Statistics - ISTAT



service (in terms of its GSBPM sub-process, GSIM input/output, etc.), and its actual implementations (in terms of the tools performing the business function).

- 2) *Compose*: in this step it is possible to design, at a conceptual level, the statistical process. CORE provides a drag-and-drop panel for connecting services already defined in the platform (“Process panel” in Figure 1).
- 3) *Configure*: this functionality allows to load a process defined in the *Compose* phase and to bind each service to one of its available implementations.
- 4) *Execute*: this is the runtime phase. In order to run a process it is necessary to specify input (data and configuration parameters) for each service. CORE provides a panel that displays the execution status (highlighting the running service).



**Figure 1 - CORE compose process**

CORE is a flexible environment that supports different types of service implementations: on the one side it allows to run command line tools (installed in CORE server), such as SAS script, R script, PL/SQL procedures developed in the Institute; on the other side it gives the possibility to invoke web-services, eventually provided by international organizations.

### **3. CASE STUDY: SURVEY ON RESIDENTIAL CARE FACILITIES**

Different statistical surveys have been taken into account as possible candidates for a complete re-engineering in CORE. Such surveys are made of many ‘building blocks’, implemented by heterogeneous tools and often laying on legacy systems. The survey on “Residential care facilities” has been chosen as a first case study.



This survey is made up of the following phases, all coordinated by means of a *custom* GUI application written in PHP:

- 1) *Data collection*: data is manually retrieved from one of the data collection systems used in the Institute. Then an Oracle PL/SQL procedure migrates data in the staging area.
- 2) *Data check*: raw data is checked against a set of rules, by means of an Oracle PL/SQL custom procedure.
- 3) *Data Transformation*: reclassification of data is made by well-established SAS procedures.
- 4) *Imputation*: an R script calculates missing data, based on previous year micro-data.
- 5) *Dissemination*: the export of data for dissemination (.csv or MS Excel files), is performed by means of a PHP library.

It is clear that such a process, involving heterogeneous technologies and managed by manual operations, can be completely optimized and re-engineered with Core, by setting up:

- a) A *generic service* for acquiring data from the data collection system: this service is potentially re-usable by any survey collecting data through such system.
- b) A set of *custom services* which wrap all the tools in use in the process (R, SAS, PL/SQL procedures, etc.)
- c) A *workflow* which composes the services that implement the business process in a completely automatic way. This makes it possible to configure the process once and replicate its 'run' infinite times by simply changing the input data.

#### 4. CONCLUSIONS

In recent years National Statistical Institutes (NSIs) are following an innovation trend, both in technological solutions and in organizational aspects, according to the Enterprise Architecture principles. In this scenario, CSPA offers a "reference architecture" for a wide spread of the "plug and play" approach in designing, implementing, sharing and reusing statistical software solutions, largely based on the "service-oriented architecture" model.

CORE is a concrete implementation of the CSPA principles. As we have shown in the case study, CORE meets important goals, such as process automation, software sharing and reuse, support for collaborative work. Moreover, CORE, through a user friendly interface and a step-by-step workflow, fosters the adoption of CSPA principles among statistical researchers, promoting software sharing, both inside a national statistical institute and at an international level.

#### REFERENCES

- [1] N. Mignolli, G. Barcaroli, P. D. Falorsi, A. Fasano, Business Architecture model within an official statistical context, Meeting on the Management of Statistical Information Systems (MSIS 2014)

- [2] CSPA Specification: Common Statistical Production Architecture, version 1.0, available at URL: <http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.0>
- [3] Generic Statistical Business Process Model, GSBPM, <http://www1.unece.org/stat/platform/display/GSBPM/Generic+Statistical+Business+Process+Model> (accessed 24 October 2014)
- [4] Generic Statistical Information Model, GSIM, <http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model> (accessed 24 October 2014)
- [5] M. Bruno, M. Scannapieco, C. Vaccari, G. Vaste, A. Virgillito, D. Zardetto, CORE: a Standard Platform for Statistical Production Processes, New Techniques and Technologies for Statistics (NTTS 2013)
- [6] E. Baldacci, Innovate or perish – Italy’s Stat2015 modernisation programme, 59<sup>th</sup> ISI-World Statistics, Congress, Hong Kong, 25-30 August 2013. <http://www.statistics.gov.hk/wsc/IPS085-P3-S.pdf>.

# On the Development of a CSPA Error Correction Service: Design and Implementation Issues

Donato Summa<sup>1</sup> ([donato.summa@istat.it](mailto:donato.summa@istat.it)), Marco Silipo ([silipo@istat.it](mailto:silipo@istat.it)), Monica Scannapieco ([scannapi@istat.it](mailto:scannapi@istat.it)), Diego Zardetto ([zardetto@istat.it](mailto:zardetto@istat.it)) and Mauro Bruno ([mbruno@istat.it](mailto:mbruno@istat.it))

**Keywords:** CSPA, Statistical Services, SOA

## 1. INTRODUCTION

CSPA (Common Statistical Production Architecture) [1] is a template architecture for supporting the industrialization of Official Statistics production processes. CSPA includes some specifications intended to define interfaces of services in a standard way, with a focus on service inputs and outputs. In the paper, we describe an experience related to the implementation of the CSPA service “Error Correction” service, realized within the CSPA Implementation Project, an international project within the High Level Group Modernization program [2]. The realized service wraps functions that are offered by the R package “rspa” developed at Statistics Netherlands [3].

### 1.1. The CSPA Concept

CSPA provides template architecture for official statistics, describing:

- What the official statistical industry wants to achieve
- How the industry can achieve this, i.e. principles that guide how statistics are produced
- What the industry will have to do, compliance with the CSPA

The principal aims of this template architecture are: (i) Provide guidance for building reliable and high quality services to be shared and reused in a distributed environment (within and across statistical organizations); (ii) Enable international collaboration initiatives for building common infrastructures and services, and (iii) Foster alignment with existing industry standards such as the GSBPM and GSIM (Generic Statistical Information Model).

CSPA is based on Service Oriented Architecture (SOA). Statistical services are self-contained and can be reused by a number of business processes (either within or across statistical organizations). A statistical service will perform a task in the statistical process, at different levels of granularity: (i) an atomic or fine grained service may, for example, support the application of a methodological step within a GSBPM sub process, (ii) coarse grained or aggregate services may encapsulate a larger piece of functionality, for example, a whole GSBPM sub process

## 2. CSPA ERROR CORRECTION: THE DESIGN

In this section we describe two relevant pieces of the design of the error correction CSPA service, namely the CSPA Service Definition and the CSPA Service Specification.

---

<sup>1</sup> ISTAT, Istituto Nazionale di Statistica

## 2.1. Definition

According to CSPA Specification [1], the design of a service has to be carried out by producing: (i) a conceptual-level *definition* of the service, where the principal business functionalities of the service are described, as well as the inputs and outputs of the service expressed according to GSIM; (ii) a logical-level *specification*, in which some design issues are addressed with respect to the methods of the service (e.g. how such methods will be invoked) and GSIM inputs and outputs implementation is specified according to a defined “logical” model (e.g. SDMX or DDI).

More specifically, the CSPA Error Correction Definition is described in **Error! Reference source not found..**

**Table 1. CSPA Error Correction Specification**

Name	Error Correction
Level	Atomic
GSBPM	5.3 Review, Validate & Edit
Business Function	This Statistical Service corrects erroneous values in records.
Outcomes	A consistent repair of records
Restrictions	None
GSIM Inputs	Unit data set, unit data set structure (Initial raw data) Unit data set, unit data set structure (Output of the CSPA Error Localization Service) Rule (Edit rules)
GSIM Outputs	Unit data set, unit data set structure, Process Output (Type=Transformed Output) (Corrected data) Unit data set, unit data set structure, Process Output (Type=Process metric) (Unsuccessful cases)
Service dependencies	The services is expected to be invoked after the invocation of the CSPA Error Localization Service
Process Method	Erroneous values that are present in records are adjusted according to the “Least Change Approach”

Notably: all the inputs and outputs of the service are GSIM objects, and the Process Method field of the CSPA Specification template does remark the importance of specifying the statistical method underlying the service even at the conceptual stage.

## 2.2. Specification

The CSPA Error Correction Specification is described in **Error! Reference source not found..** It is relevant to note that: (i) the service is invoked remotely, i.e. as a Web service thus following CSPA recommendations; (ii) input and output data structures are specified according to JSON Table Schema (JTS), thus following CSPA recommendation on explicitly separate data and metadata.

### 3. CSPA ERROR CORRECTION: THE IMPLEMENTATION

Once a formal Service Specification is carried out, then you can produce several Service Implementations of it in order to provide the business function to the community.

Normally it is possible to use different ICT technologies to wrap different statistical techniques. One of the tools able to implement the Error Correction task is “*rspa*” R package and we decided to use this one although any other could have been used.

**Table 2. CSPA Error Correction Specification**

Name		Error Correction
Protocol For Invoking the Service	Service location: I	This service is invoked as a Web service.
	Parameters passing mode:	All parameters are passed “by reference”.
	List of parameters	<ol style="list-style-type: none"><li>1. Input data set</li><li>2. Input data set structure</li><li>3. Localization data set</li><li>4. Structure of the localization data set</li><li>5. Edit rules file</li><li>6. Output corrected data set</li><li>7. Output corrected data set structure</li></ol>
Input Messages		Parameters 1) and 3) are rectangular text files following the case by variable structure with no header included. Parameters 2) and 4) are files JSON Table Schema compliant. Rule: rules are specified one by line in text format
Output Messages		Parameter 6) is rectangular text file following the case by variable structure with no header included. Parameter 7) is a file JSON Table Schema compliant.
Adopted methodology		Erroneous numeric values are corrected by finding the weighted least square adjustment of values in localized fields

In detail, we have wrapped the R script into a RestFul web service<sup>2</sup> running on a node.js<sup>3</sup> infrastructure, powered by the Restify library. Such a wrapper is specified by a file named *service.yaml*, in which you can find all service input and output parameters with their types and all other associated metadata, and the syntax for running the wrapped tool.

In order to invoke the service a client must make an HTTP POST request to the server, by using any kind of client-side technology such as curl<sup>4</sup> or a web browser but also including another CSPA compliant service in a chain fashion. The content of the POST

<sup>2</sup>See [http://en.wikipedia.org/wiki/Representational\\_state\\_transfer](http://en.wikipedia.org/wiki/Representational_state_transfer)

<sup>3</sup> See [nodejs.org](http://nodejs.org)

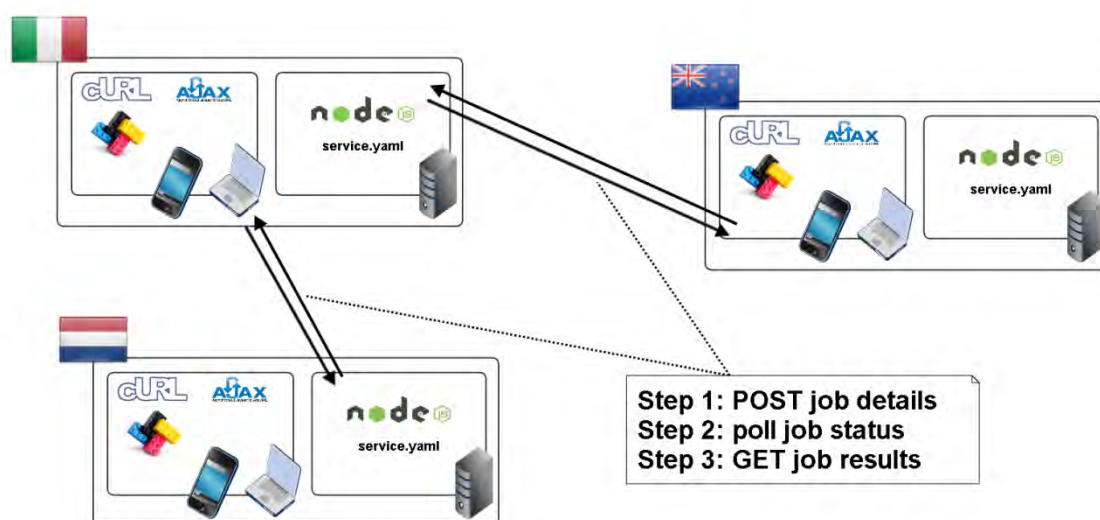
<sup>4</sup> See <http://curl.haxx.se/>

request consists of a JSON structure which contains all the actual paths or links to the files on which the tool will work, namely all input and output parameters.

Once the request is made, a process is spawn on the server to handle asynchronously the execution in a non-blocking manner and a job identifier is returned to the client. At any time a client can poll the status of the job by sending additional GET requests for the returned job id. When the job status is ‘finished’ the results are stored on the server and made available to the client, which can obtain them by making a GET request for each produced output resource.

Starting from the initial work done by CBS which created the overall infrastructure, we joined the development team of the existing GITHUB project adding the LEC service. It is now available at [https://github.com/edwindj/cspa\\_rest/tree/master/LEC](https://github.com/edwindj/cspa_rest/tree/master/LEC).

As shown in Figure 1, each NSI or organization that implements a CSPA compliant architecture can be both a service consumer or provider.



**Figure 1 - CSPA architecture and technologies**

#### 4. CONCLUSIONS

The implementation of the CSPA Error Correction Service was much more than a technological exercise: (i) issues on the format/model of input and output data emerged; (ii) issues on the protocol and on the interface to implement were also raised within the Architecture Working Group (AWG) following CSPA work<sup>5</sup>. The work presented in this paper shows a service compatible with the solutions provided by the AWG to such issues. More importantly, this work contributed, with the feedbacks deriving from its development, to the enhancement of the CSPA guidelines, aimed to develop better and better statistical services.

#### REFERENCES

- [1] CSPA Specification: Common Statistical Production Architecture, version 1.0, available at URL: <http://www1.unece.org/stat/platform/display/CSPA/CSPA+v1.0>

<sup>5</sup> See [http://www1.unece.org/stat/platform/display/pandp/\\*Home+Page](http://www1.unece.org/stat/platform/display/pandp/*Home+Page)

- [2] High Level Group (HLG): High Level Group for the Modernization of Statistical production and Services, see also: <http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services>
- [3] M. Van der Loo. rspa: Adapt numerical records to fit (in)equality restrictions with the Successive Projection Algorithm. R package version 0.1-4., available at URL: <http://CRAN.R-project.org/package=rspa>, 2013

# Reversing the flow: from an integrated system of administrative microdata to an infrastructure for the users

Simone Ambroselli<sup>1</sup> ([ambrosel@istat.it](mailto:ambrosel@istat.it)), Giuseppe Garofalo<sup>2</sup> ([garofalo@istat.it](mailto:garofalo@istat.it))

**Keywords:** administrative data, data integration, microdata collections, statistical process, policy makers.

## 1. INTRODUCTION

Starting from 2013 the Italian National Statistical Institute (ISTAT) centralized some functions for the acquisition, storage, integration and administrative data quality evaluation. In the new system SIM (Integrated System of Microdata) is realized the microdata integration and the attribution of the unique identification codes for: individuals, units (e.g. local units, schools), places and relationships among individuals and units. The aim of the system is to support all the statistical processes (like registers, and surveys) that use administrative data. At the end of its development the system will contain about 70 administrative sources (demographic, fiscal, social security, instruction) integrating several hundreds of millions of records and thousands of administrative variables in a unique Data Warehouse.

Having this amount of information, Istat developed a new project “*ARCHivio Integrato di Microdati Economici e Demografici*”<sup>3</sup> (ARCHIMEDE) with the aim of exploring the informative contents of the SIM, in order to produce new statistical microdata collections made available to social and economic researches, to sectorial and territorial planning for the evaluation of public policies at the national, regional and local levels.

## 2. METHODS

Considering a "generic" statistical production process [1], the first step identifies the specific needs ("*Specify needs phase*") where the objectives of the statistical outputs, the relevant concepts and the variables for which data are required, the sources that can meet the needs are identified. This conceptual framework, *that from the "information needs" identifies the "necessary data" to be collected*, is invariant, whether it refers to a process based on statistical surveys or administrative data or a mixed model.

The ARCHIMEDE project define a paradigm shift in the use of administrative data for statistical purposes. The exploration of the “existing data” allows the identification of the statistical information that can satisfy a need: *from the data to the needs*. This *scouting* approach changes the paradigm in relation at least to two aspects. The first concerns with the fact that the definitions and classifications can be determined in the process of exploration and therefore *are not fixed a priori*. The second, derived from the first, is that the consistency of the statistical information produced with the conceptual statistical schemes can be *evaluated only ex-post*.

The Figure 1 shows the macro functions of both SIM and Archimede.

---

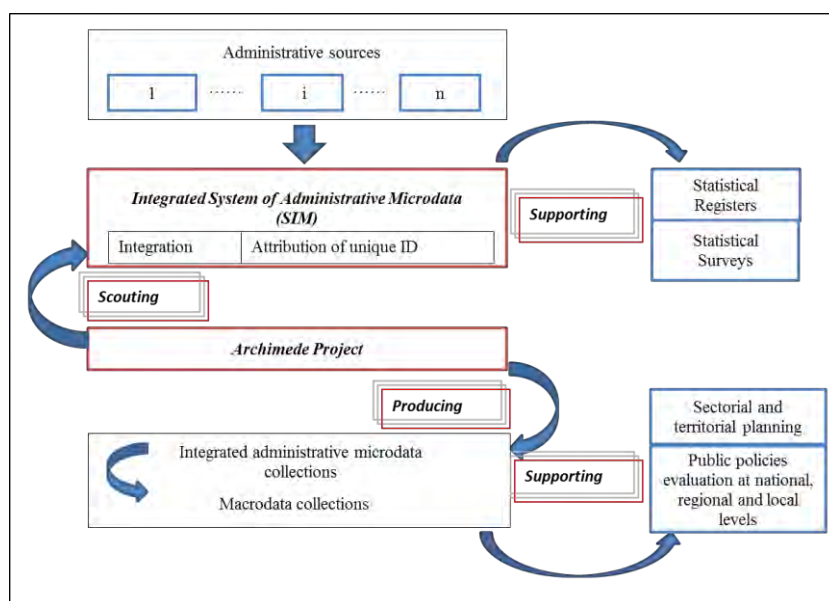
<sup>1</sup> Istat - Italian National Institute of Statistics

<sup>2</sup> Istat - Italian National Institute of Statistics

<sup>3</sup> Integrated Register of Economic and Demographic Microdata



**Figure 1. The relationship between SIM and Archimede**



The first step of the SIM process (Figure 1) is the population of the DBs for the administrative sources included in the centralized system. Admin data are not manipulated but they are transformed to guarantee the compatibility of the administrative data files with the SIM system. The step of integration and identification refers to the process of linkage and physical integration of microdata recorded in different sources according to specific strategies of integration. In that way each object in the system is identified with a unique and stable (over time) ID number.

Crossing the SIM border, the downstream statistical outputs will be automatically integrated because the objects (elementary units) have been submitted to the same process of integration and identification. ARCHIMEDE project benefits of the integration process carried out above identifying specific areas of research to support researchers and policy makers by building microdata collections.

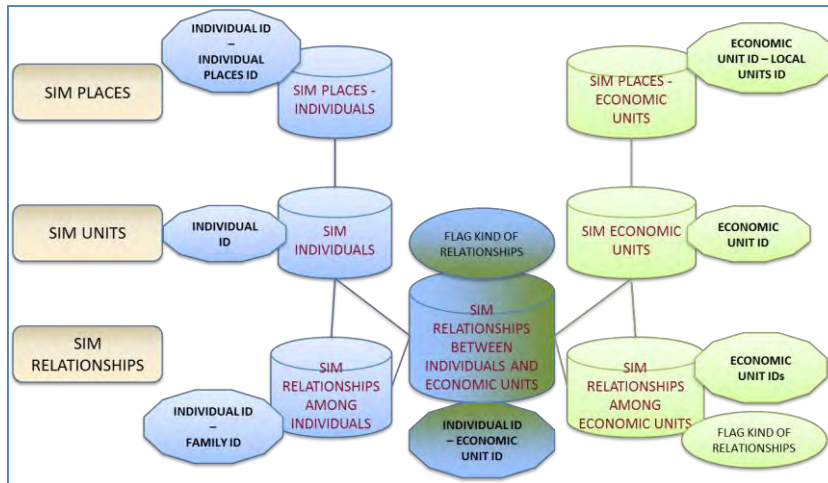
### 3. RESULTS

#### 3.1. SIM: the structure

SIM is the Istat repository of the integrated administrative microdata built with the aim of supporting the statistical production processes both for social and economic statistics. Integration means: (i) identifying each object in the administrative sources data sets with a unique and stable (over time) ID number; (ii) to define, for each object, the logical and physical relationships among administrative data coming from different sources.

SIM includes social and economic data. Seven subsystems (Figure 2) are under development; they can be grouped in: subsystems of the units; subsystems of the places; subsystems of the relationships. The traditional socio-demographic domain is shown on the left while the economic domain is on the right. In the center is placed the subsystem able to link information both from individuals and economic units.

**Figure 2. The subsystems**



The basic units for the entire system are identified in “SIM individuals” and “SIM economic units”. These two subsystems need to be developed before starting the procedures for all the others. Both from a logical and a technical point of view the two primary keys, respectively the *individual id* and the *economic unit id*, need to be assigned to all the basic objects of the system before the development of the subsystems of the places and the relationships.

The second kind of subsystems developed is referred to the places of the economic units and individuals: “SIM economic units places”; “SIM individual places”. The first contains all the locations of the economic units identified in the administrative sources acquired while the second covers the information on residence, fiscal domicile (address), other addresses associated with the household (addresses for gas, electricity, water bills) and so on, for all persons recognized by the sources stored in the system. Every address has its identifier (*individual places id* and *local units id*) and the presence of the primary keys (*individual id* and *economic unit id*) allows to link data among different subsystems.

The subsystems developed for the relationships among different entities are: “SIM relationships among economic units”; “SIM relationships among individuals”; “SIM relationships between individuals and economic units”. In the first case, the main purpose is to point out some relationships of the economic units such as the events of transformation and the control/ownership relations among them. In the second case the main purpose is the identification of the private households. In the latter case, the subsystem integrates information on the links among the individuals and the economic units where they perform their activity or spend their time. It contains the identifiers of the individuals and the economic units to tie everything together. The subsystem includes the LEED (Linked Employer Employee Data base) structures in which information from both enterprises (employer) and individuals, seen as workers, are brought together. The flag for the kind of relationships allows to refine the research activities. At the moment, ten kinds of links attributable to the macro typologies “Job”, “Business role” and “Education” are admitted.

### 3.2. Archimede project: experimental results

To evaluate the Archimede informative potentialities three experimentations were designed and conducted:

- *The city users* – the experiment aims to identify, classify and quantify the population using the territory through the use of integrated information recording data on the registry office of the population, job location of the employees and school and university location of the students. With refer to the issue on mobility, the developed experimentation makes available information on an annual basis, instead of the information available with the census every 10 years;
- *The precarious workers* – aims to identify, classify and quantify having precarious conditions in their employment. The concept of “*precarious worker*” is not well defined at international level; ILO suggest that the precarious *employment is a work relation where employment security is lacking*. This experimental approach, using more than one informative dimension, does not consider exclusively the different typologies of precarious employment contracts (like outworker contracts or the fixed-term contracts). Integrating these information with the income level, the education level, the family conditions, to have single/multi clients, the experiment allows to characterize this type of workers in an alternative way and to make a more careful reading of such phenomena.
- *The socio-economic condition of families* – aims to build a system of information on the families which allows to analyse various aspects of their socio-economic status. The integration of various administrative information enables to associate a wealth of information to families in order to highlight their lack of resources in more key dimensions (employment, income, education, disabilities, etc.). This approach allows on one hand an overall assessment of the socio-economic conditions of households, on the other hand allows the study of risk scenarios more complex than those that use the one-dimensional approach e.g. the income.

#### 4. CONCLUSIONS

SIM and ARCHIMEDE are two basic infrastructure of the statistical production of ISTAT. From one side, SIM supports the internal statistical processes sharing an integrated view of the administrative data acquired by the NSI. In fact, all the internal users, including ARCHIMEDE, benefits of the upstream identification service of SIM. From the other side, ARCHIMEDE emerges as a tool to satisfy new demands, also from outside the NSI, starting from the data available. The two projects complement each other providing the necessary microdata structures to the external users, only exploiting the potentialities of an integrated administrative microdata system without increasing the statistical burden.

#### REFERENCES

- [1] UNECE, Generic Statistical Business Process Model (GSBPM), Version 5.0, December 2013.

# Dealing with measurement and integration errors in administrative data: the case of the Italian multi-source system on small and medium enterprises

Orietta Luzi<sup>1</sup>, Marco Di Zio<sup>1</sup>, Ugo Guarnera<sup>1</sup>, Roberta Varriale (varriale@istat.it)<sup>1</sup>

**Keywords:** Multi-source statistical systems, administrative data, integration errors, under-coverage, statistical data editing

## 1. INTRODUCTION

Traditionally, in Italy annual Structural Business Statistics (SBS) on small and medium enterprises (about 4,3 million of units) are estimated based on a sample survey (referred to as SME survey hereafter) collecting information on about 100,000 sampled units, complemented with administrative data used as auxiliary information. The increasing stability, timeliness, coverage and accuracy of firm-level information available in some administrative and fiscal data (AD hereafter) sources on businesses' economic accounts has made it possible to move toward a new estimation system, based on the direct use of AD as primary source of information. A statistical information system (called *Frame SBS*) has been developed [1], where firm-level data for the main economic aggregates (e.g. Turnover, Changes in stocks, Purchases of goods and services, Intermediate costs, Labor cost, Value added) are directly obtained from integrated AD sources, covering a large portion (about 95%) of the whole SME's target population: Financial Statements, Sector Studies Survey, Tax Return data and Social Security data. The other SBS variables, which are characterized by an inadequate coverage rate, are estimated based on data observed in the SME survey using the main economic aggregates as auxiliary information [1].

This paper focuses on the methods which have been applied for dealing with some types of measurement and integration errors (in particular, consistency and coverage errors) for the main economic aggregates of the enterprise's economic accounts. Methods adopted include *harmonization* for the reduction of conceptual inconsistencies in combined administrative sources, *model-based selective editing* for the identification of possibly influential measurement errors in unit-level linked data, and *model-based predictive approaches* to deal with under-coverage in the integrated dataset.

## 2. METHODS

Moving from the traditional SBS production strategy to the new estimation system implied high initial costs for both methodological developments and data analysis, especially relating to the management of non-sampling errors characterizing the integrated AD archives (see [2] for an overview of error types in register-based statistics). For the purposes of this paper, our focus is on errors due to harmonization, measurement and coverage problems (the latter related to both units and variables under-coverage). It has to be mentioned that in the context of the *Frame SBS* no unit identification errors were possible, as the enterprises are uniquely identified in each administrative archive based on a procedure performed at the Business Register construction stage, as well as their structural characteristics.

---

<sup>1</sup> Istat - Italian National Statistical Institute

It has to be remarked that each source involved in the new estimation system actually covers different but partially overlapping sub-populations of SMEs, and that some sources provide information on (partially overlapping) variables. This “common” information has been primarily used in the data analysis phase for assessing the quality of input data of each archive. Then, as each integrated AD source uses different concepts and definitions than those required for the specific SBS purposes, the “common” information has been also used in the harmonization of data classifications and variables definitions w.r.t. the concepts described by the SBS regulation: a system of different indicators and quality measures at both micro and aggregate level were used to compare and harmonize information on target variables coming from the different sources.

Concerning measurement errors, as in the case of statistical surveys, they were identified looking at possible consistency errors in the data. A two-phases data editing strategy was implemented: at the first stage, editing activities on micro-data observed in each AD source were performed to identify logical/formal data inconsistencies (e.g. balance errors and other kind of invalid information). At the second stage, specific analyses were devoted to assess and resolve inconsistencies between variables integrated from different sources: in this analysis, inconsistent data originating outliers and influential errors were prioritized. The identification of outliers was based on a *trimming* approach, based on the analysis of the distribution of economic indicators built using information from different sources (such as the per-capita labor cost), and in rejecting those values exceeding pre-defined thresholds, by domain. Concerning influential errors, they were identified using a model-based robust selective editing approach for continuous variables [3]: in particular, the selective editing methodology implemented in the R package SeleMix (*Selective Editing via Mixture models*) [4] has been considered. The SeleMix multivariate editing approach is based on the use of contamination models. A score function strictly related to the expected error in data is defined: differently from most selective editing methods, the threshold identifying the subset of influential units can be statistically interpreted and associated to estimates accuracy (in other words, the estimated error remaining in not edited units after selective editing). In the *Frame SBS*, the target estimates in the model were the population totals of *Value added* and *Intermediate costs* by economic activity, while as auxiliary information the *Turnover*, *Number of employees*, *Personnel cost* (for enterprises with at least one employee) were used. As an example, for a 2% threshold, for the reference year 2012 about 0.3% of the total number of SME’s were classified as potentially affected by influential errors. It has to be remarked that, besides the possible identification of measurement errors, the manual inspection of influential units mainly contributed to the identification of systematic classification and harmonization errors. In general, the adoption of a selective editing approach in the context of the *Frame SBS* has highlighted a problem of costs associated to the interactive/manual inspection of the influential units: a revision of the models is needed for further optimization of results.

Model-based statistical approaches were used to deal with *under-coverage* errors, consisting of both *unit non-response* (deriving from the fact that the integrated AD sources relate to sub-populations which do not cover the overall SMEs population as defined for the SBS purposes), and *item non-response* (mainly due to the incompleteness of information, for some units, of some AD sources, which do not observe all the target variables required for SBS estimation). To deal with under-coverage, a predictive approach based on imputation has naturally allowed to build a complete micro-data file for those variables which are extensively covered by the (integrated) AD sources: in this approach, the not available information is predicted (imputed) based on the available administrative information using a combination of different techniques (including Predictive Mean Matching, Nearest Neighbor Donor, other approaches based on logistic and linear regression), which have been applied to separate groups of variables taking

into account their distributional characteristics and their relationships with other variables (see [5]).

### 3. RESULTS

In Table 1 some general results are highlighted for three of the main SBS variables: *Turnover*, *Value Added* and *Labour Cost*. In particular, the percentage difference  $d = \frac{(Y_{Frame} - Y_{Sample}) \times 100}{Y_{Sample}}$  between the variables estimates based on the new estimation system ( $Y_{Frame}$ ) and the corresponding estimates based on the SME survey ( $Y_{Sample}$ ) are reported, by year (2010-2012). Note that  $Y_{Sample}$  are based on the calibration estimator currently used for the SME survey, while  $Y_{Frame}$  are the estimates computed on the entire (partly imputed) archive by summing up all the values.

**Table 1. Percentage difference (d) between frame and survey estimates, by variable and by year**

Variable	Year		
	2010	2011	2012
Turnover	5,49	7,28	3,80
Value Added	-0,61	0,28	0,05
Labour Cost	-0,36	1,45	0,51

Some evaluation analyses have shown that actually the total difference between the survey and frame-based estimates is mainly due to a “sampling component”: in other words, the frame may ensure final estimates of the main SBS variables which are free of the traditional levels of sampling errors, deriving from the necessity of estimating target parameters for a very large population based on a relatively small survey sample size.

### 4. CONCLUSIONS

The implementation of the new information system for SBS estimation has shown that additional costs are to be paid to ensure the quality (in terms of consistency and completeness) of the administrative and fiscal sources integrated in the system are balanced by an increased accuracy of the final estimates, essentially due to the removal of the high levels of sampling errors affecting traditional survey-based estimates.

Based on the new system, starting from 2011 as reference year, estimates for key SBS will be calculated at a very refined level of detail, thus facilitating the dissemination of more detailed and better focused data to end-users. The system represents an advanced “intermediate output” which is expected to ensure higher levels of consistency between annual statistics on enterprises and National Accounts, as well as better consistency among SBS estimates and related statistical domains in the economic area.

## REFERENCES

- [1] Luzi O., Guarnera U., Righi P., The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European Conference on Quality in Official Statistics (Q2014) (2014). Vienna, 3-5 June.
- [2] Zhang, L.-C., Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* 66 (2012), 41–63.
- [3] Di Zio, M., Guarnera, U., A Contamination Model for Selective editing. *Journal of Official Statistics*, (2013), Vol. 29, No. 4, 539-555
- [4] Di Zio M., Guarnera U., SeleMix: an R Package for Selective Editing via Contamination Models, *Proceedings of Statistics Canada Symposium 2011*. Ottawa, 3-6 November (2011).
- [5] Di Zio, M., Guarnera, U., Varriale R., Imputation with multi-source data: the case of Italian SBS. Paper presented at United Economic Commission for Europe, conference of European statisticians, Paris, France, 28-30 April (2014).

# Defining usual environment with mobile positioning data

Rein Ahas ([rein.ahas@ut.ee](mailto:rein.ahas@ut.ee))<sup>1</sup>, Janika Raun<sup>1</sup> and Margus Tiru<sup>2</sup>

**Keywords:** Usual environment, mobile positioning, tourism, domestic tourism, BIG data.

## 1. INTRODUCTION

Tourism is defined as the activities of people travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes not related to the exercise of an activity remunerated from within the place visited [1]. Therefore the measurement of usual environment is essential for defining domestic tourism and domestic tourists' consumption. But understanding and measuring people's everyday movement areas is an important task in very different live spheres like transportation, taxation, public administration etc.

In tourism studies the term usual environment is employed and defined "As the geographical area (though not necessarily a contiguous one) within which an individual conducts his/her regular life routines" [2, 3]. In geographical literature the idea of activity space as the "the subset of all locations within which an individual has direct contact as a result of his or her day-to-day activities" [4] is more common. Usually it is measured with travel diaries, questioners and more recently with GPS or mobile phone based tracking technologies [5]. The latter BIG data sources enable researchers more easily and automatically gather people's everyday movement data.

Our aim is to provide a new way for determining the usual environment based on mobile positioning data and to discuss about the theoretical and methodological underpinnings related to that.

## 2. METHODS

We use Call Detail Records (CDR) (also passive mobile positioning) from mobile network operator about the times and places of call activities. The CDR database used for the current study consists of the mobile phone call activities of domestic users of the largest Estonia's Mobile Telephone Operator EMT. The market share of EMT is estimated to be 45% and its network covers 97% of the country with 4G internet. The call activities were recorded during active use of a mobile phone in the EMT network: outgoing calls and SMS messages; using internet or data services.

Every telephone in EMT network has a randomly selected pseudonym ID which is not related to the user's phone number and provides the respondent full anonymity. The random ID remains the same for all of the respondent's call activity records, even if he or she leaves Estonia and comes back at some other time which allows us to track the spatiotemporal behaviour of tourists [6, 7].

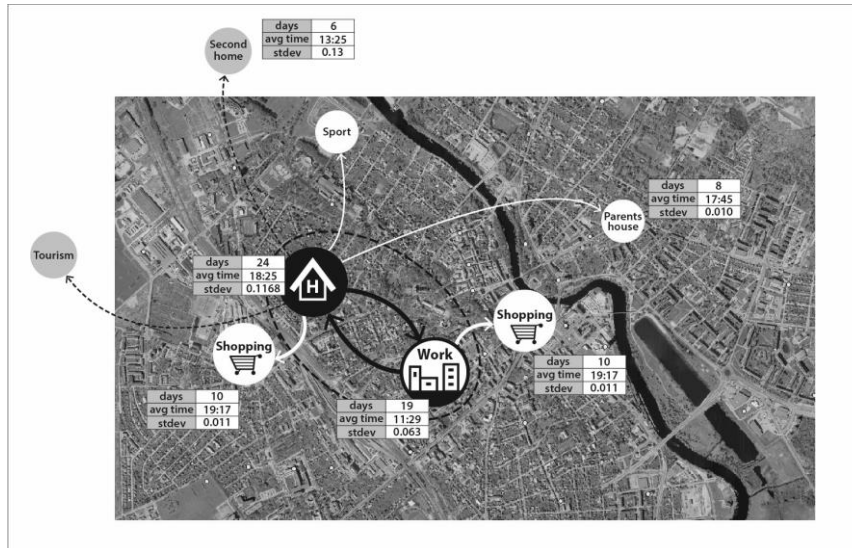
---

<sup>1</sup> Department of Geography, University of Tartu, Estonia, <http://mobilitylab.ut.ee/eng/>

<sup>2</sup> Positium LBS, Estonia, <http://www.positium.com/>







**Figure 2. Anchor point method for measuring usual environment [9]**

The third approach for defining usual environment is based on borders of administrative units. We compare distribution of CDR points on a level of a) local community (1-5 km); b) municipality (5-30 km), c) county (30-80 km). This method has less accuracy and has problems with measuring cross-border activities and selecting appropriate spatial resolution. The positive side of this method is compatibility with administrative unit based official statistics.

#### 4. CONCLUSIONS

We demonstrate the empirical results of using the described three methods for determining usual environment and defining domestic tourism in Estonia. We discuss about the theoretical, methodological and empirical strengths and weaknesses of these methods and about the practices in European Statistical System.

#### REFERENCES

- [1] OECD. Glossary of Statistical Terms. (2014) <http://stats.oecd.org/glossary>
- [2] United Nations World Tourism Organization (UNWTO). (2010). International Recommendations for Tourism Statistics 2008. New York: UNWTO. [http://unstats.un.org/unsd/publication/Seriesm/SeriesM\\_83rev1e.pdf](http://unstats.un.org/unsd/publication/Seriesm/SeriesM_83rev1e.pdf)
- [3] Govers, R., Van Hecke, E., Cabus, P. Delineating tourism Defining the Usual Environment. *Annals of Tourism Research*, 35 (4) (2008), 1053–1073.
- [4] Golledge, R.G. & Stimson, R.J. Spatial behaviour: A geographic perspective. Guilford Press, New York (1997).
- [5] Positium LBS, Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report Eurostat Contract No 30501.2012.001-2012.452 (2014), 31p. [epp.eurostat.ec.europa.eu/portal/page/portal/tourism/documents/MP\\_Consolidated%20report.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/tourism/documents/MP_Consolidated%20report.pdf)

- [6] Ahas, R., Aasa, A., Roose, A., Mark, Ü., Silm, S. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management* 29(3) (2008), 469–486.
- [7] Tiru, M., Saluveer E., Ahas, R., Aasa, A. Web-based monitoring tool for assessing space-time mobility of tourists using mobile positioning data: Positium Barometer. *Journal of Urban Technology*, 17(1) (2010), 71-89.
- [8] Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L., Discovering personally meaningful places: an interactive clustering approach. *ACM Transactions on Information Systems* 25(3) (2007).
- [9] Saluveer E, Ahas, R. Using Call Detail Records of Mobile Network Operators for transportation studies, In: Timmermans H. & Rasouli S. (eds.) *Mobile Technologies for Activity-Travel Data Collection & Analysis*, IGI Global, (2014), 325.
- [10] Vent K., Finding human activity places from smartphone gathered behavioral data, Master's thesis, department of Geography, University of Tartu (2014).
- [11] Järv, O., & Saluveer, E., Positium anchor point model. Tartu, Estonia: University of Tartu. (In Estonian) (2009).
- [12] Ahas, R., Silm, S., Järv, O., Saluveer E., Tiru, M. 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones , *Journal of Urban Technology*, 17(1) (2010), 3-27.

# Using mobile positioning data for official statistics: daydream nation or promised land?

Christophe Demunter (christophe.demunter@ec.europa.eu)<sup>1</sup>, Fernando Reis (fernando.reis@ec.europa.eu)<sup>2</sup>

**Keywords:** *mobile positioning data, big data, tourism statistics, official statistics.*

## 1. INTRODUCTION

In a changing world, statistics shall not be a static but rather a dynamic phenomenon that is in a constant relation with exogenous and endogenous factors that impact on the production methods for these (official or other) statistics. A number of such factors are currently pushing statisticians to explore the possibilities of mobile positioning data, and big data in general. *Who's on the phone? It's the 21<sup>st</sup> century calling!*

Firstly, changes in the geo-political environment condition the way data can be collected. For example, in the European context of free movement of persons, the enlargement of the Schengen area is considered a blessing for intra-EU tourism, the common labour market and international student mobility, but at the same time jeopardizes the possibilities of conducting border surveys to collect data on inbound and outbound mobility of citizens.

Secondly, technology is evolving and many tools or devices have entered citizens' everyday life. This fact, combined with decreasing prices to use such devices (e.g. roaming costs) and with continuously growing capacity to process and analyse the gigantic volumes of data generated in the back-office of service providers, is creating a whole new range of data sources that cannot be ignored.

Thirdly, the working environment of statisticians, in particular in official statistics, has changed. Political and managerial voices call for a makeover of the production methods [1], among other reasons to better match the need for more and faster data with budget cuts and respondent burden reduction. Part of the strategy is to ripe the opportunities brought by the big data era, with its abundance of data coming from social networks, sensors and IT systems in general, of which mobile phone data is a notable example. The heads of the statistical offices in the European Statistical System (ESS) agreed to jointly develop a strategy to integrate big data in official statistics [2].

In order to address the issues raised regarding the use of mobile positioning data as a source for statistics, Eurostat commissioned a study assessing the feasibility of accessing and using such data [3]. While the study focused on the specific case of tourism statistics, the findings can be generalised to other areas of statistics (and to other types of big data). This abstract/paper presents the main findings and conclusions of this study, carried out by an international, multidisciplinary consortium.

---

<sup>1</sup> European Commission, Eurostat, Unit G-3 – Short-term business statistics and tourism

<sup>2</sup> European Commission, Eurostat, Task Force Big Data

## 2. ACCESS

Mobile positioning data refers to the stored records of activities of mobile devices by mobile network operators (MNOs). While such data can also include data detail records (e.g. internet usage), location updates or technical data, this abstract/paper discusses the access to (and use of) *call detail records* (text messages, calls, data) that MNOs use for e.g. billing purposes.

The study revealed three main potential barriers to accessing mobile phone data: privacy protection and legislation, technical feasibility and financial and business related aspects.

### 2.1. "Hello, is it me you're looking for?"

While the end result of processing and compiling personal data for statistical purposes is by itself anonymous (i.e. aggregated tables), the actual processing and compiling operates in a grey area. European (and nationally transposed) legislation [see [3], report 2] lays down the conditions for accessing (directly or indirectly identifiable) personal data for statistical purposes, but the interpretation of the legal framework is still very ambiguous. As a consequence access to data is seldom granted – e.g. in the context of the feasibility study, (test) data could eventually be accessed in only one of the four covered countries.

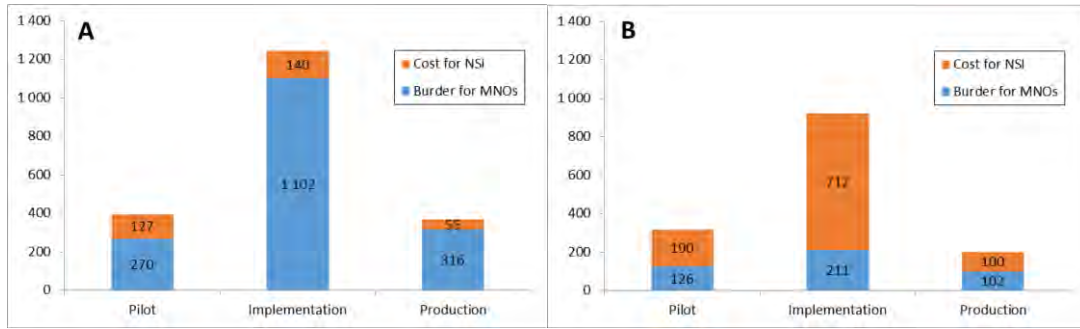
Legislators, policy makers and data protection agencies tend to be reluctant to grant access, partly due to fear of the public opinion which appears not to be ready for the exploitation of big data sources containing personal information on the whereabouts of citizens.

This fuzzy legal environment is currently the Achilles heel and in order to enable statisticians to use mobile phone data (and many other sources of big data), a harmonised legal (and methodological) framework for statistical authorities to access data held by e.g. mobile network operators shall be a priority area of work.

### 2.2. Dial M for Mobile positioning data?

Accessing mobile positioning data is obviously not a simple push on the button, as it involves very large datasets. This poses technical challenges in terms of data processing and transmission. However, even if challenging, the technical feasibility is in general not considered to be a hard barrier. The process is complex but - as some case studies have shown [see [3], report 1] - not impossible.

The technical choice for a decentralised or centralised system (most work done by the MNO and national statistical institutes (NSIs) receive the output, or statisticians receive raw data for processing into statistical indicators) has an important impact on the overall cost (see Figure 1).



**Figure 1. Estimate of the cost (in 1000s euro) of implementation and regular production of mobile positioning based statistical indicators (the case of a country with 3 MNOs with 16 million subscribers and a 15 day latency/timeliness of the data) - A: decentralised approach; B: centralised approach.**

### 2.3. Win - win?

Besides the vague legal framework - and, linked to it, the fear of the public opinion - for granting access (see paragraph 2.1), MNOs may also see financial or business related barriers.

Firstly, business secrets (e.g. share in the national roaming market) may refrain MNOs from releasing data for further external research and analysis. Secondly, while the costs for the MNO can be considerable (implementation of a data extraction system, human resources, etc), the direct benefits are not clear.

NSIs have in general the legal powers to request from any citizen or enterprise the data which are needed for the production of official statistics. However, one of the fundamental principles of official statistics is that respondents should not bear an unreasonable burden for that [4] and it should not put them in a comparative disadvantage against their competitors.

To motivate or incite MNOs, a mutually beneficial relationship should be sought. This can consist of a remuneration scheme to compensate for the response burden (however, budget cuts in NSIs will significantly reduce this option) or can include the possibility for the MNO to use the (personal) data for their own, even profit-making, purposes.

## 3. USE

Before investing in accessing mobile positioning data, interested users should have an idea of the potential benefits of using such data. This section will have a look at some methodological issues and dig into the quality aspects, including an evaluation of the coherence with current, more traditional data sources.

### 3.1. Methodological issues

Rather than wondering about privacy protection and other access related challenges, many potential users will jump immediately to more pragmatic questions related to representativeness or coverage. What about those not using mobile phones? What about the border noise of connecting to foreign networks without being abroad? What about tourists buying foreign SIM cards when travelling?

While there certainly are shortcomings that are inherent to the use of mobile phone data, most over and undercoverage issues prove not to be very significant, or at least not more

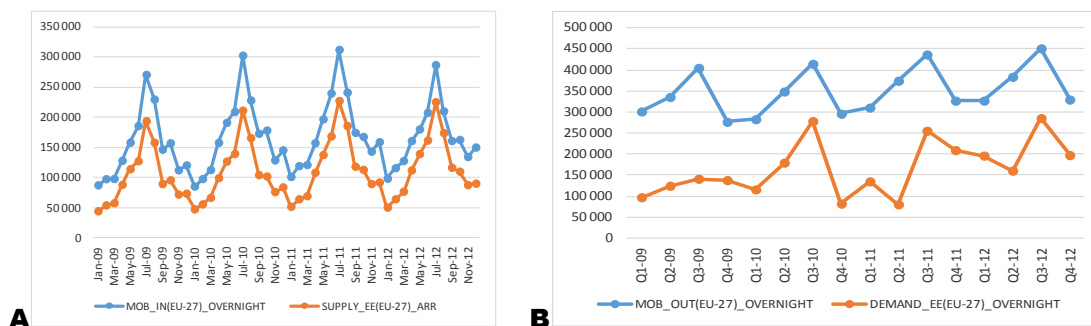
significant than similar shortcomings of traditional sources, e.g. the penetration rate of mobile phone use versus the representativeness of CATI interviews using a phone directory of fixed lines or, for tourism statistics, the very important recall bias. Furthermore, if quantified, these shortcomings can be taken into account when calibrating the final results (given sufficient reference information about the total population is available).

The likelihood that big data sources can reproduce the existing statistical indicators is another common concern. However, to answer this question, one should first assess the origins and relevance of the currently available statistics. Often, the statistical requirements laid down in statistical legislation were heavily influenced by the available data sources at the time of legal drafting, e.g. limiting data on inbound tourism to tourists staying at rented accommodation such as hotels (instead of covering all inbound tourism). Therefore, an essential choice to be made when thinking of absorbing big data as a source for official statistics will be between an evolution (partially replacing existing indicators within the current framework) or a revolution (rethinking indicators, thinking out of the box delineated by the current framework and applying a zero-based budgeting approach rather than looking for small incremental changes).

In terms of data availability, advantages and disadvantages need to be properly assessed. The first include the possibilities of new breakdowns, finer regional detail and superior timeliness. The latter include the absence of socio-demographic breakdowns, motivations (e.g. purpose of tourism trips) or domain-specific variables (e.g. tourist expenditure, means of transport used). An important challenge for users and producers is to find a compromise between timeliness and completeness.

### 3.2. Coherence

Tests comparing mobile positioning data with existing data have given very promising results. The data appears to be strongly correlated and differences in level can often be explained through known shortcomings or specificities of one of the two sources under comparison. Figure 2 gives two examples comparing mobile positioning data with accommodation statistics and tourism demand statistics respectively.



**Figure 2. Coherence of mobile positioning data with traditional data sources in Estonia - A: number of inbound tourism trips (compared with accommodation statistics); B: number of outbound tourism trips (compared with tourism demand statistics).**



#### 4. CONCLUSIONS

A replacement of the existing production methodology of official statistics with mobile positioning data (or other big data) is not likely to happen overnight, but any efficient and competitive player in the world of statistics sooner or later will have to be prepared to rely on big data for part of its business.

The feasibility study outlined in this abstract/paper can be a starting point for NSIs and researchers wanting to embark on the use of mobile positioning data to compile statistics. In parallel, however, a focus should also remain on the need for horizontal (i.e. across domains) and international cooperation in the area of big data. Implementation actions covering several countries and several domains are most likely to be the successful ones.

Finally, it should be kept in mind that big data does not only have the potential to improve the quality of existing statistics but also to stimulate a new, data driven statistical system.

#### REFERENCES

- [1] Eurostat, *The ESS Vision 2020*.
- [2] European Statistical System, *Scheveningen Memorandum – big data and official statistics*.
- [3] Eurostat, *Feasibility study on the use of mobile positioning data for tourism statistics*.
- [4] United Nations, *Fundamental Principles of Official Statistics*.



# Using Passive Mobile Positioning Data in Tourism and Population Statistics

Laura Altin (laura.altin@positium.ee)<sup>12</sup>, Margus Tiru<sup>1</sup>, Erki Saluveer<sup>1</sup>, Anniki Puura<sup>2</sup>

**Keywords:** passive mobile positioning, Estonia, tourism statistics, population statistics

## 1. Introduction

Recent developments in information and communications technologies (ICT) have left their mark on tourism, travel and everyday activities: individual GPS tracking, internet-based picture uploading websites, location-based social media check-ins, and interactive tour-guides.

One of the ways of studying the movement and behaviour of people is through the use of mobile telephones. Mobile positioning data in this paper refer to the large-scale location data of subscribers of mobile network operators that are processed and stored in operators' systems. This is highly sensitive but also very valuable data that could be used anonymously and aggregated thus ensuring the privacy of the subscribers and providing valuable insights in fields like tourism and population statistics (Eurostat Feasibility Study 2014).

The aim of this paper is to assess the possibilities of enhancing tourism and population statistics through the integration of positioning data from mobile communication networks.

## 2. Methods

In this paper passive mobile positioning data is used. Using mobile positioning data is a relatively new method in the area of tourism and population statistics (Tiru et al 2010).

The data from Estonia's biggest mobile operator EMT (Estonian Mobile Telephone) was used. EMT covers nearly 99,9% of total land area of Estonia. Market studies show that EMT has a 46% share of the local mobile phone market (TNS Emor 2008). The method for data collection and analysis has been developed in Estonia in cooperation between the private company Positium LBS, mobile operators and the Department of Geography at the University of Tartu. The database used in this paper consists of a spatial and temporal register of call detail records of domestic and foreign mobile phones using EMT's service. Call detail record is any active use of a mobile phone in networks – incoming and outgoing calls, SMS, GPRS etc. Roaming service means that mobile phones registered in countries other than Estonia can be used on the Estonian network. The register includes the following parameters for every call activity (Ahas et al 2007; Ahas et al 2008):

---

<sup>1</sup> Positium LBS

<sup>2</sup> University of Tartu

- The exact time of call activity;
- The randomly generated unique ID number for the phone (not related to the phone or SIM card number);
- The antenna ID with the geographical coordinates of the antenna;
- The phone registration country – used as the nationality of the phone owner.

The geographical precision of the data is determined by the level of the GSM network cell (Cell ID). The spatial accuracy of the location information depends on the density of the mobile network. The accuracy is higher in urban areas, where the mobile network is denser and lower in less populated rural areas, where the mobile network is sparse and where less people dwell and move. The measurements by Positium LBS show that the spatial precision in Estonia varies from 100-1,000 meters in larger cities (Tallinn, Tartu, Pärnu) to 1,5-20 km in rural areas. Quality of positioning data has been compared with accommodation statistics and a correlation between the two databases has been found.

Due to privacy issues, the database is anonymous and does not contain any back-traceable personal information about the user of the phone. To recognize a person, which is essential in order to analyze repeat visits and loyalty, a randomly generated unique ID number is assigned to every phone. The ID generated by the mobile operator enables the identification of the CDR-s made by one person during the study period.

The collecting, storage and processing of the data obtained complied with European Union requirements regarding the protection of personal data according to EU directives on handling personal data and the protection of privacy in the electronic communications sector. Separate approval was also sought from the Estonian Data Protection Inspectorate (Directive 2002/58/EC of the European Parliament).

### 3. Results

In the presentation perspective of tourism and population statistics will be introduced based on case studies in Estonia.

#### 3.1. Tourism statistics

The main benefit of passive mobile positioning data compared to traditional methods like population survey, border statistics *etc.* is the ability to evaluate the indicators for a much larger sample (indicators: country of origin of visitors, repeat visits, number of days/nights spent). It is also possible to distinguish same-day and overnight visits, transit visits from longer stays, tourists from long-term visitors (residents).

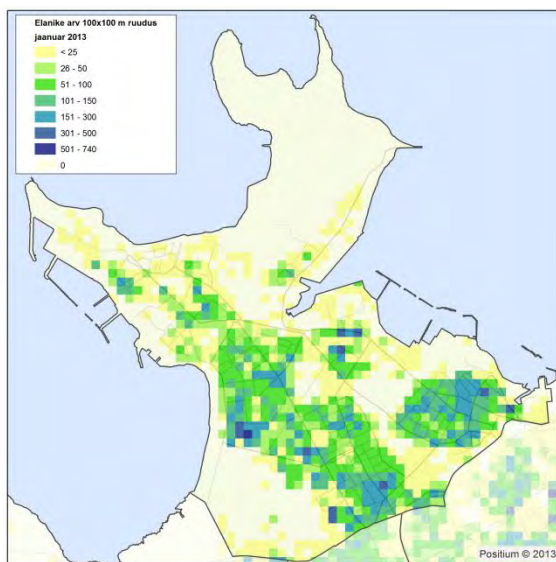
All indicators can be distinguished on the level of a particular event (concert of Robbie Williams, which took place in august 2013). The results show that visitors to this kind of event (e.g. musical performance) originate from nearby countries (while regular tourists also come from more distant countries), and have a duration of visit that is longer than in regular tourists and attract new segments of tourists (first-time visitors).

#### 3.2. Population statistics

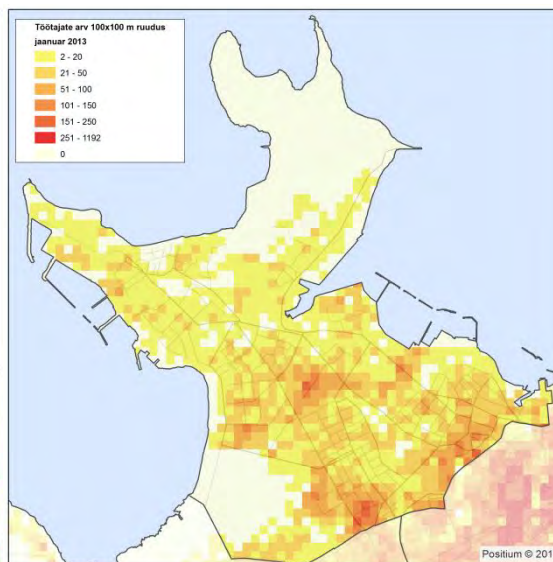
Based on passive mobile positioning data home anchor points, work-time locations and other everyday activity places can be distinguished. Compared to population census data, MPD helps to evaluate place of usual residence and also distinguish second home

anchors (where person spends more most of the time from one month in some place which is not persons home – temporary home, summer houses, holiday homes etc).

In this research the case of Põhja-Tallinn district in Estonia is introduced. For example home (fig 1) and work-time (fig 2) activity spaces are described (on the accuracy level of 100x100m grids).



**Fig 1.** Home-anchor points in Põhja-Tallinn district



**Fig 2.** Work-time anchors in Põhja-Tallinn district

#### 4. Conclusions

Despite some obstacles there exists substantial potential in using mobile positioning data as new and alternative data source in tourism and population statistics. The benefits of an increased adoption of this data are: cheaper price, timeliness, quicker delivery to end-user, bigger sample sizes, more comprehensive data.

#### References

Eurostat Feasibility Study on the use of mobile positioning data (2014). <http://mofbs.positium.ee>

Tiru, M., Kuusik, A., Lamp, M-L., Ahas, R., (2010), LBS in marketing and tourism management: measuring destination loyalty with mobile positioning data. Journal of location based services. 4 (2), 120-140.

TNS EMOR (2008) Telephone survey - CATI-bus.

Ahas, R., Aasa, A., Roose, A., Mark, Ü. and Silm, S., (2008), Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management Vol 29, Issue 3, 469-486.

Ahas, R., Aasa, A., Mark, Ü., Pae, T., Kull, A., (2007), "Seasonal tourism spaces in Estonia: Case study with mobile positioning data." Tourism Management, 28 (3), 898-910.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

# Indicator for the representativeness of linked sources

Dingeman Jan van der Laan ([dj.vanderlaan@cbs.nl](mailto:dj.vanderlaan@cbs.nl))<sup>1</sup>, Bart F. M. Bakker<sup>1</sup>

**Keywords:** representativeness, linkage quality, LR-indicator, R-indicator

## 1. INTRODUCTION

Administrative data are increasingly used in official statistics. These data have many advantages: e.g. a much smaller response burden, the possibility of large sample sizes for the production of small domain statistics, opportunities for following units through time to create longitudinal data and comparatively low collection costs. However, the wider use of administrative data has also revealed more and more quality issues [1,2]. One of the limitations of administrative data is that they usually have a small number of variables. It is not possible to produce the desired crosstables, if the two or more required variables are not in the same source. Therefore, data linkage techniques are used to combine data from different administrative sources. However, missed links could lead to biased estimates if the missed links are selective. This is more or less similar to selective non-response in surveys [3,4].

The linkage effectiveness is the most used indicator of the quality of the linkage. However, that is not always a good indicator of the quality of the linkage process of two sources. Estimates based on a linked data set with a high linkage percentage can still be biased if the missed links are selective. Other often used measures are based on the numbers of false positives (false links) and false negatives (false non-links). However, this can only be computed if the true links are already known [5] which in practice is difficult. Furthermore, these measures still suffer from the fact that the effect of missed links on estimates depends largely on the selectivity of the links.

We propose an indicator for the similarity of the linked records to the target population under investigation. We make use of the notion that missed links lead to similar errors as selective non-response in surveys. To measure representativeness of the response of a survey, the R-indicator has been developed [6,7]. The R-indicator is based on the idea that the response of a survey is representative of a target population if the response probabilities are the same for all units in the population. Note that this corresponds to the idea of Missing Completely At Random (MCAR) with respect to all variables. Because these response probabilities are unknown, a weaker version of this idea is used: the response of a survey is representative of a target population if the average response probabilities over variable  $\mathbf{X}$  (with  $H$  categories) is constant. This weaker definition corresponds to a missing data mechanism that is MCAR with respect to  $\mathbf{X}$ . For  $\mathbf{X}$  also a vector of variables could be used. The indicator is called Linkage Representativeness Indicator (LR-indicator).

## 2. METHODS

The LR-indicator is based on the concept of linkage probabilities, i.e. the probability to be linked. If the records in two sources are linked, the resulting links are representative of a target population if all units have the same linkage probability. In that hypothetical situation, it is very easy to evaluate the linkage result by measuring the amount of variation in linkage probabilities. The more variation, the less representative. However, the linkage probability is a theoretical concept that cannot be observed. What can be

---

<sup>1</sup> Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands.

observed is the value of  $R_i$ , which has the value 1 if element  $i$  links (with probability  $\rho_i$ ) and otherwise has the value 0 (with probability  $1-\rho_i$ ). The idea is to estimate the linkage probabilities using auxiliary variables, chosen in such a way that the linkage probabilities are optimally explained. If a set of explanatory variables  $\mathbf{X}$  can be found and their values  $\mathbf{X}_i$  are observed in the linked sources, the linkage probabilities  $\rho_i$  can be replaced by the linkage propensity.

$$\rho_i(\mathbf{X}) = \Pr(R_i = 1 \mid \mathbf{X} = \mathbf{X}_i).$$

To estimate the linkage propensities, one could use a logistic regression model like

$$\text{logit } \rho_i(\mathbf{X}) = \log\left(\frac{\rho_i(\mathbf{X})}{1-\rho_i(\mathbf{X})}\right) = \sum_j X_{ij}\beta_j$$

Following Schouten et al. (2009), the LR-indicator for the representativeness of the linked records of two sources can be defined by

$$\text{LR} = 1 - 2S_\rho,$$

Where  $S_\rho$  is the standard deviation of the estimated linkage probabilities. LR equals one if all linkage probabilities are equal and then there is complete representativeness. The smaller the value of LR (the minimum value is zero), the larger the lack of representativeness.

Besides an overall indicator, it is also of interest to know subpopulations are under-represented and which over-represented. This information can direct further efforts in the linkage process (e.g. the search for specific linkage variables for these subpopulations) and inform on possible biases in the analyses. For this the partial LR-indicators can be used [8], of which we currently only describe the unconditional one. Let  $Z$  be categorical variable with categories  $k=1,2,\dots,K$ .  $Z$  is a component of  $\mathbf{X}$ . Then the unconditional partial indicator for  $Z$  is defined as [8]:

$$P_u(Z, \rho_x) = \sqrt{\sum_{k=1}^K \frac{N_k}{N} (\bar{\rho}(\mathbf{X}, k) - \bar{\rho}(\mathbf{X}))^2} = \sqrt{\sum_{k=1}^K P_u(Z = k, \rho_x)^2},$$

with  $\bar{\rho}(\mathbf{X})$  and  $\bar{\rho}(\mathbf{X}, k)$  the average linkage propensity and the average linkage propensity for category  $k$  of  $Z$ , respectively;  $N$  and  $N_k$  are the number of records and the number of records in category  $k$ , respectively. The value is bounded above by 0.5 and below by zero. The larger the value, the larger the contribution of  $Z$  to the lower representativeness.  $P_u(Z = k, \rho_x)$  is the unconditional partial indicator of category  $k$  of  $Z$ . A positive value indicates an overrepresentation and a negative value an overrepresentation. The values are between -0.5 and 0.5.

### 3. RESULTS

The LR-indicator is applied to studies that were conducted by Statistics Netherlands in collaboration with a consortium of universities. In the first study the Population Register is linked to the Employee Register. The LR-indicator is determined using the micro-data. In the second study, the National Twin Register is linked to a register of a large Health Insurance Company, using only the aggregated data. Results from both studies will be presented. However, at this moment the results of only one study are available.

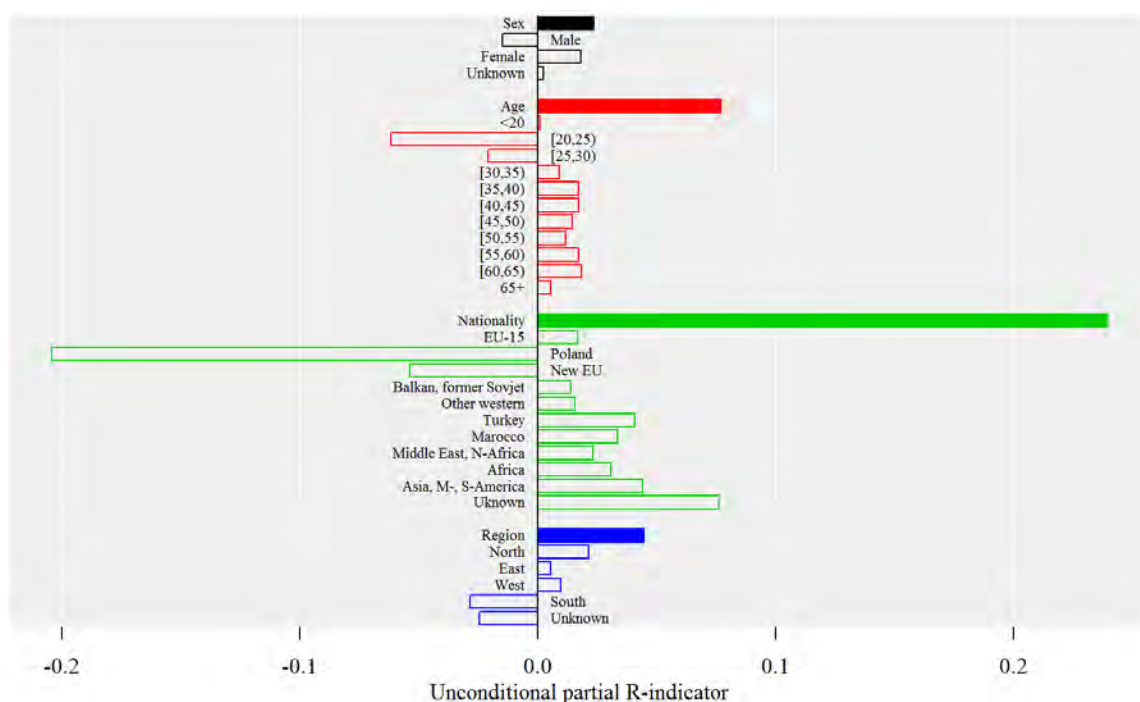
In the first example the Population Register (PR) and the Employment Register (ER) are linked. The target population was defined as the employed foreign residents in The

Netherlands, with exception of residents with a Belgian or German address. The purpose of the linkage was to enrich the ER with a number of variables from the PR, in order to answer questions on the impact of the migrant history on occupational careers. Employees with a Belgian or German address were excluded because the majority of them are borderland workers and do not have migrant history of interest for the research question.

The registers are linked with a combination of deterministic and probabilistic methods. In the first step, the records are linked deterministically on a personal identification number that is widely available in administrative data sources in the Netherlands. The remaining records are linked probabilistically [9,5]. To reduce the number of possible pairs, the data are blocked on variables that are assumed to be of very high quality: postal code or date of birth. For the probabilistic linkage date of birth, sex, postal code, house number and extension are used. In total 84.6% of the records of the ER could be linked. The probabilistic linkage leads to an increase of 0.3% of the total number of linked records.

The covariates sex, age, nationality and region were used in the calculation of the linkage propensities. Using these propensities, we arrive at an LR-indicator of 0.50 which is significantly lower than values found in some of the surveys performed at Statistics Netherlands [6]. Therefore, one should consider correction for the linkage selectivity e.g. by weighing the data, or even end the research entirely.

It is also possible to calculate partial R-indicators, which express the contributions of the various subgroups to the overall LR-indicator [9]. The unconditional partial R-indicators are shown in figure 1. Nationality has the largest contribution to the lack of representativeness of the linked data set. Especially, persons from Poland and other new EU countries (joined EU after 2004) are underrepresented. Any analysis of this data set should, therefore, take nationality into account. Furthermore, male persons, persons aged 20-30 and persons living in the south of the Netherlands are underrepresented. However, these categories will correlate with the underrepresented nationality groups.



**Figure 1. The unconditional partial LR-indicators for the linkage between the Employment Register and Population Register.**

#### 4. CONCLUSIONS

With the increasing use of linked administrative data, the need for a measure of the representativeness of the data after linkage also increases. The most frequently used indicator for the quality of the linkage process is the effectiveness. However, this is no measure for the representativeness. If all population elements have the same probability of being linked, then the resulting data are representative of the population even if the effectiveness is low.

In this paper, we present the LR-indicator to measure the representativeness of linkage of two sources. It is based on the idea that the standard deviation of the linkage probabilities estimated for a set of auxiliary variables provides sufficient indication of the representativeness: the higher the standard deviation, the lower the representativeness. It is also possible to determine partial LR-indicators to find out which variables and which categories of these variables have the largest contribution to the LR-indicator and therefore are the most misrepresented in the linked data file.

We give two examples of which one is described in this abstract. We determine the LR-indicator of the linkage of a Population Register and an Employee Register with age, sex, nationality and region. The study reveals that the representativeness is rather low. Either one should put more effort in improving the linkage, correct for the linkage selectivity or end the research.

#### REFERENCES

- [1] W. Grünewald and T. Körner, Quality on its way to maturity: results of the European conference on Quality and methodology in Official Statistics (Q2004), *Journal of Official Statistics*, 2005, 747-759.
- [2] A. Wallgren, and B. Wallgren, *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology (New York: Wiley, 2014).
- [3] B.F.M. Bakker and P. Daas, Some Methodological Issues of Register Based Research, *Statistica Neerlandica*, 2012, 2-7.
- [4] L.-C. Zhang, Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica*, 2012, 41-63.
- [5] A. Ariel, B. Bakker, M. de Groot, G. van Grootheest, J. van der Laan, J. Smit, B. Verkerk, *Record linkage in health data: a simulation study* (The Hague / Heerlen: Statistics Netherlands, 2014).
- [6] B. Schouten, F. Cobben and J. Bethlehem, Indicators for the representativeness of survey response, *Survey Methodology*, 2009, 101-113.
- [7] N. Shlomo, C.J. Skinner and B. Schouten, Estimation of an Indicator of the Representativeness of Survey Response, *Journal of Statistical Planning and Inference* 2012, 201-211.



- [8] B. Schouten, N. Shlomo and C. Skinner, Indicators for Monitoring and Improving Representativeness of Response, *Journal of Official Statistics*, 2011, 231-253.
- [9] I.P. Fellegi and A.B. Sunter, A Theory for Record Linkage, *Journal of the American Statistical Association*, 1969, 1183-1210.

# New proposals for linkage error estimation

Tiziana Tuoto ([tuoto@istat.it](mailto:tuoto@istat.it))<sup>1</sup>, Niki Stylianidou<sup>1</sup>

**Keywords:** probabilistic record linkage, linkage quality assessment, mixture models

## 1. INTRODUCTION

The use of combined data from different sources is an advantage that the National Institutes of Statistics try to take advantage of in more and more occasions.

In a context in which large information are produced by different sources, achieved through integrated and comparison operations (methods and techniques), it becomes increasingly urgent the need to equip the results of data integration and linkage with quantitative assessments of the quality of such combining operations.

As part of the integration of data at a micro level, the currently applied methodologies of record linkage, are widely used within the Italian National Institute of Statistics (Istat). These linkage methodologies generally produce good results, mainly when applied by means of the strong identification matching variables. However evaluation information ie quantitative indicators of quality of the output of the matching procedures, is rarely available. In official statistics, evaluation information regarding the output (matching data) of a procedure has a fundamental role since “certify” the accuracy and credibility of the data; especially if the aim is then to use them for further analysis or for inference.

## 2. METHODS

The following matrix identifies the possible combinations that a matching methodology may produce compared to the truly linked data. The given matrix is used for building the quality indicators of the record linkage procedure:

**Table 1. Output of the record linkage procedure**

	<i>Linked</i>	<i>No Linked</i>
Matched	<i>a</i>	<i>b</i>
Unmatched	<i>c</i>	<i>d</i>

The following indicators can be calculated:

- **Rate of false matching:**  $f = b/(a+b)$
- **Rate of missing matching:**  $m = c/(c+a)$

Usually, the variable Linked / No Linked, that determines the true state of a match, is not observed. The methods for the estimation of the Linked / No Linked variable are mainly two:

- a) with very accurate manual revision of the matching outputs by clerks;
- b) the use of statistical methods such as latent variable mixture models (Fellegi and Sunter, 1969; Belin and Rubin, 1995).

<sup>1</sup> Istat – Italian National Statistical Institute

## 2.1. The Belin and Rubin method

The most defused method for determining the matching procedure failures, mainly of those of false matches, it has been proposed by Belin and Rubin (1995). Their proposal had as a starting point a training set extracted from the records of declared up matches. Once calculated the matching weights,  $R^2$  are firstly logarithmic transformed, and thus a second transformation is applied for rescaling them in the range 1-1000. Finally, last a Box-Cox transformation is applied on the log-rescaled  $R$ 's, in order to give to the weights a normal distribution shape. At this point, the authors apply a Fitting Transformed Normal Mixtures with Fixed global parameters. The model refers to the latent variable that expresses the state of false matches for records declared matches, ie the cell b.

The proposed method of Belin and Rubin (1995) already-known problems that suffers the are:

- the difficult construction of a representative training set for which crucial information are known as the rates of truly linked matching records and thus being able to reproduce the weight distribution of the two populations Linked and No linked as observed in the universe of Matches;
- the fact that the distribution of the transformed weights of matches does not always be adapted to a normal distribution through a Box - Cox transformation.

Here in Istat a practical replication of the Belin and Rubin method has been tempted facing the above exposed problems and thus aiming no results. The attempt to replicate the proposed methodology by Belin and Rubin (1995) has been helpful for focusing on the need to seek a rule that distinguishes the True Linked vs False Linked from the matched records dataset. For this reason, we have tried to apply the methodology of Canonical Discriminant Analysis in the context of record linkage. The seek result seems to be the very similar as the one of Fisher's when in 1936 introduced this methodology regarding a dataset with two different kind of Iris flowers.

## 2.2. Discriminant Analysis

For exposing the method Discriminant Analysis, some notation is necessary:

Let's the vector of the match records be  $Y = (Y_1, Y_2)$

Where  $Y_1$  are the true linked matched records (from table 1 is *a*)

$Y_2$  is its complement alias the false-linked but matched records (from table 1 is *b*).

The covariance matrix of the matched records is :  $X = (X_1, X_2)$

denoting with  $\bar{x}_1$  the vector that contains the averages of the matched truly linked covariates

and with  $\bar{x}_2$  the vector that contains the averages of the matched false linked covariates.

Respectively are denoted with  $S_1$ = the matrix of Variance of the matched true linked and  $S_2$ = the Variance matrix of the matched false linked.

$n_1$  number, indicates how many matched true -linked records are and  $n_2$  is the corresponding number of matched false linked records.

---

<sup>2</sup> for example, the ratio of the likelihoods obtained through a procedure proposed by Fellegi and Sunter, although several other different procedures can also be considered for obtaining the weights

Calculating the Variance and Covariance matrices :

**Between clusters** 
$$B = \frac{n_1 * n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^t$$

and

**within clusters** 
$$W = \frac{S_1(n_1-1) + S_2(n_2-1)}{n_1 + n_2 - 2}$$

that are necessary for construction so called *Fishers' equation of discrimination*:

$$Y = XW^{-1}(\bar{x}_1 - \bar{x}_2)$$

The discrimination point between the two groups (Breaking point) is given by the :

$$break = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

and thus the assignment rule for each unit is given by:

$$\min \triangleq \begin{cases} |\bar{y}_1 - y_i| \\ |\bar{y}_2 - y_i| \end{cases} \quad \forall y_i$$

That is each unit of  $y$  will be assigned to the population whose center of gravity  $\bar{y}_i$  is the least distant.

Once established the effectiveness of this method, using different applications, it may be proposed to improve the data produced by the record linkage using a training set to extract the optimal rule to be applied on the whole of Matched (in order to estimate  $b$ ) but also on all the unmatched (in order to estimate  $c$ ).

### 3. PRACTICAL APPLICATIONS

The methods proposed, the Bellin and Rubin and the Discriminant Analysis, have been experimented with the aid of simulated data, for which the real state of linkage, is known through the availability of a univocal identification code.<sup>3</sup>

#### 3.1. The data in use and the linkage scenarios

The data used in this exercise have been created by Paula McLeod, Dick Heasman and Ian Forbes, of the Office of National Statistics (UK), within the European project: ESSnet Date Integration. The two methods have been tested in two different sceneries of linkage: the first one introduces a rate of false linkage slightly lower than 2% ; this it will be pointed out as Gold Scenario. In this case it has been drawn out a sample from the total of the available data, around 1500 unities. Highly identifying variables has been used for the matching procedure according to the model of Fellegi-Sunter, as implemented in the open source software for the record linkage RELAIS.

In the first matching exercise, as the dimensions of the two datasets are small, it has been applied the matching cross product of all the records. The identificative variables used in the matching procedure are: Name, Surname, day and year of birth, allowing individualizing 1300 matches of which 45 false linked matches. The matching threshold of  $p\_post$  is 0.6 .

A second scenario of linkage has been built considering all the records that compose the ons reference file that is 26000 records. As the number records is high the search space

---

<sup>3</sup> Thus the main aim of this work is to have a methodology that will guarantee us a high level quality matched records of micro data and thus an estimation of the possible false linked matches that may contain (error).

dimension has been reduced using the SimHash function in selected blocks. As matching identificative variables has been selected the same as in the Gold Scenario. This matching strategy gave as output 21560 matched records of which 1336 are false linked matches ie around 6%. The matching threshold was set at  $p\_post=0.8$ . This scenario will be indicated as Silver.

A last scenario, named Platinum, takes the outcome of the Gold Scenario that is the 1300 matched records and within them are run with RELAIS other two matching models using different identificative variables ie address and number of cap. Thus the Platinum scenario is a merge of 3 different matching models giving us an output of 1652 pair matches with 327 false linked and 1325 true linked.

### 3.2. The Results of Discriminant Analysis method

As a first approach it has been used the method indicating with Y the dummy variable of true vs false linked record regressed on only one covariable, X, the  $P\_post$  (post probability, calculated from RELAIS that a record is truly matched)

The first application on the Silver scenario, gave us as a break point  $b=0.9571004$ . In fact applying the b on the p-post we have the following results:

	foracast_true	foreasct_false		P_POST	foracast_false	foreasct_true
TRUE	17555	2669	b=0.9571004	0.8666	706	0
FALSE	16	1320		0.93581	3283	0
				0.99387	0	17571

The above results assure us that if we use as indication the above break-point (just one covariate) the estimated failure of this method is 12,5%  $((2669+16)/21,560)$  while the actual false records that are indicated as true have a rate of 0,1%  $(16/21,560)$ .

Similar results we receive from the second application, still using only one covariance the  $p\_post$  on the Gold Scenario:

	foracast_true	foreasct_false		P_POST	foracast_true	foreasct_false
TRUE	999	256	b=0.9014709	0.66549	17	0
FALSE	0	45		0.79116	64	0
				0.8376	47	0
				0.86571	173	0
				0.99982	0	999

The estimated error in this case is nearly 20% while the rate of false linked matches considered as true is 0%.

Applying the method on the Platinum scenario and thus having three covariate ie the three  $p\_post$  of each model elaborated by RELAIS, the failure rate lowers to 8% however as the break point is a combination of three different models is not easy to handle.

## 4. CONCLUSIONS

The aim of this exercise is to find a method that can help us on the estimation of the errors in the record linkage. In this particular exercise has been took under consideration two main methods the Bellin Fellegi method and the Discriminant Analysis.

The Bellin-Fellegi method gave us no results as the assumptions that must hold in this method are too restrictive and thus not applied in our datasets in hand. The Discriminant Analysis method give us reasonable results that can be considered satisfying.

# The use of uncertainty to choose the matching variables in statistical matching

Marcello D'Orazio ([madorazi@istat.it](mailto:madorazi@istat.it))<sup>1</sup>, Marco Di Zio ([dizio@istat.it](mailto:dizio@istat.it))<sup>1</sup>, Mauro Scanu ([scanu@istat.it](mailto:scanu@istat.it))<sup>1</sup>

**Keywords:** data fusion, Fréchet bounds, multivariate analysis.

## 1. INTRODUCTION: THE STATISTICAL MATCHING PROBLEM

*Statistical matching* (sometimes called *data fusion*, *synthetical matching*) aims at combining information available in distinct sample surveys referred to the same target population. Formally, let  $Y$  and  $Z$  be two random variables; statistical matching aims at estimating the joint  $(Y, Z)$  distribution function (e.g. a contingency table or a regression coefficient) or some of its parameters when: (i)  $Y$  and  $Z$  are not jointly observed in a survey, but  $Y$  is observed in a sample  $A$ , of size  $n_A$ , and  $Z$  is observed in a sample  $B$ , of size  $n_B$ ; (ii)  $A$  and  $B$  are independent and units in the two samples do not overlap (it is not possible to use record linkage); (iii)  $A$  and  $B$  both observe a set of additional variables  $X$  (for major details see [1]).

## 2. HOW TO CHOOSE MATCHING VARIABLES USING UNCERTAINTY

In Statistical Matching (SM) the data sources  $A$  and  $B$  may share many common variables  $X$ . This is the case of matching of data from household surveys where a very high number of variables concerning the household (living place, housing, number of members, etc.) and its members (age, gender, educational levels, professional status, ...) are available. In performing SM, not all the  $X$  variables will be used but just the most relevant ones. The selection of the most relevant  $X$ s, usually called *matching variables*, should be performed by consulting subject matter experts and through appropriate statistical methods.

As far as statistical methods are concerned, the choice of the matching variables  $X_M$  ( $X_M \subseteq X$ ) should be made in a “multivariate sense” [2], to identify the subset of  $X$ s connected, at the same time, with  $Y$  and  $Z$ . This would require the availability of an ideal data source in which  $(X, Y, Z)$  are observed. In the basic SM framework,  $A$  permits to investigate the relationship between  $Y$  and  $X$ , while the relationship between  $Z$  and  $X$  can be investigated in  $B$ . The results of the two separate analyses are then joined and, in general, the following rule can be applied:

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z,$$

where  $X_Y$  ( $X_Y \subseteq X$ ) and  $X_Z$  ( $X_Z \subseteq X$ ) are the subsets of the common variables that better explain  $Y$  and  $Z$ , respectively. The intersection  $X_Y \cap X_Z$  should provide a smaller subset of matching variables if compared to  $X_Y \cup X_Z$ ; this is an important feature in achieving parsimony. For instance, too many matching variables in a distance hot deck SM micro application can introduce undesired additional noise in the final

---

<sup>1</sup> Italian National Institute of Statistics (Istat), Rome.

results. Unfortunately, the risk with  $X_Y \cap X_Z$  is that most of the predictors of one target variable will be excluded if they are not in the subset of the predictors of the other target variable. For this reason, the final subset of the matching variables  $X_M$  is usually a compromise and the contribution of subject matter experts and data analysts is important in order to achieve the “best” subset.

Our proposal is to perform a unique analysis for choosing the matching variables by searching the set of common variables that are more effective in reducing the *uncertainty* between  $Y$  and  $Z$ .

## 2.1. Uncertainty

Due to the nature of the SM problem (i.e.  $Y$  and  $Z$  are never jointly observed) there is an intrinsic uncertainty: there cannot be unique estimates for the parameters describing the association/correlation between  $Y$  and  $Z$ . Approaches, such as maximum likelihood estimation, offer a set of solutions, all with the same (maximum) likelihood, usually closed, known as *likelihood ridge*. The non-uniqueness of the solution of the SM problem has been described in the different articles (see Chapter 4 in [1] and references therein).

Given that  $A$  and  $B$  do not contain any information on  $Y$  and  $Z$ , apart from their association/correlation with the common variables  $X$ , the set of solutions describes all the values of the parameters given by all the possible relationships between  $Y$  and  $Z$ , given the observed data. For this reason, [1] called this set of equally plausible estimates “the uncertainty set”. In order to reduce the uncertainty set, it is necessary to add external information (e.g. edit rules or a structural zero on a cell of the contingency table of  $Y, Z$  or  $Y, Z|X$  reduce the set of possible values).

When  $X, Y$  and  $Z$  are categorical, the uncertainty set can be computed by resorting to the *Fréchet bounds*. Let  $p_{hjk} = \Pr(X = h, Y = j, Z = k)$  for  $h = 1, \dots, H$ ,  $j = 1, \dots, J$ ,  $k = 1, \dots, K$ ; by conditioning on the  $X$ , it is possible to conclude that the probability  $p_{\cdot jk} = \Pr(Y = j, Z = k)$  will lie in the interval:

$$\left[ \underline{p}_{\cdot jk}, \bar{p}_{\cdot jk} \right] = \left[ \sum_h p_{h\cdot} \max(0, p_{j|h} + p_{k|h} - 1), \sum_h p_{h\cdot} \min(p_{j|h}, p_{k|h}) \right]$$

## 2.2. The method: choosing the matching variables by uncertainty reduction

The method proposed for selecting the matching variables when dealing with categorical  $X, Y$  and  $Z$  variables is based on an iterative procedure.

Step 0) initial ordering of the  $X$  variables according to their ability in minimizing the average widths of the bounds for the cell probabilities in the table  $Y \times Z$ :

$$d = \frac{1}{J \times K} \sum_{j,k} (\hat{\bar{p}}_{\cdot jk} - \hat{\underline{p}}_{\cdot jk})$$

computed by conditioning on each of the available  $X$  variables.

Step 1) consider all the possible combinations of the starting variable(s) with each of the remaining ones, according to the ordering provided in step (0) and evaluate the uncertainty associated in terms of  $d$ ; e.g. in the first iteration all the possible

combinations of the first variable, identified in step (0), with the remaining ones will be considered.

Step 2). Select the combination of the variables which determine the higher decrease of the uncertainty ( $d$ ) and go back to step (1).

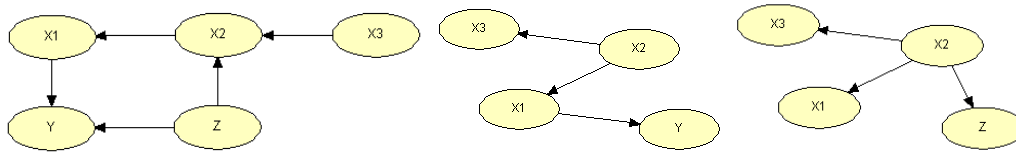
This procedure ends when all  $X$ s have been considered as explanatory of the target ones. The next section presents some results obtained by applying this procedure in two cases.

### 3. APPLICATION AND RESULTS

#### 3.1. Application of the procedure to artificial data

Bayesian networks are used to generate two artificial samples sharing 3 binary  $X$ s. Figure 1 provides the association structure among the variables. The latter two networks denote that  $Y$  ( $Z$ ) depends directly only on  $X1$  ( $X2$ ) when  $Z$  ( $Y$ ) is missing.

**Figure 1. Simulated example: complete (left),  $Y$  given  $X$  (centre) and  $Z$  given  $X$  (right) models**



Application of the step (0) of the procedure presented in Section 2.2. suggests that variable  $X1$  should be considered as the first one ( $d = 0.1703$ ), then there is  $X3$  ( $d = 0.1911$ ) and finally  $X2$  ( $d = 0.2012$ ).

**Table 1. Output of the procedure for selecting the matching variables**

$X$ variables	No. of $X$ s	$d$
$X1$	1	0.1703
$X1*X3$	2	0.1703
$X1*X3*X2$	3	0.1699

The procedure for selecting the matching variables ends with a relatively surprising result (see Table 1):  $X1$  alone is able to achieve quite the best score in terms of average width of the uncertainty bounds; adding  $X2$  does not improve the result and just a negligible decrease of  $d$  is achieved by considering all the  $X$  variables.

#### 3.2. Application of the procedure to artificial data generated from EU-SILC

As a toy example we refer to two artificial samples,  $n_A = 3009$  and  $n_B = 6686$ , generated from the EU-SILC data (data available in [3]). The ordering of the 7 common variables obtained by applying the step (0) is reported in Table 2.

**Table 2. Ability of the  $X$  variables in reducing bounds width**

	c.age	edu7	marital	sex	hsize5	area5	urb
No. of categories	5	7	3	2	5	5	3
$d$	0.0878	0.1056	0.1085	0.1097	0.1120	0.1133	0.1159
Ranking	1	2	3	4	5	6	7



In practice, step (1) starts considering “c.age”, and then adding the variables following the order presented in Table 2. The final output is provided by Table 3.

**Table 3. Output of the procedure for selecting the matching variables**

Combination of X variables	No. of Xs	$d$
c.age	1	0.0878
c.age*sex	2	0.0781
c.age*sex*edu7	3	0.0714
c.age*sex*edu7*area5	4	0.0608
c.age*sex*edu7*area5*hsize5	5	0.0411
c.age*sex*edu7*area5*hsize5*urb	6	0.0225
c.age*edu7*marital*sex*hsize5*area5*urb	7	0.0162

By comparing the results with those that would be obtained by considering all the possible combinations of the  $X$  variables, it comes out that the procedure fails to identify the “best” model with a subset of 4 of the available  $X$  variables. The identified combination precedes immediately the best solution “c.age\*edu7\*area5\*hsize5” ( $d = 0.0575$ ). The same happens when considering 5 of the  $X$  variables (best model “c.age\*edu7\*area5\*hsize5\*urb” with  $d = 0.0385$ ).

The results seem to suggest that the larger the number of matching variables the lower the uncertainty. This reasoning is jeopardized by the fact that with many matching variables increases the sparseness of the contingency tables estimated from  $A$  and  $B$ . This finding suggests that it is necessary to identify a stopping rule for the procedure which should account for the principle of parsimony.

#### 4. CONCLUSIONS

The proposed procedure goes in the direction indicated by [2] avoiding separate analysis on the data sources at hand. The procedure is fully automatic and searches for the best combination of the available categorical common variables; it appears successful in identifying the various subsets of 1, 2, 3, etc. “best” matching variables. The procedure at this stage lacks of a stopping rule which should be developed accounting for the parsimony principle.

#### REFERENCES

- [1] D’Orazio M., M. Di Zio, M. Scanu. *Statistical matching: Theory and Practice*. Wiley, Chichester (2006)
- [2] Cohen M.L. “Statistical matching and microsimulation models”, in Citro and Hanushek (eds) *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Vol II Technical papers*, Washington D.C. (1991)
- [3] D’Orazio M. “StatMatch: Statistical Matching (aka data fusion)”. R package version 1.2.2 (2014) <http://CRAN.R-project.org/package=StatMatch>

# **Environmental conditions / behaviour and income – statistical matching of EU-SILC and Micro-census Environment**

Alexandra Wegscheider-Pichler ([Alexandra.wegscheider-pichler@statistik.gv.at](mailto:Alexandra.wegscheider-pichler@statistik.gv.at))

**Keywords:** Statistical Matching, environmental conditions, environmental behaviour, household income

## **1. INTRODUCTION**

The project “environmental conditions / behaviour and income” allowed for the first time to display for different household income groups their affection by environmental domains or their environmental behaviour, using official data sources.

The micro-census special programme "Environmental conditions and environmental behaviour" by Statistics Austria contains widespread data material concerning ecological issues. The influence of income on the collected environmental characteristics is commonly assumed but could not be confirmed because the variable “income” is not part of the micro-census survey. If education and employment status are used as approximations, it can be assumed that the income of households is a crucial factor, for example for the purchase of organic products [Baud- Milota, 2011, p 77].

The project used statistical matching to add income variables from EU-SILC 2011 to the data of the micro-census environment 2011. This allowed analysing the correlation between environmental behaviour as well as environmental impacts (e.g. noise, dust) with the overall household income of the people interviewed.

One focus of the project was the comprehensive presentation of the used method for the statistical matching process. Furthermore insights into advantages or problems of the method of statistical matching are gained.

## **2. METHODS**

Income variables from EU-SILC 2011 were inserted to the data-set of the micro-census environment 2011 by statistical matching. Some focus was laid on the selection and adjustment of the variables used to link the two data files (= connecting variables). The comparability and homogeneity of the variables used are essential for the quality of the statistical matching [see Eurostat, 2013, p.13].

The overall household income was defined as the relevant variable from EU-SILC 2011 to be inserted in the micro-census data file. Additionally the net income from employment was used for two purposes: to improve the matching process and to evaluate the matching process. The review of the different matching variants represented a substantial part of the project.

### **2.1. Net income from employment**

The micro-census Labour Force Survey (LFS) contained (retrospectively) the variable net income from employment through administrative data. This variable was included in the matching process and led to a distinct improvement of the data linking. On the other hand it was possible to evaluate the different matching variants with the variable net income.

In the original files data files micro-census LFS and EU-SILC it is possible to calculate a “household net income from employment” (net income of all persons of a household aggregated). In this way each person of the household can be assigned a variable “household net income”. This variable was used for the matching process as well.

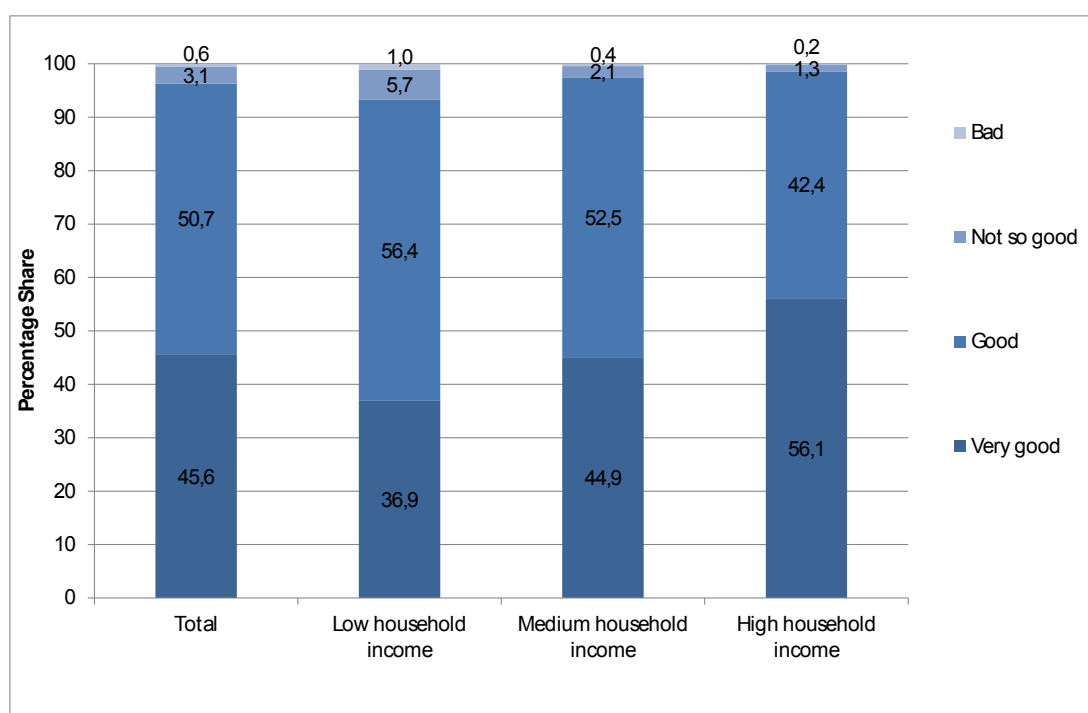
## 2.2. Statistical matching process

For the statistical matching a distance-based procedure was used. For each respondent of the micro-census environment a donor from the data from EU-SILC with minimal distance was assigned. In case of several donors with the same distance one was randomly selected. Each respondent from EU-SILC was accepted only once as donor for the micro-census.

Five different matching variants were computed to find the best option to link the variable “total disposable household income”. They differ according to the used distance function which depends on the selected connecting variables and the used weight.

For data analyses the statistical matching variant was chosen, which used the net income from employment within the household context.

## 3. RESULTS



**Figure 1. Assessment of the subjective quality of life by income (tercile)**

A distinction of responses by household income revealed significant differences in the assessment of quality of life: While 56.1% of respondents with high household income defined their quality of life as high, this was true for only 36.9% of those with low household income. Persons with low household income stated to nearly 7%, that their quality of life was poor or very poor; this information was given only by 1.5% of the group with a high household income. Persons with medium household income rated their quality of life to 44.9% as very good, 2.5% as poor or very poor.

Further findings:

43.6% persons with low household income were affected by noise disturbance, while only 35.7% of high household income earners were affected by noise.

People with a low household income were least likely to use their car for daily commuting (26.3%), medium and high income had more than 40% daily use of car.

People with high household income answered more frequently to buy "often" or "sometimes" organic food, as people in the middle or low income group.

#### **4. CONCLUSIONS**

Statistical matching still is a relatively new model-based approach to combine statistical information from at least two sources. Some advantages of the method are the cost reduction and the reduced burden on respondents [Eurostat, 2013]. Nevertheless the variables are synthetically generated “statistical twins” and not actual observations. The values obtained can thus be distorted. They depend highly on the “connecting variables”.

Therefore, much emphasis has to be laid on the comparison of the variables and the data evaluation. In this case, the data generation with statistical matching can lead to interesting results. In this project, some expectations, described in relevant literature, were confirmed by the data obtained: Environmental conditions as noise and dust showed a correlation with income. Also the connection between environmental behavior and income has been confirmed for several aspects.

#### **REFERENCES**

- [1] S. Baud and E. Milota, Umweltbedingungen, Umweltverhalten 2011, Ergebnisse des Mikrozensus“, Statistics Austria publication, (2013), Vienna.
- [2] Eurostat, Statistical matching: a model based approach for data integration, Methodologies and Working papers (2013), ISBN 978-92-79-30355-2 Luxembourg.

# The role of the auxiliary information in Statistical Matching Income and Consumption

Gabriella Donatiello ([donatiel@istat.it](mailto:donatiel@istat.it))<sup>1</sup>, Marcello D'Orazio ([madorazi@istat.it](mailto:madorazi@istat.it))<sup>2</sup>, Doriana Frattarola ([frattarola@istat.it](mailto:frattarola@istat.it))<sup>1</sup>, Antony Rizzi ([anrizzi@istat.it](mailto:anrizzi@istat.it))<sup>1</sup>, Mauro Scanu ([scanu@istat.it](mailto:scanu@istat.it))<sup>3</sup>, Mattia Spaziani ([mspaziani@istat.it](mailto:mspaziani@istat.it))<sup>1</sup>

**Keywords:** Statistical matching, Integration of surveys

## 1. INTRODUCTION

The need of new indicators that cover cross-cutting information on households economic well-being is among the current priorities of the National Statistical Institutes as well as a major goal at European level. The current process of modernization of social surveys is going towards a better integration and coordination of surveys also in order to facilitate their integration through statistical matching procedures. One of the aim of this work is to evaluate the possibility of integrating two different data sources in order to provide joint information on household income and consumption expenditures in Italy at the micro level. For this goal, we use EU-SILC 2012, with income reference year 2011, and the HBS (Household Budget Survey) 2011.

The paper focuses on the role of the auxiliary information in improving the matching outputs and overcoming the underlying assumptions. In fact, most of the statistical matching methods proposed in literature assume (i) conditional independence (CI) of the target variables given the common variables and, (ii) the observations in the available samples are independent and identically distributed (i.i.d.). Conditional independence is a very limiting assumption that rarely holds in practice. The only way to overcome the CIA is to introduce some auxiliary information in the matching procedures and this work highlights the advantages in using a reconstructed HBS household income for this purpose. The i.i.d. assumption is also difficult to be maintained when matching data from complex sample surveys. In such a case, the Renssen's approach, based on calibrations of the weights [1], seems more flexible and suitable especially when the variables under study are categorical. Such approach is also promising in matching HBS with EU-SILC.

It should be noted that an *ex-post* integration of existing micro data sets has to face several challenges due to the lack of information on wealth/consumption in SILC. In this work we prove that an important source of auxiliary information could come from the introduction of a small number of questions on food consumption and transport in SILC. Through the identification of those consumption components that are good predictors for total consumption in HBS we show how the introduction of few additional consumption variables in SILC could add valuable information on total consumption expenditures; this information can be used as additional auxiliary information in the matching process.

---

<sup>1</sup> Italian National Institute of Statistics (ISTAT), Socio-economic Statistics Directorate, Italy.

<sup>2</sup> Italian National Institute of Statistics (ISTAT), Structural Economic Statistics on Enterprises and Institutions, International Trade and Consumer Prices Directorate, Italy.

<sup>3</sup> Italian National Institute of Statistics (ISTAT), Development of Information Systems and Corporate Products, Information Management and Quality Assessment Directorate, Italy.

## 2. METHODS

It is well known that statistical matching (SM) procedures usually refer to a broad range of model-based techniques that generally aim to achieve a micro data file from different sources that have a set of variables in common but do not contain the same units or the same identifier. In the basic Statistical matching framework, the surveys to integrate, denoted as  $A$  and  $B$ , share a set of variables  $X$ , while the variable  $Y$  is observed only in  $A$ , and the variable  $Z$  is observed just in  $B$ . The final objective of SM is to explore the relationship between  $Y$  and  $Z$ . Integration at micro level can be obtained by limiting attention to a given data set (say  $A$ ) and imputing in it the missing variables ( $Z$  in this case). Many of the techniques proposed for SM at micro level are based on methods developed for the imputation of missing values: parametric (e.g. regression imputation), nonparametric (hot deck imputation) or mixed methods (e.g. methods based on predictive mean matching). Hot deck procedures consist in filling in the missing variable in the data set chosen as the *recipient* by using the other data set as the *donor*. The donation is typically based on the variable,  $X$ , available in both the data sets. Commonly encountered hot deck procedures for SM are: *random hot deck*, *nearest neighbor hot deck* and *rank hot deck* [2].

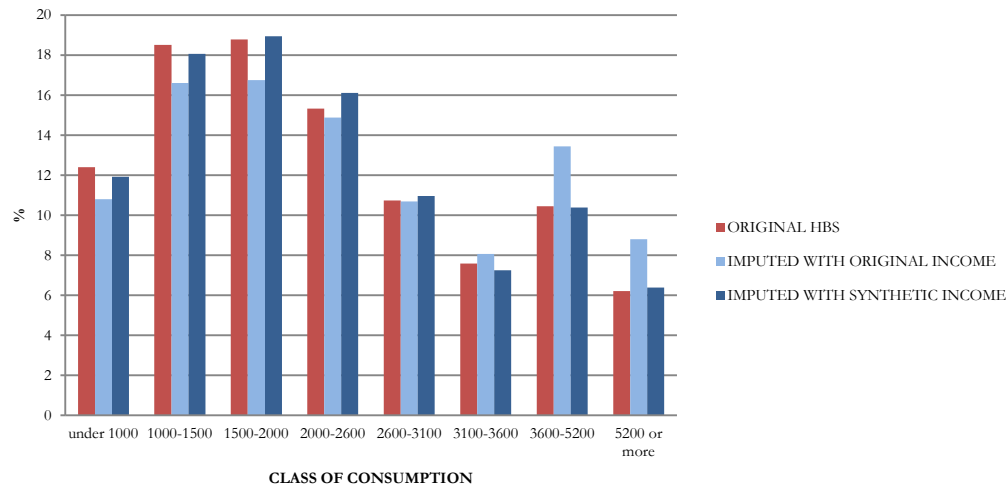
In random hot deck, the donors are chosen at random, but this choice is usually carried out within opportune subsets of donors: those sharing the same characteristics of the recipient records, e.g. in terms of geographical area, gender, typology, etc. This method is particularly suited when dealing with categorical  $X$  variables; in such case, random hot deck consists in estimating the conditional distribution of  $Z$  given  $X$  and then drawing an observation from it. A particular version of random hot deck permits to exploit auxiliary information in reducing the subset of the potential donors.

## 3. RESULTS

In order to obtain a synthetic micro data set, we use HBS as a donor data set and impute consumption classes in SILC; a first step in our matching procedure consists in the application of random hot deck under CIA using the R package StatMatch [3]. Then the exploration of SM uncertainty is also applied by calculating the Fréchet bounds for the contingency table between the variables of interest given the two common variables being considered. To overcome the CIA we use all available information included in the HBS *ad hoc* section about income and savings. The HBS income variable does not clearly have the same quality level of EU-SILC income variable. For this reason, we estimate a new income variable in order to reduce the large income discrepancy. The reconstructed income variable shows a decrease in the HBS household income underestimation, and its marginal distribution is closer to income distribution in EU-SILC.

It is worth noting that the auxiliary information concerning the reconstructed income is applied in the matching process in further restricting the subset of potential donors of the random hot deck procedure. The comparison between the imputed consumption classes in SILC using the original HBS income or using the new HBS income variable as auxiliary information displays a valuable improvement of the estimates in consumption highest classes (Figure 3.1), and, in general, the marginal distribution of the imputed consumption is really close to the reference one [4]. Moreover, the unlikely assignments between classes of consumption and income are also rather limited. In other words, there is a significant decrease of those frequencies corresponding to classes of consumption that differ more than three from the respective class of income.

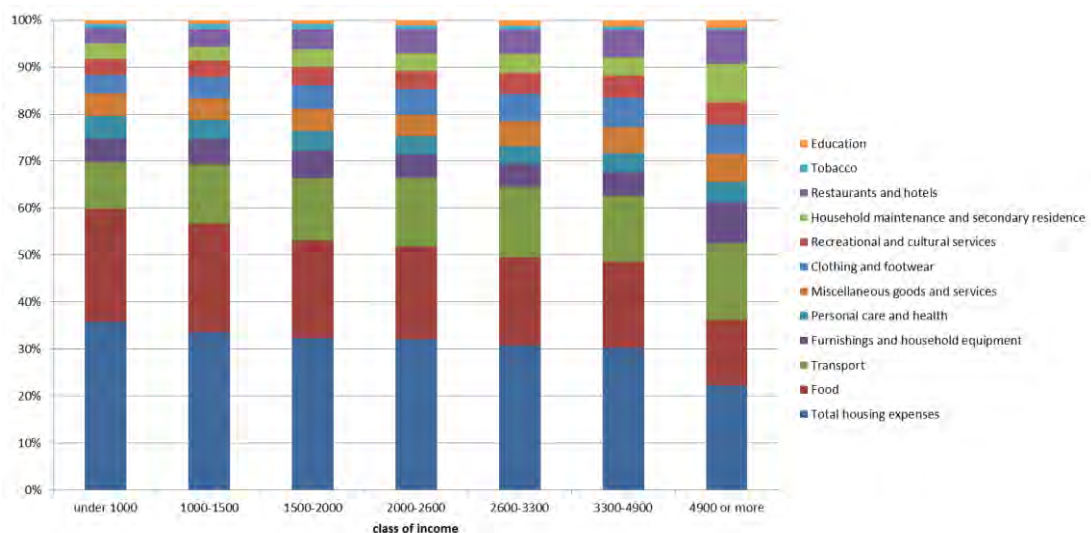
**Figure 3.1 Comparison of class of consumption imputed with synthetic variable and original variable**



In order to enhance SM of income and consumption, an ex-ante collection of information on wealth/consumption in SILC could provide new shared variables with high predictive power useful for matching procedures. Analysing the structure of consumption, we identify those components representing good predictors for total consumption. Afterward the explanatory power of each amount is investigated using a statistical model.

Looking at distribution for different income classes (Figure 3.2), we realize that there are three components (total housing costs, food, transport) contributing to 63% of total consumption. As expected the share that people reserve to food consumption decreases as income increases: from 24% for the first class of income to 14% for the last and richest class. A similar trend observable in the total housing expenses is mainly due to decreasing amount of rent payment for higher classes of income.

**Figure 3.2 HBS main consumption components by income classes**



Some simulations on HBS using different methods of classification, are performed to build a rule for identifying different consumption components, in order to allocate observations to the estimated classes. Comparing the overall classification error between models and covariates, it turns out that all the models identify the same set of variables

(food and transport expenditures among the common variables). The best model classifies correctly 56.3% of total households in HBS survey.

#### 4. CONCLUSIONS

The primary object of our work is to underline the importance of the auxiliary information in improving the estimates accuracy in statistical matching procedures. The availability of few valuable questions about the use of the household income in HBS (e.g. consumption and savings) has allowed us to reconstruct new income variable. The inclusion of one or two questions on savings in HBS can then be useful for data integration purposes, as well as for improving the quality of information on household monthly income. Similarly, another source of auxiliary information could come from the introduction of a few set of questions on food consumption and transport in SILC. These variables have a great potential and explanatory power so it can be useful to use this information for estimating a total expenditures variable in SILC to be used as auxiliary information able to improve the quality of final estimates.

#### REFERENCES

- [1] Renssen, R. H. (1998), “Use of Statistical Matching Techniques in Calibration Estimation”. *Survey Methodology*, No 24, pp. 171-183.
- [2] D’Orazio, M., Di Zio, M., Scanu, M. (2006), *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.
- [3] D’Orazio, M. (2014), *StatMatch: Statistical Matching (aka data fusion)*. R package version 1.2.2. <http://CRAN.R-project.org/package=StatMatch>.
- [4] Donatiello, G., D’Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., Spaziani, M. (2014) “Statistical Matching of Income and Consumption Expenditures”. *International Journal of Economic Sciences*, Vol. III(3), pp. 50 – 65.



# Improved Time-varying Day Adjustment in SEASABS

Jonathan Campbell and Lujuan Chen, Australian Bureau of Statistics

**Keywords:** Time-varying trading day adjustment, constrained regression, spectral analysis.

## 1. INTRODUCTION

Trading day variation is an effect present in many time series due to the differing number of each type of weekday in each month or quarter. When performing seasonal adjustment it is necessary to remove such effects to ensure comparability of estimates from one period to another. X11 and its variants use a regression method to obtain estimates of the activity levels associated with each day of the week, and use these daily weights to construct adjustment factors for each month or quarter. This method has the restriction that the daily weights are considered to be the same across the entire span of the time series, ignoring social and behavioural changes that may have occurred and affected the activities associated with each day of the week. Furthermore, the regression estimates used may result in negative daily weights, which cannot be clearly interpreted.

The Australian Bureau of Statistics seasonal adjustment package SEASABS contains improvements to the standard X11 trading day regression to deal with these issues. SEASABS allows the estimated daily weights to vary over time by performing multiple trading day regressions on sub-spans of the data. The daily weights are smoothed so that daily weight estimates are available for every year in the data span. SEASABS also checks for negative daily weights, and performs a modified regression to remove them if any occur.

This paper describes the trading day adjustment methods and algorithms used by the ABS for monthly series.

## 2. METHODS

Early methods of trading day adjustment were based on external information about the amount of activity associated with each day of the week, but such information was found to be unsuitable for the purposes of trading day adjustments [1].

The standard technique used to identify, estimate and remove trading day variation is a regression method introduced by Young [1] and fully expounded by Ladiray and Quenneville [2]. It attempts to identify systematic changes in the irregular component of the series associated with the daily composition of the month.

Moving trading day performs a static trading day regression on each 7 year span to obtain daily weight estimates for each year, and shorter spans of complete years are used at the ends of the series where a seven year span is not available. Within each separate span, the static trading day regression is performed, including the calculation of an overall standard deviation, removing extremes, and constructing the  $Y_t$  and  $Z_{it}$  values to perform the regression. These are performed exactly as for static trading day, on each separate span independently.

Each distinct span is ‘centred’ on a particular calendar year, so that the estimates from that particular regression are for that particular year. Having obtained estimates of the daily weights for each year, the moving trading day algorithm smoothes these daily

weights in a manner similar to smoothing seasonal factors, and uses these smoothed daily weights to construct trading day adjustment factors which are unique to each year. This method allows the daily weights to develop and change over time, so that the trading day adjustment factors are better suited to the year they are adjusting. Seven years of data is used since it was found to allow movement in the daily weight estimates, yet still provide estimates of sufficient quality (Sutcliffe (1999) and (2003) [3][4]). Since the calculation of moving trading day requires seven year spans, it requires a minimum of seven years of data before it can be applied.

## 2.1 Split level trading day

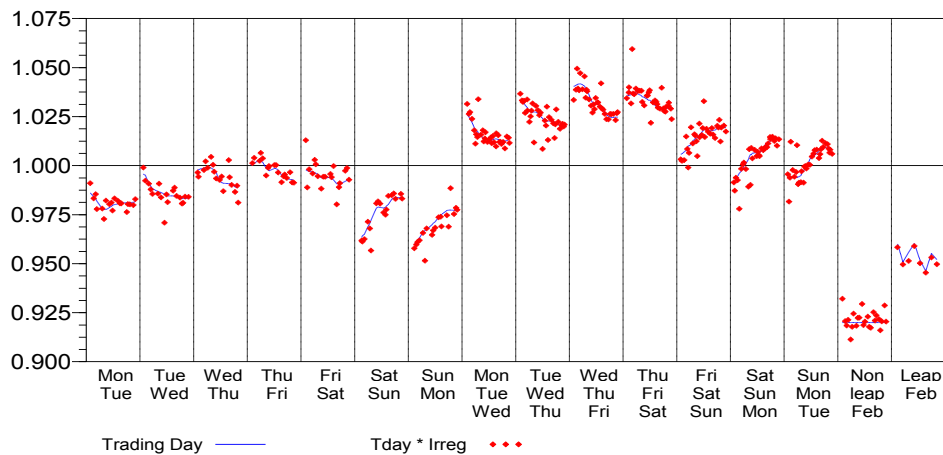
Split level trading day allows the user to specify multiple sets of prior daily weights, so that there are different sets of daily weights for distinct spans of the data. These different sets of daily weights are used to construct the prior trading day factors  $P_t$  in table A4b, and are removed as a prior adjustment. These prior factors will change across the ‘split level’ since the daily weights will change. Thus a 30 day month with an extra Saturday and Sunday before the split will have a different factor from an identical month after the split.

## 2.2 Modified regression

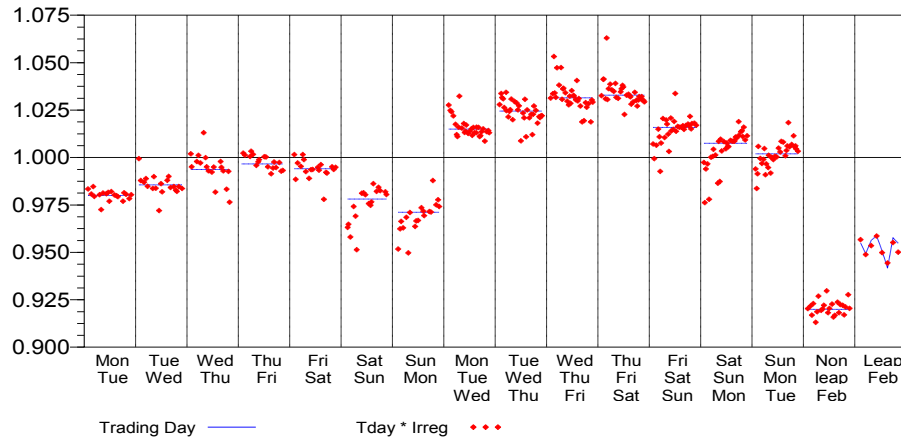
The validity of allowing negative daily weights is unclear. On the one hand, if the regression results in negative daily weights, then they are the best estimates for that regression model, and will describe the trading day variation best. On the other, a negative daily weight has no clear economic interpretation, except possibly for unusual reporting practices. SEASABS offers a different regression method, which ensures that the combined daily weights are non-negative. The method used to ensure the weights are non-negative is a basic one, and could perhaps be replaced by more complex methods, given the increased computing power now available.

## 3. RESULTS

The effects of using moving trading day rather than static trading day will be demonstrated with reference to the Australian level retail series. The series used for this analysis covers the period April 1962 to September 2014, and was analysed using SEASABS v3.0.



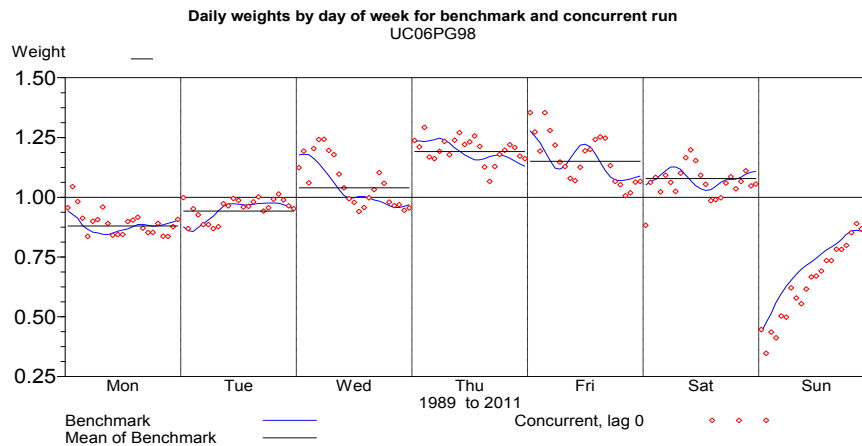
**Figure 1. Moving trading day DI chart**



**Figure 2. Static trading day DI chart**

Most evident from these DI charts are that those pairs and triplets containing Sundays are significantly higher (at the current end) under the moving trading regime as compared to static trading day, reflecting the increase in Sunday trade experienced in Australia over the past decades.

This is even more evident when examining the individual daily weights in the figure below.



**Figure 3. Comparison of trading day factors estimated static and dynamically.**

An important aspect of this improved daily weight estimation procedure is the reduction of residual trading day variation when implementing moving trading day as opposed to static trading day. Indeed, under a static trading day adjustment, with length of month, the results of the adjustment give a warning that there are patterns in the irregular, and suggests the presence of residual trading day variation.

#### 4. CONCLUSIONS

SEASABS offers a combination of procedures to perform trading day and length of month adjustments. Although based on the X11 trading day regression and adjustment,

there are notable differences in how the weights are estimated and adjustment factors obtained.

The weights can be estimated so that they are able to evolve over time, an important option since the ABS uses the entire span of a series to perform seasonal adjustments. SEASABS also offers a modified regression that ensures least squares like estimates of the daily weights be non-negative. Apart from the differences in the regression methods, the construction and application of the trading day adjustment factors is slightly different from their counterparts in X11. SEASABS always inserts the length of month or length of February correction as a prior trading day adjustment. This factor is maintained throughout since the combined trading day factors are calculated as the combination (multiplication or addition) of the prior and regression trading day factors, whereas X11 combines the regression and prior daily weights to construct the combined factor.

## REFERENCES

- [1] A.H Young, Estimating Trading-Day Variations in Monthly Economic Series, Technical Paper 12, Bureau of the Census, U.S. Department of Commerce
- [2] D Ladiray and B Quenneville, Seasonal Adjustment with the X-11 Method, Springer-Verlag, 2001, New York
- [3] A Sutcliffe, Moving Trading Day, Internal document (1999), Australian Bureau of Statistics, Canberra
- [4] A Sutcliffe, Issues Relating to Trading Day Estimation and Related Topics (2003), Internal document, Australian Bureau of Statistics, Canberra

# 1 out of 20 possible scenarios: how to perform temporal disaggregation of annual sector accounts data

**Dario Buono**, Eurostat – European Commission  
**Filippo Gregorini**, Eurostat – European Commission  
**Enrico Infante**, Eurostat – European Commission

**Keywords:** Temporal disaggregation, benchmarking, mathematical approach, regression approach.

## 1. INTRODUCTION

In the context of the European Union (EU hereafter), Member States (MS hereafter) are requested to provide Eurostat with sector account data on annual and quarterly basis. MS whose contribution to EU GDP is below 1% (hereafter “small”) have to provide less figures compared to MS above 1% (hereafter “big”) threshold. Therefore in such cases, for “small” countries only annual figures are available, but quarterly estimates are needed for the production of quarterly aggregates of Euro Area and EU.

In this paper we build on the existing literature references, and look for the best practice for an empirical application to estimate those missing quarterly series for the “small” EU countries.

## 2. METHODS

Our problem can be summarized as follows.

<i>Target series to be published:</i> Quarterly EU28 aggregate figures, for which at least three possible options are here considered.
---

<i>Series to be estimated:</i> (A) Quarterly EU28 figures, assuming that missing 5 countries behave like the aggregate of the other 23 MS for which we have quarterly figures (call it “EU23”); (B) Quarterly figures per each of the 5 missing countries; (C) Quarterly figures for the 5 missing countries aggregated (call it “EU5”).
---

<i>Needed related proxies or related indicators:</i> If (A) is to be estimated, we might use: (1) Quarterly EU23 data. If (B) is to be estimated, we might use: (1) related available quarterly figures for each missing country (e.g. gross disposable income). If (C) is to be estimated, we might use: (1) Quarterly EU23 data; (2) Seasonal component of Quarterly EU23 series; (3) Trend and irregular component of Quarterly EU23 series.
---

In our exercise we are simultaneously constrained by at least two necessary conditions (user requirements) to be satisfied.

On the one hand, the respect of the temporal constraint, which can be addressed either by applying mathematical methods (such as the naïve method of dividing annual figures by 4, Denton and Wei-Stram) or regression based methods (such as Chow-Lin, whose features are mentioned in the Eurostat Handbook on Quarterly National Accounts 2013 edition). The regression approach of Chow-Lin exploits the statistical relationship between low frequency and high frequency data via a standard regression equation subject to temporal aggregation constraint. The key advantage of regression approaches compared to mathematical ones like Denton, Denton modified or Wei-Stram is the possibility to measure the goodness of fit.

On the other hand, quarterly figures need to be benchmarked to existing annual figures (accounting constraint, to be ensured by the alternative available methods for such as Denton modified, Cholette-Dagum-Bee and Cholette multivariate).

In addition to these, it is worth to be mentioned the innovative method developed by Di Fonzo and Marini, which takes into account both the two constraints.

To develop our exercise, several software applications are available. With JDemetra+ all the models above can be tested, with the exception of Denton, Wei-Stram, Di Fonzo-Marini. Di Fonzo-Marini, in addition to the models available on JDemetra+, can be tested on JEcotrim. JEcotrim also enable to display Denton and Fame can be useful to derive estimation based on Denton modified.

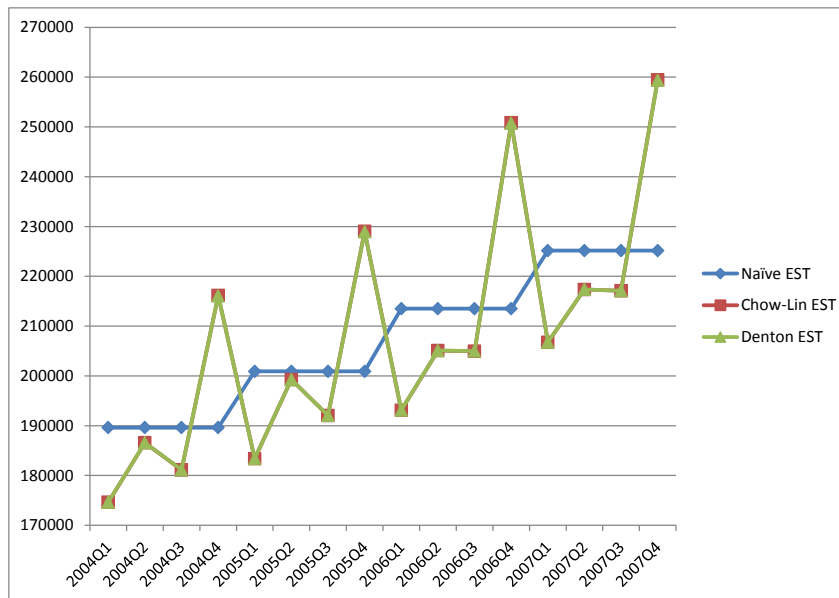
Given the available inputs, the models and the available software described above, and taking into account the peculiarity of each method, approximately 20 scenarios can be addressed to solve our exercise. Beyond that, during production rounds, the time constraint plays a decisive role in order to choose between available methods and software.

### 3. RESULTS

Our exercise (to be extended with other methods & software) provides estimates based on naïve (annual value divided by four), mathematical (Denton modified) and statistical (Chow-Lin) approaches for case (A), where quarterly EU28 figures are estimated, assuming that 5 missing countries behave like the aggregate of the other 23 MS for which we have quarterly figures.

Mathematical and statistical approaches are displayed here using JDemetra+ software.

TIME/SERIES	Naïve EST	Chow-Lin EST	Denton EST	DIFF Chow-Lin / Denton
2004Q1	189608	174618	174648	-30
2004Q2	189608	186568	186572	-4
2004Q3	189608	181108	181138	-30
2004Q4	189608	216138	216073	64
2005Q1	200920	183365	183405	-40
2005Q2	200920	199268	199268	0
2005Q3	200920	192035	192064	-29
2005Q4	200920	229012	228943	69
2006Q1	213494	193096	193139	-43
2006Q2	213494	205076	205094	-18
2006Q3	213494	204970	204996	-26
2006Q4	213494	250833	250747	87
2007Q1	225149	206747	206785	-38
2007Q2	225149	217341	217358	-17
2007Q3	225149	217087	217110	-23
2007Q4	225149	259420	259342	78



On the one hand, it is apparent that Denton modified and Chow-Lin estimates are similar. On the other hand, it has to be pointed out that the regression approach allows us to properly test the goodness of fit of the estimates.

#### Estimated regression coefficients

	coefficient	s.e.	t-stat	p-value
Intercept	-565.720366	217.920284	-2.595997	0.060925
REG_1	1.533525	0.001572	975.787816	0.000001

#### Regression diagnostics

Name	Value	Description
Number of valid cases	4	Number of observations of temporally aggregated time series
Degrees of Freedom	2	Number of observations (N) less the estimated coefficients (k)
Coefficient of determination (Buse, 1973)	1.00	Is a generalized R-squared statistic proposed by Buse (1973) in case of GLS estimation
Adjusted R-squared	1.00	Measure that imposes a small penalty in R-squared when a variable is added to the model
Standard Error of Regression	12.11	Measure of variability
Sum of Squared Totals	139536880.75	Measure of the total variation in dependent variable
Sum of Squared Residuals	293.09	Measure of the unexplained variation of the dependent variable
Sum of Squared Estimates	139536587.66	Measure of the explained variation of the dependent variable
Log-likelihood	-22.53	Log-likelihood function of the model
F-statistic	952161.86	Calculates F-statistic only if there are more than 1 regressor
Probability (F-statistic)	0.00	Displays the p-value corresponding to the reported F-statistic. Measures the significance of the F-statistic
Akaike Information criterion	5.29	Measure of the explanatory capability of the model proposed by Akaike (1974)
Schwarz Information criterion	4.99	Measure of the explanatory capability of the model following a Bayesian approach suggested by Schwarz (1978)
Durbin-Watson statistic	1.69	Measure of the first-order autocorrelation

## 4. CONCLUSIONS

Our exercise shows that with up-to-date software, different approaches are fast and easily replicable, and therefore useful for production purposes, in particular in Eurostat Units involved in production of European aggregates.

Chow-Lin statistical model is as fast as mathematical models to be displayed but also has the key advantage of being a statistical method; that is, its regression analysis enable us to measure the goodness of fit of the estimates.

## 5. REFERENCES

- (1) Chamberlain, G. (2010) Temporal Disaggregation. *Economic & Labour Market Review*, Nov 2010, 106-121.
- (2) Cholette, P.A., Dagum, E. Bee (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods of Time Series*, Lecture Notes in Statistics 186, Springer Science+Business Media, LLC, New York.
- (3) Chow, G. and Lin, A. (1971). Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. *The Review of Economics and Statistics*, 53, 4, 372-375.
- (4) Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. *Journal of the American Statistical Association*, 66 (333), 99-102.
- (5) Di Fonzo, T. and Marini, M. (2011), Simultaneous and two-step reconciliation of systems of time series: methodological and practical issues, *Journal of the Royal Statistical Society*, Vol. 60/2, pp. 143-164.
- (6) Eurostat (2013 edition) *Handbook on Quarterly National Accounts*, European Commission.



# Forecasting Evaluation with *JDemetra+*:

An application to Flash GDP growth

David de Antonio Liedo<sup>1</sup>

Jean Palate

RESEARCH & DEVELOPMENT

Statistics Department

National Bank of Belgium<sup>2</sup>

**Keywords:** nowcasting, dynamic factor models, TRAMO-SEATS, forecast accuracy, encompassing tests

## 1. INTRODUCTION

This paper presents an innovative forecasting evaluation library that takes into account the calendar of macro-economic releases in order to provide a realistic simulation of out-of-sample forecast errors in multivariate and univariate time series models. The library is meant to be an independent module in *JDemetra+* (*JD+*), which is an open source, and extensible software written in Java for time series analysis. *JD+* is mostly used in statistical agencies for the analysis of seasonality, since it enables the implementation of the ESS Guidelines on seasonal adjustment, but it supports the use of several other time series methods.

The forecasting evaluation library presented here complements the in-sample based measures of fit that are often reported. One key advantage with respect to typical evaluation exercises described in the literature is that it allows us to replace the standard concept of forecasting horizon by “information scenarios”, which are more suitable for real-time situations. Thus, we allow users to evaluate the forecasts obtained under pre-specified information assumptions that mimic the availability of data in real time forecasting scenarios. As an input, the algorithm requires the approximate publication delay for each one of the time series that enters the model. The library incorporates multiple tests to assess the statistical significance of relative forecast accuracy measures, as well as the concept *fixed-event* forecasts. The latter will enable us to understand how the accuracy of different forecasting models improves over time as they approach the actual realization.

The empirical application will start with a brief introduction to the methodology to estimate the Flash GDP for a given quarter and the role played by the data on VAT returns and industrial production at a disaggregated level<sup>3</sup>. Those indicators are not available for the three months of the quarter. In this context, alternative indicators, such as surveys, could be considered. Survey data represent *soft* information regarding the recent and expected evolution of several sectors related to industry, services, trade, and construction activities. Because of their timeliness, they can help to forecast the hard data on industrial production and VAT returns that is missing for a given quarter. Thus, we will compare the accuracy of multiple models at forecasting the missing information required for the construction of the Flash release. The models included in the forecasting competition range from multivariate

---

<sup>1</sup> Corresponding author: david.deantonioledo@nbb.be

<sup>2</sup> This paper should not be reported as representing the views of the National Bank of Belgium (NBB). The views expressed are those of the authors and do not necessarily reflect those of the NBB.

<sup>3</sup> This empirical approach based on disaggregate data is different from the model developed by De Antonio Liedo (2014). He proposes to monitor on real GDP growth Belgium and the euro area mostly relying on composite indicators obtained from surveys and hard data corresponding to relatively broad economic sectors.

dynamic factor models<sup>4</sup> to univariate ARIMA models, which are currently an important element of the current procedure used for the early estimation of the Flash. The JD+ forecasting evaluation tool will play a fundamental role at determining which strategies provide the most significantly accuracy gains. The results provide additional and valuable information regarding the reliability of the alternative models considered beyond the in-sample statistics often reported, which do take into account the quality of the predictor variables, but not their timeliness.

## 2. METHODS

### 2.1. Forecast Accuracy Measures

By default, we focus on the forecasts in levels  $Y_t$ . The aim is that the forecasts  $F_t$  is such that the resulting forecast errors,  $e_t = Y_t - F_t$ , are as small as possible over the history. The following measures of accuracy allow researchers to compare the forecasts resulting from different methods, including the Mean Square Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MdAE) and Relative measures of forecast accuracy. For instance, the relative RMSE of a given model will be expressed as  $relRMSE = \frac{RMSE_{model}}{RMSE_{benchmark}}$ . Hyndman and Koehler (2006) provide a critical view of the different forecast accuracy measures. Here, we aim to provide the user with tools to have a deeper understanding of the relative merit of all the methods and their statistical significance. The measures that we have defined can be computed for several subsamples in order to make sure the superiority of a given method remains stable.

### 2.2. Forecast Accuracy Tests

Forecasting models that turn out to have an excellent in-sample fit because of their large number of parameters may bring about very volatile forecasts, which could be translated into large forecast errors. Therefore, the analysis of the forecasts obtained in real time could add some valuable information for model comparison in finite samples beyond that of the usual in-sample diagnostic statistics. This kind of model comparisons, however, may require several years until sufficiently large time series of truly out-of-sample forecast errors are available. JD+ helps by reconstructing (pseudo) out-of-sample forecasts from all the models under consideration and implementing a number of widely used tests to evaluate the statistical significance of the results.

#### 2.2.1. Diebold-Mariano

Define two time series of forecast errors as  $e_{1,t} = Y_t - F_{1,t}$  and  $e_{2,t} = Y_t - F_{2,t}$ . Without the need to specify these forecasts errors in terms of a model or parameters, Diebold and Mariano (1995) proposed to test the null hypothesis of equal predictive accuracy in terms of a loss function differential. Under a quadratic loss function, for instance, the loss differential would be defined as  $d_t = e_{1,t}^2 - e_{2,t}^2$ . Alternatively, a loss function that considers the absolute value of the error would yield the following loss

---

<sup>4</sup> Charles et al. (2014) have recently developed a nowcasting module inside the JD+ environment, following the literature on factor models for short-term analysis along the lines of Banbura and Modugno (2012).

differential:  $d_t = |e_{1,t}| - |e_{2,t}|$ . Thus, under the null hypothesis of equal accuracy,  $d_t$  will be zero on average  $E[d_t] = 0$ . More precisely, if we are willing to admit that  $d_t$  is covariance stationary, the test statistic can be calculated by regression of the loss differential on an intercept, using heteroskedasticity and autocorrelation robust (HAC) standard errors.

### 2.2.2. Encompassing Tests

In the event that both forecast errors turn out to be significantly different, one cannot discard the possibility that the poorly-performing forecast provides some marginal information that is not contained in the more accurate forecast, say  $F_{1,t}$ . In such case, for  $\lambda \in [0,1]$ , the error  $e_{c,t}$  resulting from the forecast combination  $F_{c,t} = (1 - \lambda)F_{1,t} + \lambda F_{2,t}$  will yield a forecast error with smaller variance than that of  $e_{1,t} = Y_t - F_{1,t}$ . This will not be the case when the covariance between  $e_{1,t}$  and  $(e_{1,t} - e_{2,t})$  is zero. Thus, the null hypothesis  $\lambda = 0$  can be interpreted in terms of the statement that the second forecast does not add any further forecast accuracy gain, i.e.  $F_{2,t}$  is encompassed by  $F_{1,t}$ . Thus, Harvey et al. (2008) have proposed to formulate the test statistic as in Diebold and Mariano (1995), by formulating the null hypothesis in terms of  $E[d_t] = 0$ , where  $d_t = e_{1,t}(e_{1,t} - e_{2,t})$  instead of the loss differential. This is equivalent to the use of heteroskedasticity and autocorrelation robust (HAC) standard errors for the parameter  $\lambda$  in the regression  $e_{1,t} = \lambda(e_{1,t} - e_{2,t}) + e_{c,t}$ .

### 2.3. A new definition of forecasting horizon

Here, we define the concept of forecast horizon in terms of the information set that is available to the forecaster at the point in time the forecast is computed. Thus, rather than specifying the simple parameter “h” that would typically stand for “horizon”, we parameterize what we define as an *information assumption*. For instance, if the target variable is Belgian GDP, we could define the information set available, say, one month before the official flash release. In this case, we have economic sentiment surveys available for the first three months of the quarter and industrial production figures only for the first month. When comparing the pseudo out-of-sample performance of different models, JD+ will simulate this information assumption and recursively project GDP growth over the evaluation sample period. In the case of univariate models for GDP, which do not exploit surveys or industrial production data, the details regarding our data availability assumption remain less relevant and the forecast degenerates to the traditional one-step-ahead projection. The parameterization of the information assumption can be explained in two steps, without the need to explicitly specify a calendar of data releases.

- 1) **Defining the “publication delay” ( $D$ ) relative to the *reference period* for all variables.** For each one of the variables that are part of the information set available to the forecaster, we first need to specify the approximate publication delay of the official data releases. Remarkably, for the Business and Consumer Surveys published by the National Bank of Belgium, we define a negative delay. That is, they are published

several days before the reference month has ended, e.g. the consumer survey for October was published on October 20<sup>th</sup> (i.e. delay=-10 days).

**2) Determining the “forecasting delay” ( $FD$ ) relative to the *reference period* for our variable of interest.** Once we have modeled the flow of data releases, it is not surprising that we also need to define the publication delay of our forecast itself, i.e. the forecasting delay ( $FD$ ). In principle we will aim to reproduce pseudo out-of-sample scenarios where the forecast for a given variable is computed several months before the official publication data, so that  $FD < D$  for our variables of interest. The *information assumption* that results from a given  $FD$  will of course depend on the publication delay specified for all series in the previous step. Such model for the publication delay is used by  $JD+$  to reconstruct all the encompassing information sets available when the nowcasting delay ranges from  $FD$  to zero. Thus, we will be able to assess the extent to which the forecasts gradually get closer to the actual values the nearer we are to the official publication date.

### 3. RESULTS

The simulation of out-of-sample forecasts will prove very useful in our empirical application. We will compare the forecasting accuracy of several competing methods applied to the early estimation of the Flash GDP. All those methods rely on forecasts for VAT returns and industrial production, which are not available for the three months of the quarter. Those forecasts will be computed with alternative ARIMA and dynamic factor models specifications, which are estimated within  $JD+$ . The pseudo out-of-sample forecast evaluation performed will shed some light on the forecasting reliability of the different methods beyond the in-sample statistics typically reported, thereby providing us with further insurance against the curse of in-sample over-fitting. Nevertheless, as pointed out by Diebold (2013), we should not interpret the forecasting evaluation results as definitive evidence for model selection, since the testing methodology used simply tests for the statistical significance of accuracy gains regardless of the models.

### 4. CONCLUSIONS

The implementation of alternative nowcasting and forecasting strategies in  $JD+$ , such as the use of dynamic factor models, or ARIMA models (TRAMO –SEATS methodology), contains a number of measures of fit which are also part of the loss function both at the model selection and estimation stages. In the process of model selection, it is customary to avoid over-parameterized models by minimizing a combination of the number of parameters and the one-step-ahead squared forecast errors. In this paper, we redefine the concept of forecast horizon in terms of the information assumptions that are compatible with the practice of real-time forecasting. Proceeding in this manner, our automatic simulation of out-of-sample projections will be very useful to construct measures of forecast accuracy that contribute to further insure the user against the curse of in-sample over-fitting. At the same time, the forecast errors comparisons implemented here are widely applied in the applied time series literature, since they measure the statistical significance of the forecasting gains associated to a given method.

## REFERENCES

- [1] Banbura, M. and M. Modugno (2012), "Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data", *Journal of Applied Econometrics*, **29**, 133-160.
- [2] De Antonio Liedo, D. (2014), "Nowcasting Belgium", National Bank of Belgium Working Paper Series, N. 256
- [3] Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, **13**, 253-263.
- [4] Diebold, F.X. (2013), "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests", *NBER Working Paper No. 18391*
- [5] Harvey, D., Leybourne, S. J. and P. Newbold (1998). "Tests for forecast encompassing". *Journal of Business & Economic Statistics*, vol. 16(2), pp. 254–259
- [6] Hyndman, R. J. and Koehler A. B. (2006). "Another look at measures of forecast accuracy." *International Journal of Forecasting* volume 22 issue 4, pages 679-688. doi:10.1016/j.ijforecast.2006.03.001
- [7] Charles P., D. De Antonio Liedo, J.Palate and M. Maggi (2014). "A Nowcasting Library in JDemetra+ for reading and visualizing news"
- [8] White, H. (2000), "A Reality Check for Data Snooping", *Econometrica*, **68**, 1097-1126
- [9] Grudkowska, S. (2015), "JDemetra+ User Manual"

# Simple forecasting techniques can reduce forward-series bias and keep revisions low for benchmarked Quarterly National Accounts estimates.

Geoffrey Brent ([geoffrey.brent@abs.gov.au](mailto:geoffrey.brent@abs.gov.au))<sup>1</sup> and Alex Stuckey ([alex.stuckey@abs.gov.au](mailto:alex.stuckey@abs.gov.au))<sup>1</sup>

**Keywords:** quarterly national accounts, temporal benchmarking, forecasting, forward series

Disclaimer: Views expressed in this paper are those of the authors and do not necessarily represent those of their employer. Where quoted or used, they should be attributed clearly to the authors.

## 1. INTRODUCTION

National statistical offices publish quarterly estimates that mimic the quarterly movements of indicator series whilst adhering to high quality level estimates with annual timing. There is inevitably a delay between the availability of the quarterly indicator and the availability of the relevant annual benchmark, requiring a forecasting/extrapolation approach to produce initial quarterly estimates. The published quarterly estimates are then revised when the annual benchmark becomes available. This study aims to identify a combination of benchmarking and forecasting methods that will produce quarterly estimates that preserve the movement of indicator variables, meet annual benchmarks, and minimise revisions to published quarterly estimates.

Several different methods are available for the problem of benchmarking a quarterly indicator series to produce Quarterly National Accounts estimates (“QNAs”) and revising these estimates as later data become available. Among the most prominent are Proportional Denton-Cholette, Proportional Cholette-Dagum Regression-Based, and Chow-Lin Regression-Based[1][2][3].

Because annual benchmarks are not available for the most recent quarters, each of these methods also requires some form of extrapolation to produce the forward series. Some methods do this through the same process that benchmarks data from previous years, e.g. fitting a regression model on benchmarked years. Alternately, extrapolation can be done as a separate step, using benchmarked estimates as input for a model that forecasts benchmark-indicator relationships for the forward series, e.g. [2] pp. 15-18.

Australia’s QNAs use Proportional Denton-Cholette benchmarking. The benchmark-indicator ratio (“BI ratio”) from the last benchmarked quarter is carried forward to all non-benchmarked quarters. Our colleagues previously found that other forecasting techniques may give better predictions for BI ratios[4]. Here we test whether this improved forecasting can translate into improved initial estimates, with reduced bias and revisions, when implemented in a realistic publication cycle using only data that would be available at the time of estimation.

---

<sup>1</sup> Australian Bureau of Statistics, Methodology and Data Management Division

## 2. METHODS

### 2.1. Benchmarking and forecasting methods

We tested five benchmarking methods available in the R `tempdisagg` package [5] (Chow-Lin-maxlog, Fernandez, Litterman, Denton-Cholette, original Denton). We also tested our own implementations of Di Fonzo and Marini’s enhancement to proportional Denton[6] (“Denton-Enhanced”) and of proportional Cholette-Dagum[3] with autoregressive parameters set at 0.84 and 0.93 (“DC0.84” and “DC0.93” below).

We combined these with several forecasting methods, including random-walk (RW) and random-walk with drift (RWD), `auto.arima`, and ETS, all available from the forecast package in R[7]. The forecasting methods were applied to the quarterly BI ratios or, in the case of the enhanced Denton method, the annual BI ratios. Forecast method “none” indicates the default forward-estimate calculation for the benchmarking method in question; for Denton-Cholette and Denton-Enhanced this is equivalent to random-walk forecasting. Some benchmark/forecast combinations were excluded because of implementation issues.

In this study we simulate the timing of benchmarking in the Australian quarterly national accounts. Benchmarking is done over a five-year window of the data and extrapolation is required for up to eight quarters ahead of the most recent annual benchmark[8].

### 2.2. Data and Metrics

We tested methods on a range of Australian QNA series: 13 seasonally-adjusted Industry series and 44 Public Capital series containing seasonal effects. Several of these series show evidence of long-term structure in the BI ratio, with either gradual or abrupt changes. We also used simulated data to explore how various characteristics of data might affect model performance, e.g. whether an abrupt change in BI ratios can cause large revisions to the back series.

We assessed methods based on three main metrics:

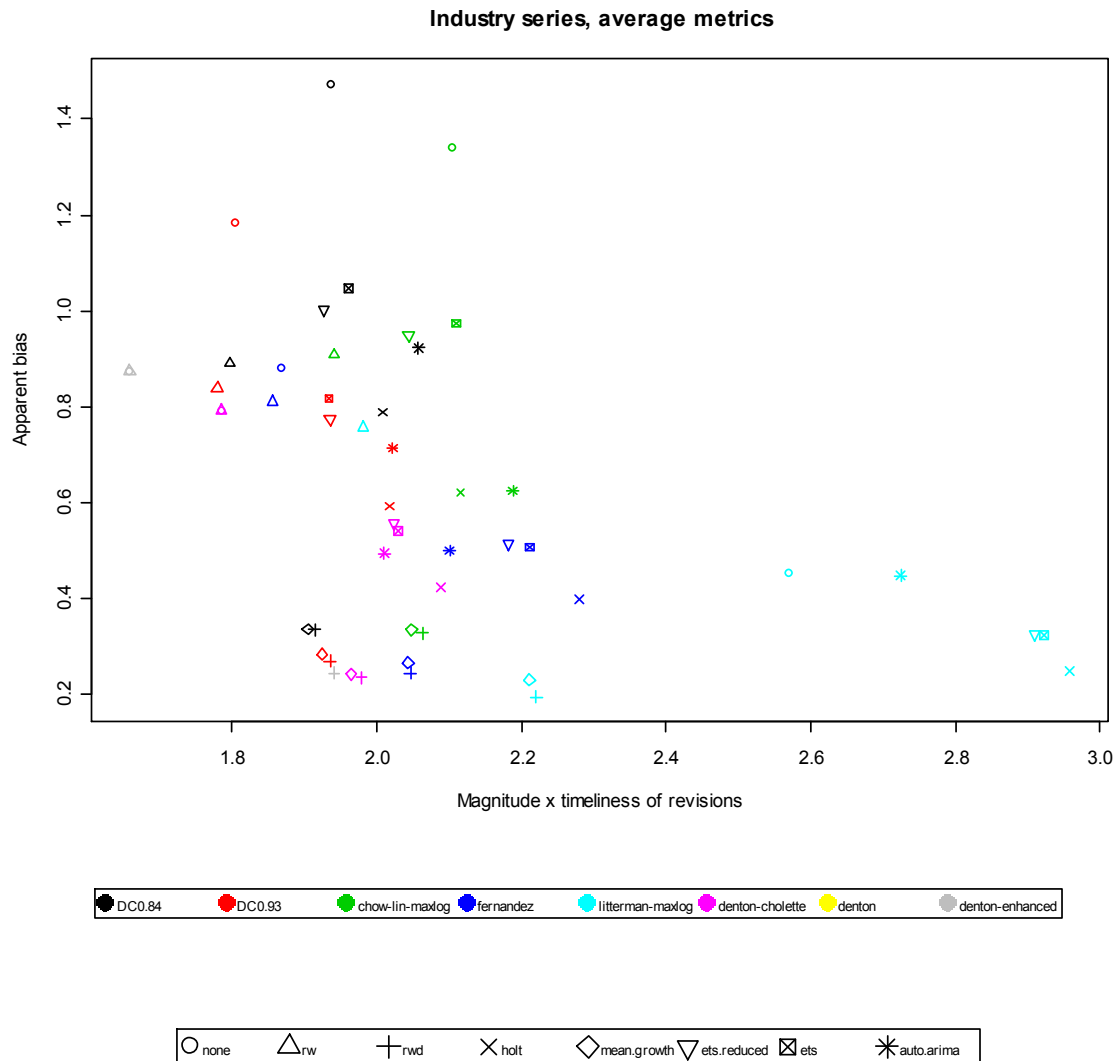
- Difference between movements in the final benchmarked estimates, and movements in the indicator
- Magnitude x timeliness of revision effects on movements (i.e. how quickly do estimates for each quarter approach the final estimate?)
- Absolute estimated bias of initial vs final estimates (i.e. for a given series, does the forecasting method consistently over/under-predict the final estimate?)

We scaled metrics relative to series volatility, using a method based on Mean Absolute Scaled Error, in order to improve comparisons and interpretation between different series. Methods are assessed based on averaged performance over multiple series, but we also plotted results for individual series results to identify outliers etc.

## 3. RESULTS

Random-walk BI forecasting consistently gave the best performance for revisions, in both seasonal and non-seasonal data, especially when combined with Denton-enhanced benchmarking. RWD and extrapolating the mean growth showed slightly worse results for revisions, but gave much smaller bias. Figure 1 shows average bias and revisions metrics for each method across the Industry series; the best compromise options appear to be Denton-Enhanced, Cholette-Dagum, or Denton-Cholette, combined with either

RWD or mean growth forecasting (mean growth not tested for Denton-Enhanced). Original Denton consistently gave large revisions and has been excluded from further consideration. These findings also hold for Public Capital (not shown).



**Figure 1. Bias and revisions for Industry series (original Denton not shown)**

All methods showed similar performance in following indicator movement for non-seasonal Industry data. On Public Capital data, Denton-Enhanced, Denton-Cholette, and Cholette-Dagum all performed equally well at following indicator movement: the mean discrepancy between indicator and estimate movements was approximately 0.9% of the quarter-to-quarter variation in movements (some of which is seasonal). Chow-Lin, Fernandez, and Litterman did worse, largely due to poor performance on two badly-behaved data sets.

Results on synthetic data sets suggested that regression-based methods work well for data where there is no long-term structure in the BI ratio, but tend to do poorly when such structure exists, unless their residuals model is capable of reflecting this structure.

**Table 1. Comparison of metrics for selected methods**



method		Industry		Public Capital	
benchmarking	BI forecasting	revisions	bias	revisions	bias
Cholette Dagum ( $\phi = 0.93$ )	none	1.8037	1.1828	0.0776	0.0890
	RW	1.7796	0.8384	0.0766	0.0608
	RW with drift	1.9359	0.2700	0.0863	<b>0.0069</b>
Litterman	RW	1.9811	0.7570	0.1365	0.0693
	RW with drift	2.2206	<b>0.1961</b>	0.1436	0.0205
Denton-Cholette	RW	1.7854	0.7923	0.0769	0.0576
	RW with drift	1.9792	0.2366	0.0882	0.0072
Denton - enhanced (Di Fonzo and Marini)	RW	<b>1.6588</b>	0.8742	<b>0.0701</b>	0.0645
	RW with drift	1.9410	0.2456	0.0821	0.0167

#### 4. CONCLUSIONS

Simple forecasting methods can be effective in reducing bias for initial estimates and magnitude/delay of revisions. For regression-based benchmarking methods, a random-walk BI forecasting approach produced the smallest/earliest revisions. (For Denton methods, this is already the default forecasting method.)

Random-walk-with-drift and mean-growth forecasting methods produce slightly higher revisions but much lower bias, and may be a good compromise. These methods are best combined with Denton-Cholette, Denton-Enhanced (Di Fonzo-Marini), or Cholette-Dagum benchmarking. We provide a matrix-form expression for Denton-Cholette RWD estimation, allowing for easy implementation in a production environment.

Further work is discussed including an investigation of the ideal data span over which benchmarking is done. Also, the possible need for prior correction of indicator series to account for changes in source data is discussed with some approaches suggested.

#### References

- [1] Eurostat, “Handbook on quarterly national accounts” (2013), chapter 5 pp. 121-171
- [2] International Monetary Fund, “Update of the Quarterly National Accounts Manual” (January 2014 draft), chapter 6
- [3] E. Dagum and P. Cholette, “Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series” pp. 80-81
- [4] Y. Joymungul Poorun, R. Mathews, J. Chien, and P. Gould, unpublished work.
- [5] C. Sax and P. Steiner. “tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series”, (2013) R package version 0.22. <http://CRAN.R-project.org/package=tempdisagg>.
- [6] T. Di Fonzo and M. Marini, “On the Extrapolation with the Denton Proportional Benchmarking Method”, IMF Working Paper, June 2012
- [7] R. J. Hyndman “forecast: Forecasting functions for time series” (2011) R package version 3.04. <http://CRAN.R-project.org/package=forecast>

- [8] Australian Bureau of Statistics “5216.0 Australian National Accounts: Concepts, Sources and Methods” (2000), chapter 13 p. 167.

# Does web anticipate stocks? Analysis for a subset of systemically important banks

Michela Nardo<sup>1</sup> ([michela.nardo@jrc.ec.europa.eu](mailto:michela.nardo@jrc.ec.europa.eu)) and Erik van der Goot<sup>2</sup>

**Keywords:** stock market trends, sentiment analysis, systemically important banks, web mining, big data.

## 1. INTRODUCTION

Are financial prices and transactions predictable? Not according to the Efficient Market Hypothesis (EMH, [1]) that models stock markets as random walks, where shocks are temporary and largely driven by new and unexpected information. The critical reading of EMH, offered by Behavioral Finance [2], suggests instead a certain degree of predictability. Investors could be subject to waves of optimism and pessimism, causing prices to deviate systematically from their fundamental values ([3]) or may be systematically overconfident in the ability to forecast future stock in prices or earnings [4]. Recently the geometric increase in digital information (on line journals, dedicated blogs, social networks, etc.) made possible to address the predictability of financial markets from a different perspective, that of Big Data. Terabits of data on financial transactions can be matched to a comparable amount of on line news to dig into the mechanism of decision making. The literature relating web mining with financial prediction is relatively recent. To the best of our knowledge the first study is due to Wysocki [5]. He proved that, between January and August 1998, the Yahoo! posting volume related to 50 companies did forecast next day trading volumes. About the opposite result, namely internet buzz cannot predict trading volume, is obtained by Tumarkin and Whitelaw [6], among others. More recently Preis et al. [7] and Moat et al., [8] show that weekly transactions volumes of S&P500 companies are indeed preceded by web searches of financial words or Wikipedia views.

The aim of this paper is to analyse whether information coming from the web has some predictive power on the stock market behaviour of a set of systemically important European banks. The research questions are: *(i) is web buzz able to lead stock behavior?* *(ii) Are stock movements sensitive to the source location of web buzz?* Using the Europe Media Monitor (EMM) engine we scrape the World Wide Web between December 2013 and April 2014, and retrieve the public information related to a set of 10 banks: Barclays, BBVA, BNP Paribas, Cr dit Agricole, Deutsche Bank, HSBC, Royal Bank of Scotland, Santander, Soci t  G n rale and Unicredit). Our working hypothesis is that the amount and the mood (sentiment) associated to the web buzz can be a proxy of the interest a given financial institution is attracting. This interest, in turn, should be linked to the stock price of the institution.

---

<sup>1</sup> Unit for Financial and Economic Analysis, European Commission, Joint Research Centre, <https://ec.europa.eu/jrc/>

<sup>2</sup> Unit for Global Security and Crisis Management ([erik.van-der-goot@jrc.ec.europa.eu](mailto:erik.van-der-goot@jrc.ec.europa.eu)), European Commission, Joint Research Centre, <https://ec.europa.eu/jrc/>

## 2. METHODS

Data: Between Dec. 5<sup>th</sup> and April 30<sup>th</sup> 2014 we collect daily news coming from electronic media sources (obtained via Europe Media Monitor) for 10 systemically important banks. Weekends (and non-contracting days) are excluded. For the same period daily data on stock prices (open, close, highest, lowest) and volumes exchanged are gathered from New York Stock exchange and from various European Stock exchanges (Frankfurt, London, Madrid, Milan, Paris). The relationship between stock data and web news is analysed via cross-correlation function, Granger causality, rank-sum test, Factor and Cluster analysis for each combination of 8 stock prices variables, 12 web buzz variables, 4 set of sources (with different geo-tagging), various stock markets. The Europe Media Monitor (EMM) provides near real-time (update frequency measured in minutes) monitoring/scraping of over 4000 electronic media websites in 60 languages. EMM analyses the retrieved web texts in the form of entity recognition, entity extraction, recognitions of quotes, sentiment/tonality analysis. The tonality/sentiment of an article is determined using 4 sets of ‘tonality’ words per language, denoting highly positive, positive, negative and highly negative words. These tonality dictionaries are currently available in 14 languages, including the main EU languages.

## 3. RESULTS & CONCLUSIONS

In line with the literature, we find an average cross-correlation between stock variables and web buzz, in the range 0.33-0.37 at lag  $\delta=0$  (contemporaneous correlation), significant at 1%. Gloor et al. [9] find a positive correlation at  $\delta=0$  (highest equals 0.45 significant at 5%) between a set of web variables constructed via semantic social network analysis and the prices of 21 stocks. Significant cross correlation (around 0.3) is found between search data and volume traded for some specific terms and only for instantaneous correlation by Preis et al. [10]. Bordino et al. [11] find on average 0.31 at  $\delta=0$  with peaks of 0.83. Checking for individual banks we find correlations up to 0.73 for Barclays and between 0.6 and 0.68 for Unicredit and the Royal Bank of Scotland. Web buzz seems to have a poor association with New York stock data for all banks analysed no matter which set of web sources is considered, cross correlation is systematically lower when New York stock data are used. We further explore the issue regressing stock prices (volumes) from the NY stock market onto its past values and on present and past values of web buzz. The web variables almost always result to be non-significant. A further look to the data confirms that New York stock values reacts much more to the corresponding movements in European stocks (NYSE opens 5/6 hours later) than to web buzz. The correlation between opening prices ranges from 0.91 to 0.98 for all banks considered. A Granger causality test on opening prices confirms that association goes one-way from European to NY stock exchanges

*Are stock movements sensitive to the source location of web buzz?* To answer this question we estimate the equation:  $S_t = \alpha + \beta_1 S_{t-1} + \beta_2 W_t + \beta_3 W_{t-1} + \varepsilon_t$  for each bank in the sample and each of the four different information sets for the web buzz ( $W$  denotes web variables and  $S$  stock variables). Web variables are calculated from web texts coming from: 1) a source located in the USA and in the European Union (EU+US); 2) a source located in the European Union (EU); 3) sources all over the world (ALL); 4) sources located in the country where the bank has its main seat (Country). For each estimated model we calculate the percentage change in the model fit ( $R^2$ ) using option 4 as baseline. Our analysis shows that the location of the source matters. Web buzz derived from EU+US sources or from world sources improves the predictive power of the regression up to 27.5%, if compared with the same regression but with web buzz obtained from Country sources. Overall European stock markets seem to respond to news

reported at the international level, rather than locally. The importance of the news is probably the explanation. Main news, those more likely to drive stock prices, are those finally reported by the international (financial) journals. As EMM is unable to distinguish between “important” and “unimportant” news (as soon as the required keywords are in it), the use of sources with international echo eliminates some of the noise introduced by irrelevant texts at the country level.

*Is web buzz able to lead stock behaviour in our dataset?* Not on average, according to our data. Granger test fails to support an average association that goes one-way from web to stocks. We nevertheless find a statistically sound anticipation capacity for single banks, particularly Unicredit, Deutsche Bank and Crédit Agricole but also in some cases for BBVA, Royal Bank of Scotland, Société Générale with gains in RSS ranging from 4 to 12%. The explanation offered by the literature for this poor average performance is that new information is rapidly incorporated into agents’ information set so excessive returns rapidly vanish: only very short (ideally intraday) stock price movements can be capitalized ([12], [13]). In our analysis U-rank test confirm the association between web buzz and intraday price movements making this topic a potential candidate for future research.

A simple hierarchical clustering on the price variables Close(t)-Open(t) shows that euro-area banks tend to cluster together very fast while English banks are far apart and move differently. However, this is not the case for web buzz variables, where the differentiation between continental and UK banks is not clearly defined. Principal Component Analysis (PCA) confirms the findings of Cluster analysis. PCA on stock prices and volatility shows that, while euro area banks are all robustly loaded (with the same sign) by the same single factor, UK banks tend to be loaded by multiple factors (especially HSCB which stands out as the most *diverse* bank). Euro area banks show a unique common driver explaining 74.06% of the total euro area variance, all the remaining variance practically represents idiosyncratic bank-related noise. If web buzz were to reflect/anticipate stock movements we should expect a grouping in the PCA similar to than found for the stock variables. This is not the case: the PCA on the web variable *Number of articles* reveals at least 5 different (orthogonal) relevant factors, the first of which explaining only 15.55% of the total variance (the first PCA factor on the stock variable represented about 60% of the total variance).

Results therefore seem to indicate that while supra-national decisions/facts at the Euro area level are in fact driving stock behaviours, web news about single banks is only episodically making a difference in stock movements. Most likely in these times of financial turbulence announcements of the BCE or of other international authorities are likely to play an decisive role in explaining trade behaviours. Results do not say that web buzz could not be relevant in explaining stock behaviour but rather than web buzz about individual banks cannot. Yet, we believe that the web hosts valuable information thus in future works we will investigate general economic/financial trends based on web information. Our analysis does not suggest a clear advantage of measures of web buzz based on tonality with respect to other count variables (e.g. relative number of messages). This could be partly due to the algorithm calculating tonality. During the test phase we realized that the tonality failed to identify some important financial news (like for example the downgrade of Deutsche Bank on the 19<sup>th</sup> of Dec.). Currently the tonality algorithm is being upgraded to provide entity based sentiment. Even so tonality and sentiment analysis on financial texts are the latest and most promising advances in this type of literature (see [14], [15], [16]). Finally another limitation of our analysis is surely the restricted set of bank analysed. Enlarging the group of banks would lead us to face the trade-off between wide coverage but lower number of daily web texts extraction (e.g.

we obtain very few web texts and not every day for the Portuguese Banco Espirito Santo, the Finnish Pohjola and the Belgian KBC). Aggregation at the weekly level could be a solution worth exploring.

## REFERENCES

- [1] Fama E., (1965), The Behavior of Stock Market Prices, *The Journal of Business*, vol. 38(1): 34-105.
- [2] Della Vigna S., (2009) Psychology and Economics: Evidence from the Field, *Journal of Economic Literature*, vol. 47(2):315–372.
- [3] DeBond W.F.M., Thaler R., (1985), Does the Stock Market Overreact? *Journal of Finance*, vol. 40: 793-805.
- [4] Kahneman D., Tversky A., (1979), Prospect Theory: an Analysis of Decision Under Risk, *Econometrica*, vol. 47(2):263-291.
- [5] Wysocki P.D., (1998), Cheap Talk on the Web: The Determinants of Posting on Stock Message Boards, *Working Paper n. 98025*. University of Michigan Business School.
- [6] Tumarkin R., Whitelaw R.F., (2001), News or Noise? Internet Message Board Activity and Stock Prices, *Financial Analysts Journal* vol. 57:41-51.
- [7] Preis T., Moat H.S., Stanley H.E., Bishop S.R., (2012), Quantifying the Advantage of Looking Forward, *Scientific reports*, 2, Article number: 350.
- [8] Moat H.S., Curme C., Avakian A., Kenett D.Y., Stanley H.E., Preis T., (2013), Quantifying Wikipedia Usage Patterns Before Stock Market Moves, *Scientific Reports*, 3, n. 1801.
- [9] Gloor P., Krauss J., Nann S., Fischbach K., Schroder D., (2009), Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. 2009 International Conference on Computational Science and Engineering, Aug. 29-31, Vancouver, Canada.
- [10] Preis T., Reith D., Stanley H.E., (2010), Complex Dynamics of our Economic Life on Different Scales: Insights from Search Engine Query Data, *Philosophical Transactions of the Royal Society A*, vol.368:5707-5719.
- [11] Bordino I., Battiston S., Caldarelli G., Cristelli M., Ukkonen A., Weber I., (2012), Web Search Queries Can Predict Stock Market Volumes? *PLoS One* vol. 7(7): e40014.
- [12] Schumaker R., Chen H., (2006), Textual Analysis of Stock Market Prediction Using Breaking Financial News: the AZFinText System, *12th Americas Conference on Information Systems* (AMCIS-2006), Acapulco, Mexico.
- [13] Schumaker R., Chen H., (2009), A Quantitative Stock Prediction System Based on Financial News, *Information Processing and Management* vol. 45(5):571-583.
- [14] Zhang X., Fuehres H., Gloor P., (2010), Predicting Stock Market Indicators through Twitter ‘ I hope it is not as bad as I fear’. *Procedia – Social and Behavioral Science*, 2010.
- [15] Zhai J., Cohen N., Atreya A., (2011), Sentiment Analysis of News Articles for Financial Signal Prediction, *mimeo*, University of Stanford, USA.
- [16] Tulankar S., Athale R., Bhujbal S., (2013), Sentiment Analysis of Equities using Data Techniques and Visualizing the Trends, *International Journal of Computer Science Issues*, vol. 10(4): 265-269.

# Building a Cross Border Data Access System for Improved Scholarship and Policy: The Case of the German IAB Network of RDCs

Joerg Heining ([joerg.heining@iab.de](mailto:joerg.heining@iab.de))<sup>1</sup>, Warren A. Brown<sup>2</sup>, William C. Block<sup>2</sup> and Stefan Bender<sup>1</sup>

**Keywords:** Confidential Micro Data, Remote Data Access, International Data Sharing

## 1. INTRODUCTION

The Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) provides data on individuals, households and establishments, as well as data that comprise both establishment and personal information. The FDZ data originate from three different sources. From the notification process of the social security system and the internal procedures of the Federal Employment Agency process-generated data are obtained. Furthermore, the IAB acquires data by conducting own surveys. The use of weakly anonymous data is subject to restrictions concerning data protection legislation. Due to these regulations the data can be analyzed only via on-site use. For this purpose, the FDZ provides separate workplaces for guest researchers in Nuremberg and further locations in Germany and in the USA. The benefits and the means for documenting these benefits is a central focus of the paper.

## 2. METHODS

The Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg (see Heining 2010), Germany provides researchers with access to confidential micro labour market data. The available data at FDZ on individuals, households and establishments come from several sources. Administrative data are obtained from the notification process of the social security system and the internal procedures of the Federal Employment Agency. The IAB also conducts its own surveys of households and establishments. The administrative and survey data can be analyzed separately, but all IAB surveys can also be linked to the respondent administrative records, resulting in, for example, linked-employer-and-employee data.

At the moment, three data access modes are available at FDZ:

- **Scientific Use Files:** Scientific Use Files are specially prepared data sets for off-site access. After the conclusion of a use agreement with FDZ, researchers can download a copy of the data to their local computer.
- **Remote Execution:** Data with higher research potential so called weakly anonymized data may be accessed off-site via remote execution. Researchers submit codes to FDZ which are processed with the data. The results are returned via email to the researcher after disclosure review.

---

<sup>1</sup> Institute for Employment Research (IAB), Regensburger Strasse 104, 90487 Nuremberg, Germany

<sup>2</sup> Cornell Institute for Social and Economic Research (CISER), 391 Pine Tree Road, Ithaca, NY 14850, USA

- On-site Access: Weakly anonymized data may also be accessed on-site at the FDZ in Nuremberg.

Especially to users of weakly anonymized data, on-site access is very important. This is the only access mode which allows researcher to actually see the data. Although the FDZ provides documentation of the data and test data in order to prepare codes for remote execution, the disadvantage of not seeing the data cannot be compensated. Costly trips to access the data on-site at the FDZ in Nuremberg have been the only possibility for researchers if they wanted to actually “see” weakly anonymized data.

In 2010, FDZ introduced the Research Data Center-in-Research Data Center approach (RDC-in-RDC, see Bender and Heining 2011) which intends to overcome these problems and to bring data access in Germany closer to the idea of remote data access. For the first time, researchers may access weakly anonymized FDZ data on-site at locations other than Nuremberg.

The RDC-in-RDC approach is technically implemented by using a so called Citrix-thin client technology (see Heining and Bender 2012). A thin client computer serves as an interface for the user in order to establish a secure communication link to a FDZ server in Germany where all data processing is done. The RDC-in-RDC approach was initially implemented at four sites in Germany and at one site in the US. The RDCs of the Statistical Offices of the Länder (i.e., states) in Berlin, Bremen, Düsseldorf and Dresden, as well as at the Institute for Social Research (ISR), University of Michigan in Ann Arbor, MI, USA. Over the years, additional access points in Germany (University Applied Labour Studies of the Federal Employment Agency in Mannheim) the US (Cornell University at Ithaca, NY, University of California at Berkeley, and Harvard University) have been opened. Future sites will be at Princeton University, Princeton, NJ, USA and at the Statistical Office of Niedersachsen, Hannover, Germany. Moreover, in the context of the Data without Boundaries project (DwB), both the University of Essex in Coulchester, UK and at the L'Institut national de la statistique et des études économiques (INSEE) will host IAB access points.

### **3. RESULTS**

As this network of research data centres/access points is built out it is with the expectation that improved access to restricted data managed with a high degree of data security will yield benefits to policy makers. So far, these expectations have been more than fulfilled.

Researchers in Germany and abroad value the possibility of decentralized access to a resource which allows for cutting edge research in Economics, Statistics and the Social Sciences. Because of the RDC-in-RDC approach, researchers can easily access IAB's detailed longitudinal data (administrative data as of 1975) to study the mechanisms governing labour markets and the incentive effects of labour market interventions. This is impressively proven by the number of new projects and the utilization of IAB's access points as well as by the numerous research papers using IAB data and published in peer reviewed academic journals. Especially users from abroad show a high interest in studying the German economy since it is the third largest economy in the world with a workforce of over 47 million people and an elaborate system of unemployment insurance and benefit schemes. The derived insights from studying Germany may thus be of broader impact and may be directly applicable to other countries.



Because of the facilitated access to high quality administrative data, the RDC-in-RDC supports training and education at the host institution, too. Since average approval time for student projects is less than two months, students may use IAB data for class assignments as well as longer term projects such as master's theses and doctoral dissertations. Furthermore, group approvals are possible and students may use restricted data for class (projects still take place at IAB on-site access points, but can be built into the regular classroom curriculum). By using IAB data, students and scholars learn how to work with administrative data and how to cope with problems and shortcomings typical to data of this type. They acquire skills and experiences which are transferable when they want to work with comparable data from different data producers/populations centers (for example, data from the U.S. Census Bureau, Longitudinal Employer-Household Dynamics (LEHD) data, etc.).

Another dimension of benefits from the RDC-in-RDC approach is that it helps to realize positive externalities between data producers. In general, staff of the hosts acts as supervisor of IAB's external access points. Because of this, they learn about access rules, legal institutional backgrounds, available data, and procedures at IAB. This eventually results in a dialogue between different national and international data producers and a knowledge transfer between international experts on data access and dissemination.

A direct result of these externalities is the joint project between IAB and the Cornell Institute for Social and Economic Research (CISER) on hosting FDZ's Scientific Use Files. As mentioned before, FDZ prepares so-called Scientific Use Files which are specifically designed for off-site use/access and are characterized by a higher degree of anonymity compared to weakly-anonymized data. In the past, users could download these Scientific Use Files from a secure exchange server after concluding a use agreement with FDZ. However, this procedure bared the problem that Scientific Use Files were kept illegally by users after the expiration of the use agreement. In order to solve this problem, Scientific Use Files will be stored at the Cornell Restricted Access Data Center (CRADC), a wholly-contained secure computing environment at CISER. Users can connect to CRADC after the concluding of a use agreement with IAB. Once this agreement has expired, their CRADC account is disabled. As a consequence, no local copies of Scientific Use Files may be kept illegally.

#### **4. CONCLUSIONS**

IAB's network of RDCs has been a true story of success so far. It provides a cheap and scalable solution for international data sharing and shows benefits to both data producers, local hosts and scholars. Positive externalities may be realized by all parties involved. However, this story of success has not come to an end yet. IAB will continue to expand its network of access points/RDCs in the future. The idea of a Reciprocal RDC-in-RDC, i.e. IAB will host thin clients/data of other data producers, eventually leading to a true transnational data access network will be a challenge for the future.

#### **REFERENCES**

- [1] Jörg Heining, The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009, *Zeitschrift für ArbeitsmarktForschung* 42,4, (2010), 337-350
- [2] Stefan Bender and Jörg Heining, The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing, *IASSIST Quarterly* 35, 3 (2011), 10-16

- [3] Jörg Heining and Stefan Bender, Technical and organisational measures for remote access to the micro data of the Research Data Centre of the Federal Employment Agency, FDZ-Methodenreport 08/2012 (2012), 14 pp.

# Microaggregation for the masses: non-confidential enterprise-level data for analytical and research purposes

Pierre Lamarche (pierre.lamarche@ecb.europa.eu)<sup>1</sup>, Sébastien Pérez-Duarte (sebastien.perez-duarte@ecb.europa.eu)<sup>1</sup>

**Keywords:** enterprise data, micro aggregation, distributed micro-data, statistical disclosure control, clustering.

## 1. MOTIVATION

There has recently been a renewed interest in the use of partially aggregated data in cross-national analysis of enterprise data. Partially aggregated data – which we term “meso-aggregated” to indicate a finer aggregation than macro statistics – has found a niche for enterprise data in international comparisons.

The oldest yet still active example is probably the BACH dataset of the European Committee of Central Balance Sheet Offices<sup>2</sup>, initially sponsored by the European Commission (DG ECFIN), which consists of financial information of European corporations by country, sector and size class. More recent initiatives, pioneered by Bartelsman and co-authors [1] (who term it “distributed micro-data analysis”), have pushed for the construction of indicators by country, sector, and time. Even more recently, the Competitiveness Research Network (see [4]) of European central banks and statistical institutes has pooled resources from a large number of countries in providing even more detailed but still aggregated information from firm-level data. Although less fine, one could also name the DynEmp and the EUKLEMS initiatives.

Some of these analyses seem to consider that aggregating enterprises by e.g. sector allows treating these sectors as if made of homogeneous agents and thus drawing conclusions from the study of ratios and/or regressions, treating the aggregate “cell” as statistical unit. This is of course well known to be an over-simplification.

One distinguishing feature of the BACH, the CompNet and the DynEmp datasets is that in each cell, in addition to the mean or total of the variables of interest, information is provided on the distribution within this cell – in the case of BACH and DynEmp, the quartiles; in the case of CompNet, deciles, top and bottom percentiles, standard deviation, and some higher moments. It would thus seem that, with this approach, the users of the dataset do not need to assume the homogeneity of the cells, and can take the heterogeneity into account.

We first argue in this paper that, although meso-aggregated data allow working with otherwise confidential information, their use is mostly limited to the initial purpose with which they were designed: if a correlation is of interest that was not foreseen in the meso-aggregation procedure, it is not always possible and rarely practical to recover it from the data. This severely limits the usefulness of meso-aggregated data and is therefore not the go-to solution one could hope for.

---

<sup>1</sup> European Central Bank, 60640 Frankfurt, Germany

<sup>2</sup> <https://www.bach.banque-france.fr/>

The ongoing work of Eurostat and the European Statistical System in establishing cross-border remote access facilities is certainly the preferred solution for researchers across Europe. However, this solution is probably some years in the making and an alternative plan, which would allow more users to access and work on some form of aggregate data is required, or at least desired.

Therefore, in this paper we consider part of the existing literature on micro-aggregation, and argue for it as a superior alternative to meso-aggregated data.

## **2. MICROAGGREGATION AS A POTENTIAL IMPROVEMENT ON MESO-AGGREGATION**

Microaggregation is the process of joining several statistical units into a composite one, and reporting only the total (or equivalently the averages; for more details see for instance [2], [3] and [5]). The units to be grouped are selected with the idea that they are sufficiently similar and that the composite unit behaves not dissimilarly to each of the constituents. Each group or cluster should contain a minimum number of units to comply with statistical confidentiality rules. One of the main difficulties in microaggregation is the clustering process of how to select similar units (i.e. reducing intracluster variance as much as possible) while ensuring a sufficient but not too high number of units in each cluster. Clustering in this manner is an NP-complete problem, and hence is computationally very expensive with the large samples of the kind that are commonly encountered for business data.

We review the existing literature and the different clustering techniques, and propose an application of them that allows backwards compatibility with the current BACH and CompNet endeavours. We then implement different variations of this microaggregation approach on Bureau van Dijk's Amadeus dataset of corporate balance sheet and income statement information, by considering different variables in the clustering and minimum cluster size.

We then compare the results of analysing the original microdata and the microaggregated data, as well as with data meso-aggregated in the BACH manner by sector, size class and country, showing the advantages as well as the limits of microaggregated data, and their superiority over meso-aggregated data.

## **3. PRELIMINARY RESULTS**

In this paper we will concentrate on a basic but well-tested algorithm of micro-aggregation, namely the MDAV-generic (Maximum Distance to Average Vector), described for instance by [3]. It is not expected that the results are sensitive to the exact aggregation method but this is left for future work.

In a first step we analyse the effect of microaggregation carried out with only a few variables (cost of labour, fixed assets, turnover, profit and loss, and capital) on the distribution of these and the other variables of the balance sheet and income statements. These results on Amadeus data show that variables that are excluded from the computation of the distance may be poorly aggregated if the microaggregation variables have little explanatory power for these variables. As a consequence their distribution may be badly reproduced in the microaggregated data: for instance, whilst distortion for quantiles is generally no more than 10% for variables that are included in the computation of the Euclidean distance (although with some notable exceptions in the tails of the distribution), errors for the other variables may reach 50% and are higher along the total distribution. Likewise, the correlation between the different variables may

be disturbed. However the few regressions performed on microaggregated data do not show significantly higher bias for covariates that were not included in the computation of distance (in line with the existing literature, see e.g. [5] and [7]). There is a trade-off between the number of variables that are used to compute the Euclidean distance and the errors due to the microaggregation for the two sets of variables. Indeed increasing the number of variables in the computation of the distance implies higher bias with respect to the full distribution. In this respect, it would be important to choose the variables used in the calculation of the distance based on the analysis that will be conducted on the microaggregated data. Unfortunately, for general purpose microaggregation, it is not possible to know in advance all the analyses that will or may be carried out and one must therefore settle for a generic selection of the microaggregating variables.

In a second step, we look into stratification of the sample before microaggregation, as a technique to decrease the computational complexity of the procedure and possibly improve the results. If the strata are well chosen, this technique does indeed not decrease the utility of the data. For microaggregation to be compatible with already existing meso-aggregated data such as BACH or CompNet, stratification by industry and/or size is needed. We test the hypothesis that stratification by size is not needed since microaggregation was expected to capture size effects and find that, contrary to our expectations, defining sectors (at the NACE 2-digit division level) or even macro-sectors (at the NACE 1-letter section level) rather than size classes (here defined by the number of employees) as strata during the microaggregation process often has only an undetermined effect on the distribution (quantiles). We follow [6] and look for the stratification that implies the least loss of variance. We find that stratification by size preserve variance very slightly more efficiently than by sector. However the best solution remains a stratification combining both size and macro-sectors, which reduces the loss of variance by 1% on average.

In a third step we consider a few possible transformations that can be applied to the variables in the microaggregation in order to improve the efficiency of the procedure. For instance, applying a logarithmic transformation to the variables is equivalent to performing their geometric mean. Moreover financial variables are often highly skewed and in economic research the regressions are most of the time performed on the logarithms of the nominal variables. This transformation is thus appealing and natural. First results indicate that the logarithmic transformation preserves better the distribution but tends to lower the variance. We also investigate the question of financial ratios whose analysis is often performed by analysts on business data. For instance, users may be interested in the return on equity, or in the unit labour cost. Reproducing this ratio in microaggregated data is done in a more efficient way, mostly for the tails of the distribution, by including the ratio directly in the computation of the distance.

Finally, we perform some simulations in order to confirm our results. We estimate the parameters of a multi-variate lognormal distribution on the Amadeus data by sector, and generate variables following this law. The simulations confirm our findings that there is no higher bias for variables that are excluded from the computation of the distance.

#### **4. CONCLUSIONS**

Micro-aggregation is a promising alternative to the other currently used approaches to construct and disseminate cross-country corporate sector data. However it requires careful analysis of both the characteristics of the data and the possible uses, and is computationally intensive on large datasets. There is still much work to make

microaggregation easily accessible and implementable, and an interesting application might be a European microaggregated dataset of enterprise balance sheet information.

Possible extensions to our work would take into account the panel dimension of the firm data, which is completely missing in the current meso-aggregated datasets. Naïve extension of microaggregation to the case of longitudinal data would quickly be affected by the curse of dimensionality, and in line with our results, this would limit significantly the usefulness of the microaggregated data. Enterprise demography would also need to be considered. Finally, the results in [7] could be extended to the more general case of multivariate microaggregation.

## REFERENCES

- [1] Bartelsman, E., J. Haltiwanger, and S. Scarpetta, Cross-Country Differences in Productivity: The Role of Allocation and Selection, *American Economic Review* 2013, 103(1): 305–334
- [2] Defays, D., Protecting Micro-Data By Micro-Aggregation: The Experience In Eurostat, *Qüestió*, Vol. 21, 1 I 2, p. 221-231, 1997
- [3] Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2), 195-212.
- [4] European Central Bank, Competitiveness Research Network: First Year Results, June 2013.
- [5] Feige, E. L., & Watts, H. W. (1972). An investigation of the consequences of partial aggregation of micro-economic data. *Econometrica*, 343-360.
- [6] Mateo-Sanz, J.M., and J. Domingo-Ferrer, A Comparative Study Of Microaggregation Methods, *Qüestió*, vol. 22, 3, p. 511-526, 1998
- [7] Schmid, M., Schneeweiss, H., & Küchenhoff, H. (2007). Estimation of a linear regression under microaggregation with the response variable as a sorting variable. *Statistica Neerlandica*, 61(4), 407-431.

# Occupational mismatch impact on earnings

Monica- Mihaela Maer Matei (matei.monica mihaela@gmail.com)<sup>1</sup>

**Keywords:** over skilling, occupational mismatch, propensity score, earnings.

## 1. INTRODUCTION

This study investigates the impact of over-skilling within UK labour market. The research uses micro data available in the Programme for the International Assessment of Adult Competencies (PIAAC) database. This is a household study collecting information about educational background, work experience and skills of adults around the world and represents one of the most useful initiatives for understanding the integration of higher education graduates into the labour market. It provides the necessary data for skill mismatch estimation because it involves the direct assessment of literacy, numeracy and problem solving in technology-rich environments adult's competencies.

## 2. METHODS

The analysis is developed for higher education graduates who are in full employment. Data collected by the Survey of Adult Skills (PIAAC) is used to estimate the size of skill mismatch. The data collection for the Survey of Adult Skills (PIAAC) took place from 1 August 2011 to 31 March 2012 in most participating countries. The Survey of adult skills measures the essential competencies for information- processing in three domains: literacy, numeracy and problem solving in technology rich environments ([1], [2]). The measure of job mismatch that was analysed in this paper uses the PIAAC assessment regarding the proficiency for numeracy skill dimension.

Skill mismatch is concept based on whether workers have the actual skills needed to carry out successfully the tasks required by their current job. In order to identify the skill mismatch, the procedure indicated in *The survey of adult skills: Readers companion* was used [2]. Hence, first were identified the workers who self-report being well- matched using the answer to the following questions " Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job?" , "Do you feel that you need more training?" After that, for each skill dimension the minimum (5th percentile) and the maximum (95th percentile) proficiency by each 1 digit ISCO code were defined. The competency scores (plausible values), representing the distribution of a respondent's proficiency in each field, were taken into account in this stage. Finally those cases, for which the proficiency level exceeds the maximum, were classified as over-skilled. In order to take into account the replicate weights and the plausible values, R packages: survey, svyPVpack were installed and used ([3], [4]). The labour market mismatch was measured for UK dataset for the higher education graduates whose occupations are included in Major Group 1 and 2 (Managers and Professionals), according to the International Standard Classification of Occupations.

In order to estimate the impact of over skilling on earnings, an approach based on the principles of Propensity Score Matching (PSM) was used following the study of McGuinness [5]. The over skilled adults represented the treatment group while the well

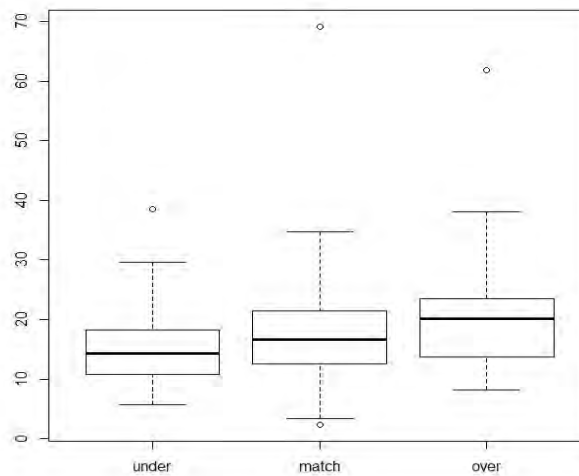
---

<sup>1</sup> Bucharest University of Economic Studies, National Scientific Research Institute for Labour and Social Protection

matched workers were included in the control group. The propensity scores were estimated by a logit model. The matching characteristics include: gender, age, area of study, economic sector, type of contract, current job industry and employment tenure (years).

### 3. RESULTS

Our estimation results, obtained for the UK data, regarding the measures of skills mismatch in numeracy show that 5.43% of higher education graduates within ISCO group 1 and 2 have the skills to cope with more demanding duties than those required by current job. In order to identify the occupational mismatch impact on earnings, we first compared the earnings distribution for the two groups. According to the representation in Figure 1, there are no strong evidences that over-skilled workers are suffering earning losses compared to matched workers.



**Figure 1. Impact of skill mismatch (numeracy) on earnings**

This conclusion was also supported by the estimation based on PSM, revealing that the differences in earnings for the matched pairs are not statistically significant.

**Table 1. Average treatment effect for the “treated”**

Estimate	-0.48
Standard Error	1.77
T-stat	-0.27
p.val	0.79

Table 1 presents the summary of the matching procedure. The treatment effect for the treated (over-skilled) is estimated to less than 0.5 decrease of hourly earnings. According to our results, this effect is not statistically significantly different from zero given that we obtained a high p-value (0.79).



Thus when comparing the outcome between treated (over-skilled) and control (well-matched) observations that are very similar to each other according to a propensity score, we found that the average effect of the treatment (over- skilling) is not significant.

#### **4. CONCLUSIONS**

The research undertaken in this study investigates the impact of job mismatch on labour market outcomes for higher education graduates. Previous studies proved that over education has a negative impact on higher education earnings. However earnings losses of overeducated adults could be explained by a lack of skills. Analysing mismatching phenomena using PIAAC data enabled a more detailed analysis given it provides a direct measure of skill mismatch. This approach overcomes the drawbacks of the methods based on qualifications which are not always a precise indicator of skills. Our preliminary results obtained for UK dataset, show that occupational mismatch measured by over skilling does not have a significant impact on the higher education graduates earnings.

#### **REFERENCES**

- [1] Berlingieri, F., Erdsiek, D. (2012). How Relevant is Job Mismatch for German Graduates?, Discussion Paper No. 12-075, <http://ftp.zew.de/pub/zew-docs/dp/dp12075.pdf>
- [2] OECD. (2013a). The methodology of the Survey of Adult Skills (PIAAC) and the quality of data, in OECD, The Survey of Adult Skills: Reader's Companion, OECD Publishing. doi: 10.1787/9789264204027-6-en
- [3] Reif. M., Peterbauer, J. (2014), Package 'svyPVpack'.
- [4] Lumley, T. (2014), Package 'survey'.
- [5] S. McGuinness, How biased are the estimated wage impacts of overeducation? A propensity score matching approach, Applied Economics Letters, Volume 15, Issue 2, (2008) , 145-149.

# Bayesian estimation approach in frameworks, integration of compilation and analysis

Jan W. van Tongeren<sup>1</sup> [jwvtongeren@gmail.com] and Ruud Picavet<sup>2</sup>

**Keywords:** Bayesian estimation, Frameworks, Reliability coefficients, National Economic Accounts, Satellite accounts.

## 1. INTRODUCTION

The Bayesian estimation approach in frameworks was initially designed for the purpose of national accounting. It makes use of five elements of the compilation of national accounts:

- (i) It has been developed in frameworks in which relations exist between the variables to be estimated in national accounting
- (ii) The relations include identities between variables (e.g. identities based on then definitions of GDP, disposable income, saving, etc.), and ratios variables (e.g. input-output ratios) that are conventionally used in national accounts compilation
- (iii) It incorporates the data available for variables and ratios as an integral part of the compilation.
- (iv) And finally it assigns reliabilities to these data and ratio values, which are also implicitly used in national accounting to adjust these values in order to arrive at framework consistency of the final estimates. See also Van Tongeren 2011 [1]

The compilation methodology conventionally used in national accounting has been formalized into a Bayesian estimation method, which subsequently has been computerized through the application of specifically developed software. As in conventional national accounting, the Bayesian estimation methodology is applied to frameworks of data.

The formalized national accounts compilation procedure was used in training sessions in a Netherlands project for Central American Central Banks, in order to develop an alternative to conventional compilation methods, speed up and improve the accuracy of the compilation through computerized procedures and provide means of incorporating over time updates of data and thus arrive at new estimates quickly. It was also used as an alternative to conventional national accounting methods in other projects in the Caribbean and Angola and in Kurdistan, the autonomous region of Iraq for which no national accounting estimates were made until now. Comparable methods are presently being used by Statistics Netherlands. [2].

Through the formalization of the method used in national accounting, it was possible to extend the method to satellite systems in which not only economic data are incorporated, but also physical and other data, as long as frameworks can be defined and relations between the variables of the framework can be incorporated (as in frameworks for national economic accounts, and also e.g. in frameworks for environmental accounts). Recently the method is being applied to the National Transfer Accounts Framework [3], first proposed by Lee and Mason [4] and subsequently described in a UN Manual in collaboration with the UN Population Division [5]. Here demographic and economic data are combined and the interaction between demographic and economic developments can be analyzed – e.g. the impact of the graying population on the economy - through inter-generational transfers that take place.

---

<sup>1</sup> Jan W. van Tongeren is ex-Chief of the National Accounts Branch, UN Statistics Division in New York

<sup>2</sup> Ruud Picavet is ex-Assistant Professor of Development Economics at Tilburg University, Netherlands

## 2. METHODS

The frameworks discussed here afterwards, are logical frameworks of data, variables to be estimated, ratios and identities between the variables, and prior reliability coefficients of data and ratios values. The variables are selected and the framework is constructed in order to estimate posterior values of all variables and their posterior reliability coefficients, and to serve predefined types of analysis. The variables and ratios incorporated in the framework are necessary and relevant for the desired analysis.

In this method, prior values are attached to data and ratio values and prior reliabilities are assigned to those values, based on subjective assessment of data and ratio values available. The reliabilities are expressed as reliability coefficients (standard deviation  $\sigma$ /value). Posterior estimates are then made by minimizing the squared differences between the prior and posterior values, under conditions of pre-established identities and ratios between variables, and also taking into account the prior reliabilities of the values. The inverted values of variances ( $1/\sigma^2$ ) are used as weights, so that prior data values with a higher variance may be adjusted more than data values with low variance, when arriving at posterior values.

The number of information items used in this method (available data, ratio values and identities) may be much larger than the number of variables to be estimated. It can be shown that an increasing number of information items improve the posterior reliability of the estimated values of variables and ratios within the logic of the framework.

In this compilation method, ratios are linearized, and can thus be included as elements of the least square expression. The method does not only arrive at posterior estimates of variables and ratios with and without data, but also at posterior reliabilities. The latter show how national accounting frameworks with a large number of data and ratio values and identity and ratio conditions may be used to arrive at posterior reliabilities that are much more precise than the prior reliabilities.

As prior values of ratios used in the compilation can also be used in their posterior format as analytical ratios, the framework also arrives at an analysis of the posterior estimates through the compilation and use of analytical ratios. In general, when converting in the Bayesian estimation approach prior information on compilation ratios that are traditionally used in national accounts compilation, into posterior estimates of analytical ratios, the approach may help to better integrate compilation and analysis of estimates.

## 3. RESULTS

The main potential of the Bayesian estimation approach in frameworks is that it responds effectively to the scarcity of data and to the use of those data in all kinds of analysis. To show this, the paper will include a number of applications. How this potential could be used in the future, will be discussed in the section dealing with “Conclusions”.

The quantitative assessment of the method is done by comparing values and posterior standard deviations of national accounts estimates, using Netherlands data, and frameworks in support of alternative analyses and based on comprehensive and limited data sets. The alternative analytical frameworks include the following:

1. First a framework based on a comprehensive national accounts data set for the Netherlands (2010 data) will be used to illustrate the working of the Bayesian method and to show how the reliability of national accounts estimates will be improved through this method. The latter improvement is quantified by comparing prior and posterior values and standard deviations of key variables and ratios.
2. Secondly, in a framework with 2012 data, the effect on posterior values and standard deviations of key variables and ratios will be measured, when only limited data are available. This is the case when early estimates are made for a year following 2010.
3. Using an extended framework, it will be shown how a full set of consistent estimates could be made by reconciling Integrated Economic Accounts (IEA) data on the

one hand, with financial accounts and balance sheet data on the other. The balance sheet data (for 2010 and also 2012) are inconsistent with the IEA data in the Netherlands. The effect on posterior values and standard deviations of key variables and ratios of the extension of the IEA will be quantified, as under point (1).

4. Using the extended framework as in point (3), it will be shown how projections for a future year (2012+n) could be made. The Bayesian projection would be based on a limited set of projected values of selected and strategic variables, and using identities, ratios and prior reliabilities, as used in point (3) for the year 2012.

#### Assessment of 2005 and 2006 Guatemala Bayesian estimates

	<i>Prior</i> reliability coefficient	Bayesian estimates 2005		Bayesian estimates 2006	
		Change in Value (as compared to conventional estimates)	<i>Posterior</i> coefficient of variation	Change in Value (as compared to conventional estimates)	<i>Posterior</i> coefficient of variation
GDP	3.00%	0.0%	0.01%	-0.1%	0.01%
HH Final consumption	3.00%	0.0%	0.00%	2.7%	0.00%
GOV Final consumption	0.10%	0.0%	0.01%	-0.2%	0.01%
NPI Final consumption	6.00%	0.0%	0.42%	0.5%	0.45%
Gross fixed capital formation	3.00%	0.0%	0.20%	-5.4%	0.20%
Exports	0.10%	0.0%	0.02%	-0.7%	0.09%
Minus: Imports	0.10%	0.0%	0.02%	1.1%	0.05%
Compensation of employees	12.0%	0.0%	0.13%	0.7%	0.14%
Taxes less subsidies on production	0.10%	0.0%	0.07%	0.6%	0.10%
Mixed income, gross	12.00%	0.0%	0.21%	-0.8%	0.22%
Non-HH Operating surplus, gross	12.00%	0.0%	0.17%	-0.6%	0.18%
HH Operating surplus (on dwelling services), gross	12.00%	0.0%	0.14%	0.2%	0.14%
HH disposable income	12.00%	0.0%	0.08%	1.1%	0.08%

In the present abstract only the results of the quantitative comparison between prior and posterior values and standard deviations, referred to under points (1) and (2) above, are presented below. For the time being the results of an earlier research project with Guatemala data (2005 and 2006 data) will be used; the data in the table are extracted from Van Tongeren [1]. In subsequent slides and poster sessions, the Guatemala data will be replaced by Netherlands data, and also the results of the quantitative assessments referred to under points (3) and (4) will be presented.

The quantitative results of the comparisons for 2005 Guatemala referred to under points (1) and (2) are summarized in the table above. Some conclusions that could be drawn from the quantitative results presented in the table are the following:

i. The differences between the Bayesian estimates and the national accounts estimates (Guatemala 2005 data) are very minor (all differences are 0.0%). This was to be expected, as the national accounts estimates were taken into account when making the Bayesian estimates. The differences between the prior and posterior values of the reliability coefficients, however, are significant. All estimates improve in accuracy, i.e. all posterior coefficients of variation are significantly reduced as compared to their prior values. E.g. the coefficient of variation of GDP reduces from 3.00% to 0.01%, the one of HH final consumption from 3.00% to 0.00%, the ones of exports and imports from 0.10% to 0.02%, the one of compensation of employees from 12.00% to 0.13%, and the one of HH disposable income from 12.00% to 0.08%.

ii. If only partial data are available (Guatemala 2006 data), the posterior estimates deviate more from the prior data. Thus, the largest deviations from conventional estimates refer to gross fixed capital formation (-5.4%), HH final consumption (2.7%), and HH disposable income (1.1%); all other estimates deviate less than  $\pm 1.0\%$ . The posterior reliability coefficients are higher for 2006 than for 2005, as less data are available. The differences between the values for the two years, however, are minor. Thus, Bayesian estimation within frameworks generates accurate posterior estimates, even in the case of limited data availability (Guatemala 2006 data). In the latter case, the estimated values are fairly close to the values of the data that were available from conventional national accounting. Estimates made within frameworks are generally more reliable, as identity and ratio relationships between variables are taken into account.

#### **4. CONCLUSIONS**

As mentioned above, the Bayesian estimation approach responds effectively to the scarcity of data and to the use of those data in all kinds of analysis. Furthermore, through its formalization and computerization, it results in more accurate estimates and less time to develop those than in case of conventional methods. In conventional accounting, analysis is “fixed” by the tables that are used - and generally internationally defined- in national economic and satellite accounts compilation. In the present approach, this limitation has been removed by defining alternative frameworks on the basis of variables that are needed in predefined analysis. Thus, if economic-environmental issues are discussed, the framework of the SNA is modified and extended with related environmental variables. Or if monetary-economic analysis is the objective, the framework incorporates variables of the IEA of the SNA, including the balance sheets and the financial accounts that record the changes therein. On the other hand, if economic-demographic analysis is needed to study the impact of a greying population, the economic variables are supplemented with related demographic variables, and with breakdowns by ages. The design of the framework is done in the first instance by ignoring available data, and focusing only on concepts/variables that best represent and support analysis. Thereafter the concepts/variables of the framework are adjusted, taking into account not only available data, but also the identity and ratio relations between variables that can be defined and are used in the Bayesian approach. The more such analytical relations can be incorporated the more precise posterior estimates can be made, and the richer the analysis based on the framework. On the other hand, ratios and identities can also be used to compensate for limited data availability. Thus, frameworks are defined that take optimally into account desired analyses (through identity and ratio relations), as well as data availability. These features are not available in traditional national economic and satellite accounting. Also future developments of data (e.g. in the development of “big data” sources) and of new focuses of policy analyses can thus be better supported by the Bayesian approach to frameworks.

#### **REFERENCES**

- [1] Jan W. van Tongeren, From National Accounting to the Design, Compilation and Use of Bayesian Policy Analysis Frameworks, PHD thesis defended on 14 October 2011 at Tilburg University, Netherlands.
- [2] Reinier Bikker, Jacco Daalmans en Nino Mushkudiani, Macro-integratie Deelthema: Inpassen, Statistische methoden (08003), Centraal Bureau voor de Statistiek, 2008
- [3] Jan W. van Tongeren, Review of NTA (National Transfer Accounts) from SNA perspective, appendix prepared in 2012 for an NTA Manual by The UN Population Division, and paper prepared for the US National Research Council in Washington, March 2013
- [4] Ronald Lee and Andrew Mason, Population Aging and the Generational Economy, A Global Perspective, Edward Elgar Publishing Limited (US and UK) and International Development Research Centre (Canada), 2001.
- [5] United Nations (Population Division, Department of Economic and Social Affairs), National Transfer Accounts Manual: Measuring and Analyzing the Generational Economy, E.13.XII.6, 2013

**From HOMBÁR to EAR**  
(Evolution of data processing in Hungary)  
Éva Laczka, [eva.laczka@ksh.hu](mailto:eva.laczka@ksh.hu)<sup>1</sup>,

**Keywords:** unified data processing, META databased system, data processing system governed by statisticians

## **1. Introduction**

After the turn of the Millennium Hungarian statisticians dealing with agricultural statistics applied successfully for an EU-tender aimed at developing agricultural statistics. The resources of the project made possible to review the data processing tasks of agricultural statistics. The renewal of data processing was justified by the fact that the traditional data processing system was relatively slow, it was not adequate for the efficient implementation of quality control, it was not sufficiently well-documented, and last but not least it required important human resources. Agricultural statisticians and IT developers developed a special, new data processing system that was called HOMBÁR (it means: Granary).

The concept of HOMBÁR is quite similar to the LEGO game; agricultural statisticians and IT experts created LEGO cubes (statistical operations) of different forms that were programmed by the IT experts. Using a comparison, the system functions in such a way that statisticians chose or assemble LEGO cubes according to their purpose of building “a horse or a tractor”. This means that if the statistical process changes, statisticians reorder the LEGO cubes (the system can react to the changes in a flexible manner). In this way data processing is led and managed by statisticians, the task of the IT team is “just” to secure the IT operation of the system and produce the new LEGO cubes (statistical operations). The development lasted 3 years and further 2 years were needed before the processing of agricultural (survey and admin) data with the new system became a routine for statisticians. The use of the HOMBÁR system made possible to reduce by half the time of data processing and by 30-40 per cent the human resource needs (in the case of IT people). The HOMBÁR provided a proper documentation, relation with the databases, and integrated more efficient quality controls. At that time agricultural statisticians considered the HOMBÁR not to be a suitable tool to process account type of data, like the Economic Accounts for Agriculture.

On the basis of the experiences of HOMBÁR, the HCSO decided in 2008 to extend the system to the whole statistical production of the office. As the whole statistical system is broader than agricultural statistics and is in a certain sense more complex, further development was needed. The new, extended system was called the Unified Data Processing System (in Hungarian EAR). The EAR compared to the HOMBÁR is a general data

---

<sup>1</sup> Affiliation

**Éva Laczka**  
*Deputy President*  
*h. professor*  
*Central Statistical Office*  
*E-mail: [eva.laczka@ksh.hu](mailto:eva.laczka@ksh.hu)*

processing system, which does not process the data of only one, or few statistical domains, but can be used for all the data collections and data sets of the HCSO.

The purpose of the paper is to present the concept, the functioning and the advantages of the new data processing system.

## **2. Methods**

### ***2.1 A META Database controlled system***

One of the major characteristics of statistical data processing is that it changes frequently, the scope of the data collected, the concepts linked to the processing, the methods, procedures of data processing can change through the years.

Consequently, the EAR concept was to be elaborated in a way to follow the changes without generating each time new programming work. The system created had to allow statisticians to describe changes concerning all the fields of our statistics “selecting and ordering the LEGO cubes”, without the need of new IT programming. One of the most widespread solution to follow changes in a dynamic way is the use of meta-database which led us to create a META database controlled system. The basic principle of the META database controlled operation is that the information needed for the functioning of the EAR system (the components necessary for statistical data processing, the different steps of data processing, the tasks in each step) are stored in the meta-database. The running of the EAR system relies on the use of the meta-database that contain the tasks to be implemented and the way of implementation.

### ***2.2 An integrated system***

The EAR system is linked at data level to the meta-database system of the HCSO, the data entry system and the Data Warehouse. It is linked through interfaces to tools for data analysis. The purpose is to fulfil the greatest amount of statistical processing functions, but in those cases when processing can be made more efficiently with existing software, data access or interface must be guaranteed.

### ***2.3 A system governed by statisticians***

Statisticians define and execute data processing using the meta-database and the integrated primary data processing steps. In the case of a change in the dataset and/or of the processing process, the modifications are performed by the statistician. The contribution of IT experts is “only” needed for the creation of the interfaces and primary operations.

### ***2.4 A transparent, well-documented and standardized system***

The EAR system stores the processing phases defined by the statistician in the meta-database system from where the documented information can be extracted. It is a unified, standardized processing system in the sense that different data collections, datasets and data from different periods can be processed together.

## ***2.5 The tasks of the statisticians***

In the development phase the task of the statisticians is to specify the requirements towards the system. It was necessary to define for each statistical domain the basic operations that were needed during data processing. Already at the beginning the aim was to create a comprehensive set of basic operations. An important task was to adjust the functioning META-database system of the HCSO to the needs of the EAR system which required the development of the META-database.

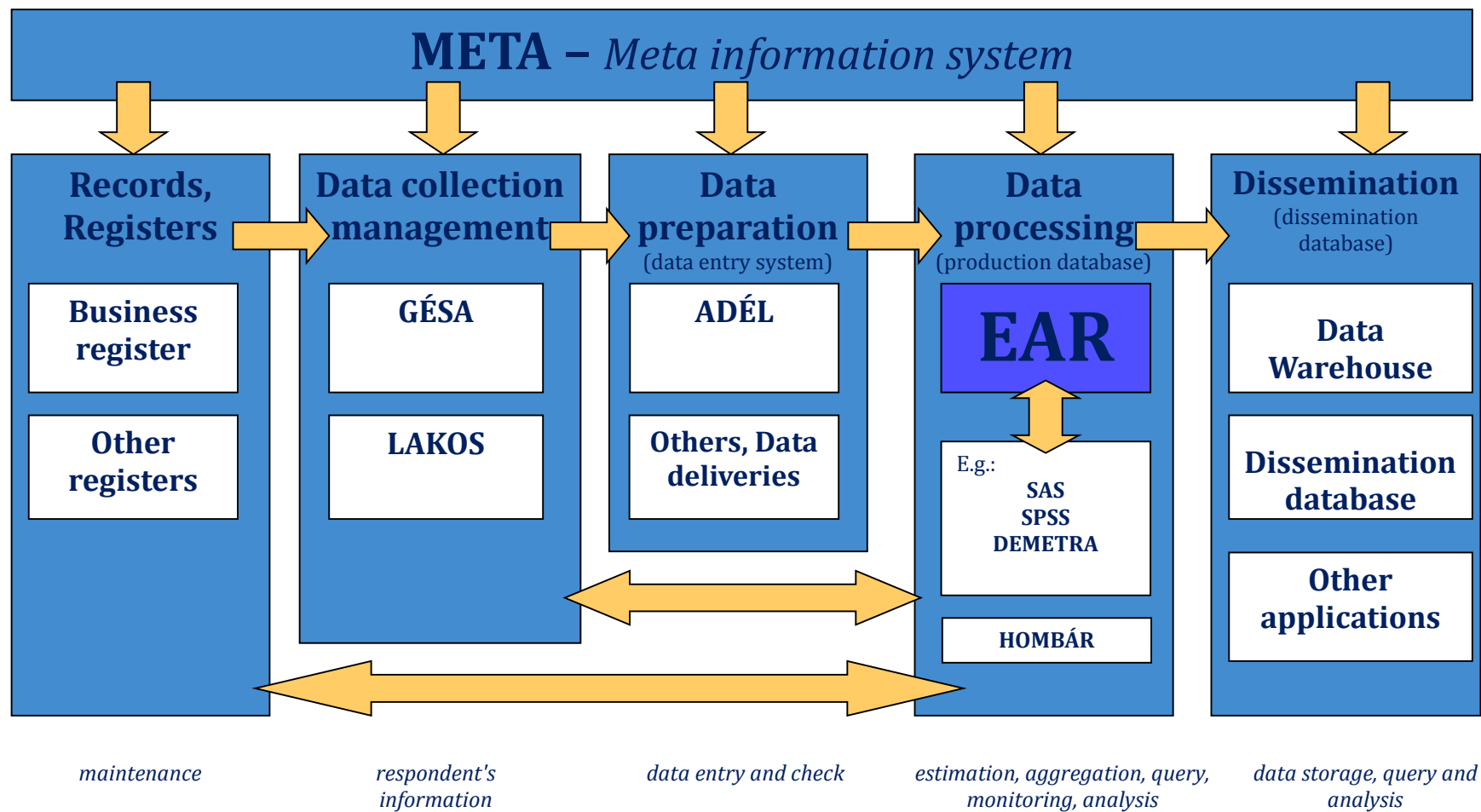
## ***2.6 IT tasks***

During the development phase of the system, the task of the IT team was to secure the connection with the systems functioning in the office, create the operating IT environment, but the development of the META-database required also the contribution of IT experts.



### 3. Results

#### EAR in the statistical data production process



#### **4. Conclusion**

After the turn of the Millennium Hungarian agricultural statisticians decided to develop a new data processing system. The renewal of data processing was justified by the fact that the traditional data processing system was relatively slow, it was not adequate for the efficient implementation of quality control, it was not sufficiently well-documented, and last but not least it required important human resources.

The concept of the new data processing system is quite similar to the LEGO game; statisticians and IT experts created LEGO cubes (statistical operations) of different forms that were programmed by the IT developers. Using a comparison, the system functions in such a way that statisticians chose or assemble LEGO cubes according to their statistical needs. This means that if the statistical process changes, statisticians reorder the LEGO cubes (the system can react to the changes in a flexible manner). In this way data processing is governed by statisticians, the task of the IT team is “just” to secure the IT operation of the system and produce the new LEGO cubes (statistical operations).

The first, new data processing system that got the name of HOMBÁR (meaning: Granary) was implemented between 2000 and 2005. On the basis of the experiences of HOMBÁR, the Statistical Office decided in 2008 to extend the system to the whole statistical production of the office.

The IT development phase of the EAR system started in 2008 and finished in 2011 including the test-phase, which was followed by the preparation and the training of the statisticians. The development work was accelerated at the beginning of 2013. It is not by chance that the department from where the idea rose (the Department of Agricultural and Environment Statistics) has been using EAR since 2011 (start of the system) for the processing of all its tasks enjoying all the advantages of the system (considerable saving in human resources, quicker data processing, better quality, well-documented processes, etc.).

Nowadays all the statistical departments of the Office (including the Department of National Accounts) use the EAR system, the routine operation by all of the departments is expected by the end of 2015.

# Robust Variable Estimation by Combining Administrative data sources

Guy Vekeman ([guy.vekeman@economie.fgov.be](mailto:guy.vekeman@economie.fgov.be))<sup>1</sup>

**Keywords:** Improving robustness through using para-data, Outlier treatment, Administrative proxies for statistical variables

## 1. INTRODUCTION

Structural Business Statistics relies on both administrative and complementary survey data in order to compile statistical variables. Among the administrative data sources, the profit and loss accounts are most valuable.

The imputation scheme developed, uses historic statistical data: previous values for the variables such as defined in the EC regulations on SBS, combined with historical and new administrative data. In this case a ratio imputation scheme most often provides the better estimations. It is however necessary to monitor the imputation process and branch to a more robust –but perhaps biased– imputation method whenever influential outliers might be generated.

The main need for imputing complete sets of SBS variables for small enterprises arises from the rotational design of the sampling scheme, whereby a few NACE sections are fully surveyed, whereas for the others only medium sized and large enterprises are included.

### 1.1. Administrative sources used for financial variables

When making estimations for financial variables, VAT tax data may be used alongside profit and loss accounts. While VAT data also refer to concepts such as turnover, purchases or investments, their definition differs from that generally used in accounting. This is the major challenge in the combined use of these administrative sources to make estimates of a complete set of SBS variables for a given enterprise.

The main problem arises for the abbreviated accounting scheme, whereby inclusion of the more essential data fields from the profit and loss account, such as turnover and purchases is not compulsory. In this case, there is no alternative but using VAT tax data. Since no cross-checking is possible on the values of the administrative proxies, we need to check coherence in an alternate way, using the value of the gross exploitation earnings from the abbreviated profit and loss account, which is the best proxy for the statistical variable “value added”.

### 1.2. Administrative sources used for personnel variables

The National Office for Social Security (ONSS) when providing the employment records is using its own employer identification number throughout the tables. In one of their tables the unique enterprise identification number, for universal use between the enterprise and any administration, is also included. The implicit hypothesis herein is that an employer corresponds to one single enterprise. Possible matching problems have been dealt with, as reported in [1]. Most of those are of a transitory nature.

---

<sup>1</sup> Statistics Belgium - (FOD Economie, AD Statistiek)

Employment characteristics can be estimated from social security data records and from yearly company accounts. Both sources are not necessarily coherent and definitions of administrative data fields differ, even though they may be labelled identically. Yet, even after carefully checking the metadata on the two administrative sources and including the data fields necessary to fully cover the definition of the statistical variable, the two sources don't yield the same result. It has been shown [2, 3] that annual accounts tend to overstate total personnel costs. The analysis method developed in these papers will be included among the SAS-tools used in the current paper.

## **2. METHODS**

The question faced when financial SBS variables need to be estimated for SME's mainly consists in the reconciliation of administrative proxies derived from VAT tax data with others found in the profit and loss accounts.

A first approach is to check how the turnover VAT proxies relate to the SBS turnover variable (12 110). Figure 1 shows the distribution of the SBS turnover divided by the VAT proxy for all enterprises surveyed. This distribution luckily peaks at unity value, confirming that the reported turnover SBS variable is most often coinciding with the VAT proxy. A close match for 60% of the cases and an acceptable match (within -20% to + 25%) in 86% of the cases are more or less reassuring.

The situation is more cumbersome for VAT purchases which often seem to exceed the SBS purchases variable (13 110). The distribution peaks around 0.66 and averages 0.76, making the VAT purchases proxy a highly biased (almost consistently overestimate) of the purchases SBS variable. The ratio distribution also is a lot broader; proving that the VAT proxy also is a much less reliable estimate for the SBS variable (figure 2).

### **2.1. Reconciliation procedure**

The vast majority of small businesses are using an abbreviated accounting scheme, whereby mentioning purchases and turnover in the profit and loss account is optional, while an operational margin (a good proxy for value added) is compulsory.

Having an input of three related proxies allows prior checking of the coherence between VAT proxies and value added as reported in the profit and loss (P&L) account. Whenever the P&L account also reports turnover and purchases, the VAT proxies will be disregarded, since P&L account values not only are coherent but also are thought to be the more reliable estimates, since their definition is closer to that of the SBS variables.

If turnover is absent from the P&L account, the first option is to use the VAT turnover and approximate purchases by subtracting value added. This imposes coherence of the proxies, but we completely disregard VAT purchases, albeit the less reliable proxy. As with any "one tool fits all" approach, we soon learnt that quite a few anomalies were generated in this way. In few aberrant cases value added exceeds VAT turnover, yielding a negative estimate for purchases.

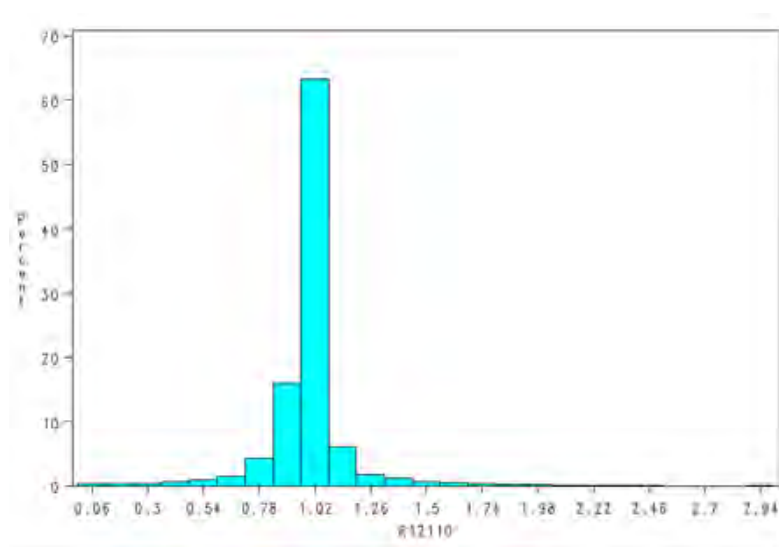
This may result from a mismatch between VAT turnover and accounting turnover. The former may include credit notes paid or revenues from the sale of assets. In both cases VAT turnover differs from accounting turnover. Apart from being defined differently, the value of VAT turnover is not consistently smaller or larger than accounting turnover, but the difference for any individual enterprise may vary over time.

In order to optimize proxies, their coherence: how well ‘purchases’ (Prch) plus ‘value added’ (VA) equals ‘turnover’ (TO), is what needs to be monitored. The quantitative procedure developed is based on the heuristic of what probably causes the mismatch between turnover and the sum of value added and purchases. A relative value difference (VD) between the two is calculated:

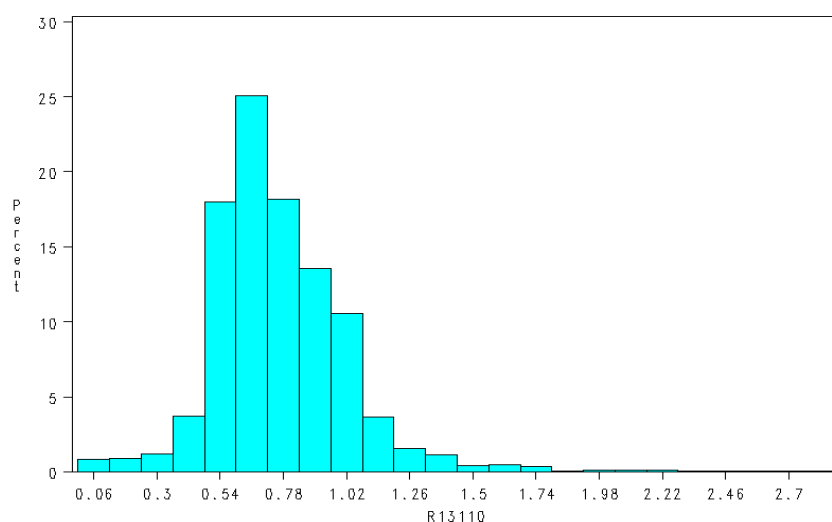
$$VD = 2 * (TO - (VA + Prch)) / (TO + (VA + Prch))$$

If all elements are positive this necessarily is constraint between -2 and 2. As we also need to cope with completely aberrant cases (negative values), the resulting difference is forced between -2 and 2. Subsequently, a different approach is followed depending on where the VD is situated.

### 3. RESULTS



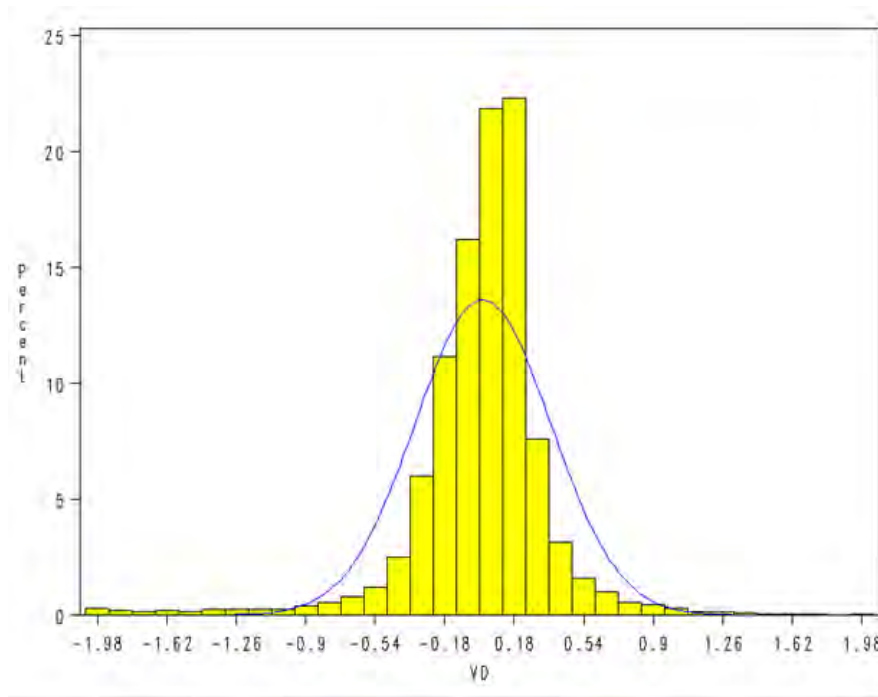
**Figure 1. :** Ratio of turnover as reported (SBS var. 12 110) to the VAT turnover proxy



**Figure 2. :** Ratio of purchases (SBS var. 13 110) to the VAT purchases proxy.

VAT purchases are biased relative to the SBS defined variable: most often VAT purchases are much higher. It has been pointed out that VAT data do not correctly discriminate purchases of goods and services from those of investment goods. In the

analysis of VD, the relative difference, VAT purchases first need to be corrected for the bias observed. As can be imagined, straight forward *scaling of VAT purchases is not the correct way* of handling this problem. The ratio of investments to total spending is higher for large industrial enterprises than for small retailers, just to name the extremes.



**Figure 3. :** Distribution of the relative value difference VD used in the proxy calculation.

The distribution analysis of this relative difference then shows what cases will prevail and what fraction needs to be corrected.

#### 4. CONCLUSIONS

Avoiding implausible proxies for turnover and/or purchases and eliminating spurious ratio imputations whenever runaway values tend to be produced, considerably reduces the risk of producing flawed aggregates.

#### REFERENCES

- [1] G. Vekeman, Using Social Security Administrative Records for Statistical Employment Variables, NTTS-2011 conference (2011), <http://www.cros-portal.eu/content/s2-presentation-4-ntts-2011-s2>
- [2] Frank Verschaeren et al., “Guide to checking the usefulness and quality of admin data”, ESSNet Admin Data – WP2 deliverable, Chapter 3.5, pages 144-159 [http://www.cros-portal.eu/sites/default/files//SGA%202011\\_Deliverable\\_2.4\\_b.pdf](http://www.cros-portal.eu/sites/default/files//SGA%202011_Deliverable_2.4_b.pdf)
- [3] Guy Vekeman, Confronting various administrative data sources to estimate employment variables, (2012) – A contribution to Q2012, Athens, Session 18., See: [http://www.q2012.gr/articlefiles/sessions/18.3\\_Guy%20Vekeman\\_AdminDataSource\\_sestimateEmployment.pdf](http://www.q2012.gr/articlefiles/sessions/18.3_Guy%20Vekeman_AdminDataSource_sestimateEmployment.pdf)

# Constructing Structural Earnings Statistics from Administrative Datasets

Kevin McCormack (Kevin.McCormack@cso.ie)<sup>1</sup>, Dr.Mary Smyth<sup>1</sup> (mary.Smyth@cso.ie)

**Keywords:** Structure of Earnings Survey (SES) – Administrative Data Project (SESADP), Big-data, Census, Modeling data, Unique Identifier, Identity Correlation, Central Statistics Office (CSO).

## 1. INTRODUCTION

The Central Statistics Office (CSO) in Ireland is currently engaged in compiling a complex Structure of Earnings Survey (SES) – Administrative Data Project (SESADP). This big-data project, which commenced in October 2013, will radically change how the CSO compile earnings statistics. Previously to provide SES statistics the CSO deployed an enterprise based survey of 50,000 individuals, titled “National Employment Survey”, which ran for 6 years (2002, 2005 to 2009).

The SES has at its core, the measurement of the relationship between remuneration and the individual characteristics of employees (i.e. gender, age group, level of education, etc.), and this sets the challenge for the SESADP. Fundamental to the SESADP is the design of a robust architecture. This will facilitate the better exploitation of existing business and social administrative data sources for the development of a master data source. This is titled the SES-Administrative Data Source (SESADS). This data source will contain information for approximately one million employees, firstly for 2011 and annually thereafter.

## 2. METHODS

The scope of the project consists of five sequential modules:

1. Identification of data sources and inventory of available and non-available characteristics for both individual and employer,
2. Linking of data sources,
3. Modelling of non-available characteristics,
4. Construction of the SESADS and
5. Publication of results.

### 2.1. Module 1 - Data sources and inventory of available characteristics

In the first module of the project a comprehensive review of existing data sources was undertaken. It was determined that in total seven primary data sources had potential to provide data for approximately 90% of the required SES variables, for both the individual and the employer. These data sources were five CSO data sources (two business and three social), one Dept. of Social Protection (DSP) and one Revenue Commissioners.

Primary Data Sources:

- Central Business Register (CBR),

---

<sup>1</sup> Central Statistics Office, Ireland

- Earnings, Hours & Employment Costs Survey (EHECS),
- 2011 Census of Population (COP),
- Quarterly National Household Survey (QNHS),
- Survey of Income and Living Conditions (SILC),
- Dept. of Social Protection (DSP) and
- Revenue Commissioners (P35L)

These data sources have been developed with varying attention to common identifiers. The business sources performed satisfactorily, while the social sources performed less favourably, as they have been developed using a stove-pipe approach.

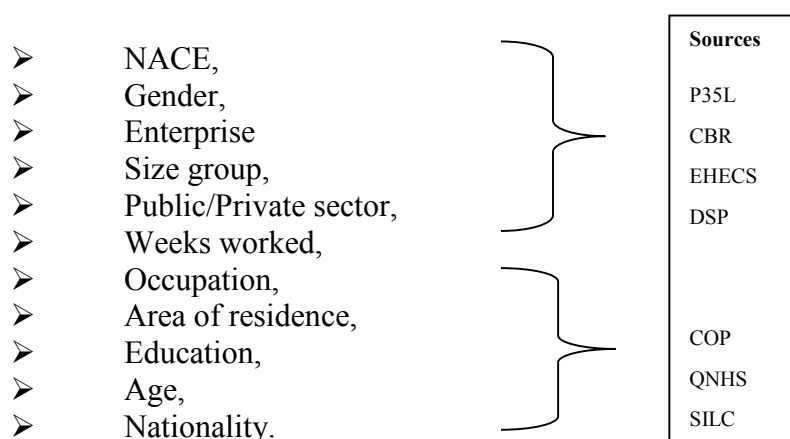
The Revenue P35L<sup>2</sup> file is the most accurate source of remuneration and, therefore, is the foundation of the SESADS. The 2011 COP and the DSP datasets will be the initial source for the individual characteristics of employees, with the QNHS and SILC being used in the inter-censal periods.

## 2.2 Module 2 - Linking of data sources

An analysis was undertaken of the data fields contained within the SESADS sources. The Personal Public Service Number<sup>3</sup> (PPSN) for employees and Enterprise Number (ent\_nbr) for employers were identified as the most suitable unique identifiers (UI) to link CSO's data sources, the DSP and Revenue Commissioners P35L data files.

Linking to the social data sources is a greater challenge for the CSO as they do not contain Unique Identifiers (UIs), such as a PPSN. UIs were developed by following an *identity correlation approach*, e.g. combining date of birth, area of residence and NACE<sup>4</sup>. This identity correlation approach enabled the social data sources to be linked with each other.

On completion of Module 2 the SESADS will contain for all employees in the State, Gross Annual Earnings classified by:



[1] The P35L dataset provides a complete register of all employees and therefore would provide a census of employees and eliminate any bias caused by sample selection.

[2] The Personal Public Service Number (PPSN) is a unique reference number assigned to individuals for use in transactions with public bodies or persons authorised by those bodies to act on their behalf.

[3] NACE is the "Statistical Classification of Economic Activities in the European Community"



### 2.3 Module 3: Modelling of non-available characteristics

The significant employee characteristics to be modelled are: (1) Annual bonuses, (2) BIK (benefit in kind) and (3) the determination of *full/part-time* employment status for employees.

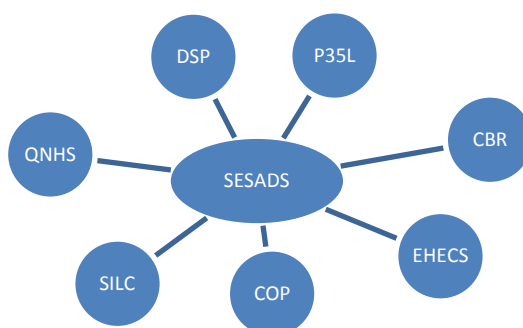
A multiple imputation methodology will be employed to carry out this stage of the Project. The QNHS and SILC data sources will be leveraged to provide the base information.

Once this model is completed, the SESADS will fulfil both the Eurostat annual and four yearly Eurostat SES earnings requirements.

### 2.4 Module 4: Construction of the SESADS

The SESADS will be constructed in the CSO's Administrative Data Centre (ADC) with its structures (known as layers) consistent with those as outlined in the ESSnet5 on micro data linking and data warehousing in statistical production.

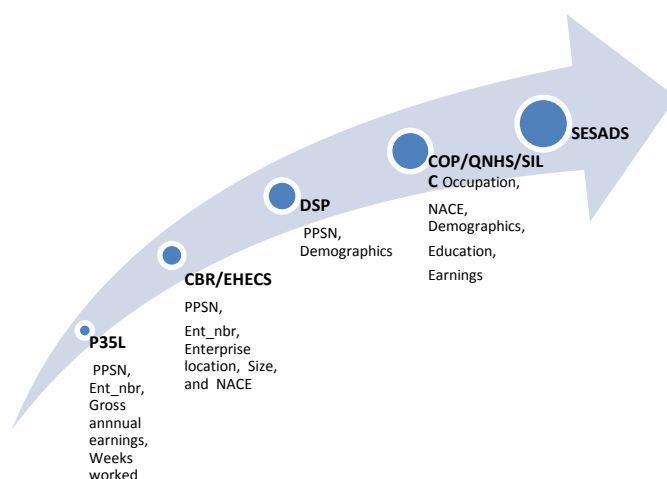
## 3. RESULTS



**Figure 1: SESADS primary data sources – Module 1: Data sources**

---

[4] During the Palermo meeting in September 2002, the DGINS (Directors General of NSI) expressed the need to find synergies, harmonization and dissemination of best practices in the European Statistical System (ESS). They proposed to create an adequate instrument: the Centres and Networks of Excellence (Cenex, now called ESSnet) projects, for putting together expertise distributed throughout the ESS organisations in order to develop specific actions which would benefit the whole system.



**Figure 2: Module 2: Linking of Data Sources**

**Table 1: Preliminary Results generated from SESADP**

Mean Annual Earnings Data (€) in 2011 - Nace Economic Sector, Full/Part-time status and Sex (only includes employees working 50 weeks or more)

NACE Rev.2	Fulltime employees			Part-time employees			Total employees		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
	€ Mean	€ Mean	€ Mean	€ Mean	€ Mean	€ Mean	€ Mean	€ Mean	€ Mean
<b>B-E</b>	51,270	40,539	48,547	26,786	24,391	25,293	26,786	24,391	25,293
<b>F</b>	46,057	40,301	45,486	20,478	20,734	20,571	20,478	20,734	20,571
<b>G</b>	43,625	32,459	38,340	15,899	15,361	15,480	15,899	15,361	15,480
<b>H</b>	46,070	38,125	44,683	20,222	25,189	23,080	20,222	25,189	23,080
<b>I</b>	34,659	28,489	31,434	14,626	13,013	13,590	14,626	13,013	13,590
<b>J</b>	60,293	45,014	56,242	26,641	24,183	24,864	26,641	24,183	24,864
<b>K-L</b>	74,606	50,586	62,031	46,550	28,222	29,637	46,550	28,222	29,637
<b>M</b>	64,142	44,003	56,825	33,975	15,055	18,965	33,975	15,055	18,965
<b>N</b>	44,990	35,825	40,242	20,733	20,021	20,151	20,733	20,021	20,151
<b>O</b>	53,003	45,356	49,690	25,172	28,142	27,745	25,172	28,142	27,745
<b>P</b>	61,934	53,764	56,121	27,373	24,666	25,012	27,373	24,666	25,012
<b>Q</b>	53,827	42,959	45,479	21,943	21,844	21,850	21,943	21,844	21,850
<b>R-S</b>	43,713	36,051	39,351	17,465	17,050	17,181	17,465	17,050	17,181
<b>Total</b>	52,442	42,986	48,094	20,636	20,565	20,578	20,636	20,565	20,578

NACE Rev.2
B-E Industry
F Construction
G Wholesale and retail
H Transportation and
I Accommodation and Food
J Information and
K-L Financial, insurance,
M Professional, scientific &
N Administrative and
O Public administration &
P Education
Q Health & social work
R-S Arts, entertainment,

## 4. CONCLUSIONS

The SES Administrative Dataset Project (SESADP) provides the main Eurostat annual earnings requirements. Eurostat's *Gentleman's Agreement* to deliver earnings data for the *Gender Pay Gap* and *Annual Gross Earnings* are provided. The regulatory requirement to deliver the four-yearly EU SES microdata (SES 2014) is provided by the addition of the EHECS, SILC and QNHS variables, and modelling these variables to provide full employee coverage.

# The sensitivity of job reallocation measures to longitudinal linkage problems

Karen Geurts (karen.geurts@kuleuven.be)<sup>1</sup>

**Keywords:** Job creation and destruction, firm record linking.

## 1. INTRODUCTION

In the past decades, research into firm and employment dynamics has received strong impetus from improved accessibility to large-scale official datasets. The pitfalls associated with the use of these data are well understood [1][2]. One of the main difficulties is that firms may change identifier for a variety of reasons such as a change in ownership, legal form, merger or split-up. This leads to an overestimation of aggregate firm and employment dynamics and a bias in firm-level estimates [3][4]. To address these problems, statistical agencies have developed record linking methods to identify and repair missing links between firm identifiers. Two sets of methods, which take an entirely different approach to the firm linking problem, are now commonly implemented. The first relies on conventional record linking methods and uses information from additional business data sources and probabilistic matching techniques to link firm identifiers. The second relies on linked employer-employee data and uses one key input factor of the firm, the workforce, to identify changes in identifiers and firm structure. Both conventional and employee-flow methods are found to remove a substantial amount of ‘spurious’ firm and employment reallocation from the data [5][6]. Less is known, however, about the bias that is left in dynamic measures after the linkage methods have been implemented.

This paper empirically evaluates the sensitivity of widely-used measures of firm and employment dynamics to the longitudinal linkage problem. We consider entry and exit rates, the employment distribution at entry and exit, job creation and destruction rates, and firm-level growth rates. We aim at empirically comparing how well the two types of record linking methods perform in producing reliable means of these measures. Using panel data on the population of Belgian private employer firms in 2003-2012, we apply two linking methods that are illustrative examples of the conventional and the employee-flow approach. We calculate the empirical measures before and after application of each individual method, and compare them to benchmark results obtained by both methods combined. The benchmark results enable us to evaluate the bias left in the measures by each method separately.

## 2. METHODS

The first set of firm linkages we implement relies on conventional record linking methods and has been developed by Statistics Belgium. The approach follows the Eurostat/OECD recommendations for the construction of harmonised business registers and statistical indicators on firm demography [7]. To identify changes in firm identifiers and firm structure, the method makes use of information on firm continuity from supplementary data sources, such as Commercial Court files and social security data. In addition, firm identifiers are linked by a probability-based matching procedure, which exploits similarities in firm name, address and 4 digits industry code.

---

<sup>1</sup> Center for Economic studies, University of Leuven (KU Leuven)

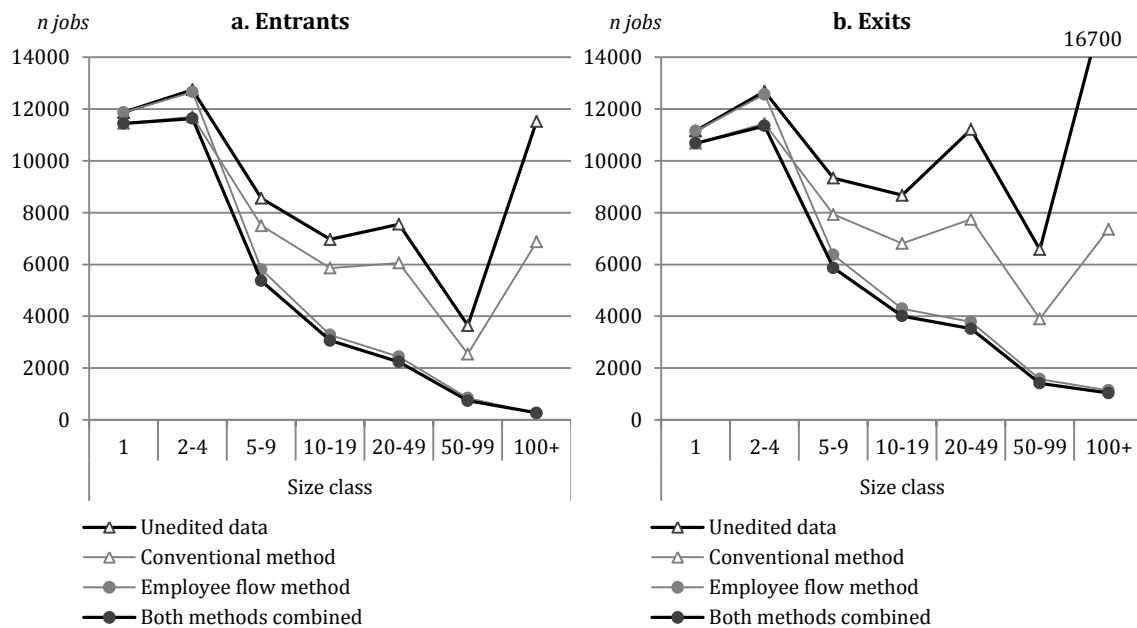
The second set of linkages is based on an employee-flow approach [8][9]. If a firm changes identifier but continues its operations, one of the main production factors, the stock of employees, is likely to remain largely the same. Similarly, if firms are merged, split up or parts are sold to another firm, this will be reflected in a division or merger of workforces. Continuity of the workforce can thus be used to trace the continuity of the firm. The method identifies large clusters of employees that ‘move’ between two firm identifiers in a short period of time, i.e. between two quarterly observations. The employee flows are used to signal changes in identifiers or firm structure. The method applied in this paper has been developed in collaboration with the National Social Security Office taking, similar methods in other countries as an example.

We finally construct longitudinal firm linkages that incorporate all information provided by both the conventional and the employee-flow method. Links edited in this way are the most accurate longitudinal firm records that can be obtained with the available methods. They are used to calculate benchmark measures against which results obtained by the two individual methods are compared.

### 3. RESULTS

We here report the results for two key measures of employment dynamics: the employment distribution at entry and exit, and job creation and destruction rates. In the full paper, additional results are discussed with respect to entry and exit rates, annual variance of job reallocation rates, and firm-level growth rates.

Panel a. of Figure 1 presents the distribution of employment created by new firms at entry, and panel b. the employment distribution of firms in the year of exit. The top lines represent the results based on the unedited data, and the other three lines show the results after implementation of the linkage procedures, i.e. after ‘spurious’ entrants or exits have been removed. Spurious entrants are established firms that are misclassified as entrants due a change in identifier or firm structure. Spurious exits are continuing firms misclassified as exits for similar reasons.



**Figure 1. Employment distribution of entrants and exits**

Note: Annual averages over the 2003-2012 period

The benchmark results (lower lines) show that job creation by real entrants and job destruction by real exits is highly concentrated in the smallest size categories. In the unedited data (top lines), the distribution are strongly shifted to the right: missing links between firm identifiers introduce an important upward bias in the employment shares of medium and large size classes. The conventional link method leaves a substantial bias in the data. The method fails to identify an important part of spurious entrants and exits in medium and large size classes, which disproportionately affect the results. The employee flow method, by contrast, is effective in identifying misclassified entrants and exits in the most critical size classes above 10 employees. The method strongly reduces the initial biases and reveals entry and exit patterns that closely resemble the right-skewed distributions obtained by the benchmark results.

Table 1 presents summary measures of job reallocation before and after the linkage procedures. Net employment growth is decomposed into job creation by entrants and expanding firms, and job destruction by exiters and contracting firms, following [10].

**Table 1. Annual job creation and destruction rates**

	Net growth rate	Job creation rate			Job destruction rate		
		Total	By entry	By expansion	Total	By exit	By contraction
<b>a. Rates (%)</b>							
Unedited data	1.03	9.01	2.52	6.48	7.98	3.06	4.92
Conventional method	1.03	8.17	2.09	6.08	7.14	2.24	4.89
Employee flow method	1.03	7.24	1.49	5.75	6.21	1.64	4.57
Both methods combined	1.03	7.06	1.39	5.67	6.03	1.52	4.51
<b>b. Percent bias vs. both methods combined</b>							
Unedited data		28%	81%	14%	32%	102%	9%
Conventional method		16%	50%	7%	18%	48%	9%
Both methods combined		3%	7%	1%	3%	8%	1%

Note: Annual averages over the 2003-2012 period

Total employment increased by 1.03% on average per year in period of observation (2003-2012). Statistics based on both methods combined show that the net growth rate is the result of an average annual job creation rate of 7.06 per cent and a job destruction rate of 6.03 per cent. If based on unedited data, the rates are overestimated by about 2 percentage points, or 28 percent and 32 per cent respectively. The bias is mainly due to an overestimation of the job creation rate by entry (81%) and the job destruction rate by exit (102%). Job reallocation rates by expanding and contracting firms are less strongly overestimated. In line with the results discussed before, the employee flow method yields job creation and destruction rates that are close to the benchmark results, leaving only 3 percent bias in the total job reallocation rates. The conventional method, leaves a substantial upward bias in the measures. Most noticeable is the 50 percent overestimation of the job creation rate by entry and the 48 percent overestimation of the job destruction rate by exit.

#### 4. CONCLUSIONS

Our main findings are summarized as follows. Larger firms are disproportionally affected by the longitudinal linkage problem and introduce an upward bias in job creation and destruction rates, especially at entry and exit. Missing linkages lead to employment distributions at entry and exit that are strongly biased towards larger firms, to spurious

variation in annual job reallocation, and to an underestimation of firm-level growth rates in medium and large size classes. The two linkage methods yield markedly different results. The employee-flow method is most effective in capturing missing links in the critical size classes above 10 employees and strongly reduces the initial biases. The results from this method are close to the benchmark measures and reveal highly right-skewed employment distributions at entry and exit, moderate job creation and destruction rates, and remarkably stable job reallocation by entrants and exits over time. The conventional method, by contrast, leaves an important bias in all measures of employment dynamics because it misses an important share of longitudinal linkages in larger size classes.

Our findings have important implications for policy measures based on administrative registers. Employment statistics based on poorly edited longitudinal data support the common perception that entrants and small firms are the engine of job creation. Instead, improved longitudinal linkages reveal that job creation by entrants is extremely modest, and that large established firms contribute more positively to employment growth than smaller ones..

## REFERENCES

- [1] T. Dunne, M.J. Roberts and L. Samuelson, Patterns of Firm Entry and Exit in U.S. Manufacturing Industries, *RAND Journal of Economics*, 19:4 (1988), 495-515.
- [2] S.J. Davis, J.C. Haltiwanger and S. Schuh, Job creation and destruction. Cambridge: MIT Press (1996).
- [3] J.R. Spletzer, The contribution of establishment births and deaths to employment growth, *Journal of Business and Economic Statistics*, 18:1 (2000), 113–126.
- [4] K. Geurts and J. Van Biesebroeck, Job creation, firm creation, and de novo entry, CEPR Discussion Paper no. 10118 (2014) .
- [5] J.C. Pinkston and J.R. Spletzer, Annual measures of job creation and job destruction created from quarterly microdata, Bureau of Labor Statistics Statistical Survey Papers (2002).
- [6] G. Benedetto, J.C. Haltiwanger, J. Lane and K. McKinney, Using worker flows to measure firm dynamics, *Journal of Business and Economic Statistics*, 25:3 (2007), 299–313.
- [7] Eurostat-OECD, Manual on Business Demography Statistics, Luxembourg: Office for Official Publications of the European Communities (2007).
- [8] J.R. Baldwin, R. Dupuy and W. Penner, Development of longitudinal panel data from business registers: the Canadian Experience, *Statistical Journal of the United Nations* 9 (1992), 289-303.
- [9] id. [6]
- [10] id. [2]

# New Approach to Gross Domestic Product Decomposition

ANTE ROZGA<sup>1</sup>, ([ANTE.ROZGA@EFST.HR](mailto:ANTE.ROZGA@EFST.HR)), ELZA JURUN<sup>1</sup>, IVAN ŠUTALO<sup>2</sup>

**Keywords:** Törnquist, Fisher and Lloyd-Moulton model, GDP decomposition, superlative indices, elasticity of substitution

## 1. INTRODUCTION

In the focus of this paper is a new methodological approach to upgrading the statement of GDP growth rates and implicit GDP deflators – on annual and quarterly bases. For a long time in the practice of national statistical agencies the chain-linking methodology has been used. By means of chain linking index number drift has been resolved partially in the sense of the second best solution. As time passes Laypeyres index with fixed base substantially overestimates Paasche index as further as index base is being left in the past. Paasche price index is lower compared to its Laspeyres counterpart but it is the most appropriate GDP deflator due to statistical (Cauchy theorem) and economic (substitution-transformation effect) reasons. Putting together Lloyd-Moulton with Törnqvist and Fisher indices authors have constructed Lloyd-Moulton-Törnqvist-Fisher (LMTF) model. LMTF model improves GDP price-volume decomposition due to more precise substitution measurement. Fisher index supported by LMTF model has been also built and it resolves the problem of additive (absolute and relative) inconsistency in GDP data. Another significant achievement of the paper is keeping product test identity (volume = volume times price). An integral part of the survey are testing results which prove that Fisher index supported by LMTF model can be considered as "ideal" in the practical applications.

## 2. METHODS

### 2.1. Lloyd-Moulton-Törnqvist-Fisher index and Fisher index supported by Lloyd-Moulton-Törnqvist-Fisher counterpart

The central point of this paper is construction of LMTF index, which measures GDP decomposition better than classic chain-linking methodology does. The complete estimation procedure has been carried out on the case study of Croatia. Original data sources used for LMTF calculation are Croatian annual and quarterly GDP data from q1 2000 to q4 2007 shown in data files: AGDP current prices, QGDP current prices, AGDP chain linked and QGDP chain linked. The four mentioned data files are shown the most up-left in "Fig 1". The most demanding part of LMTF (I) calculation, the first variant of LMTF model, has been done by econometric Lloyd-Moulton (LM) estimation. The central point of this estimation was calculation of 28 elasticities of substitution  $\sigma_j^{LM}$ , one for each q1 2000 – q4 2007 quarter. In order to calculate these elasticities, QGDP relative price deflators ( $I_i$ ) and relative QGDP shares  $s_i^j$  – at previous year's prices – have to be calculated. Both of the two just mentioned sets of indicators consist of 1540 pairs (56 NACE classes =>  $56 \cdot (56-1)/2 = 1540$  pairs) relative  $I_i$  and  $s_i^j$ .

---

<sup>1</sup> Faculty of Economics, University of Split, Cvite Fiskovića 5. 21000 Split, Croatia

<sup>2</sup> Zagreb School of Economics and Management, Jordanovac 110, 10000 Zagreb, Croatia

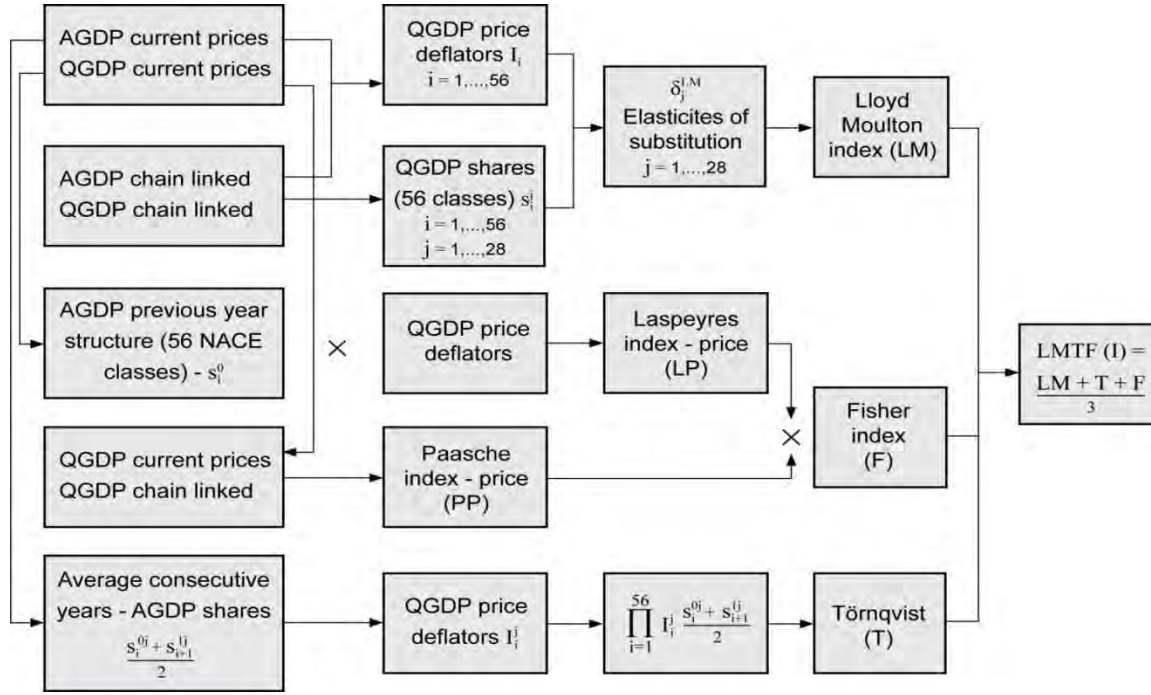


Figure 1. Scheme of the estimation procedure with the original data sources and intermediary tables for calculating Lloyd-Moulton-Törnqvist-Fisher index of type I.

Changes of GDP shares, among 1540 industries and between the two consecutive years (the same quarter of the current year through the same quarter of the previous year) and QGDP price deflators (just among 1540 industries) are in reverse order what is consistent with substitution behaviour of the Croatian producers. Namely, if GDP in industry  $j$  is getting “relative more expensive” compared to industry  $i$ , GDP share in  $i$ -th industry has to go down compared to industry  $j$ , and vice versa. Elasticities of substitution  $\sigma_j^{LM}$  are derived from econometric estimation of equation (1):

$$\ln \left[ \frac{\left( \frac{s_i^{qt}}{s_j^{qt}} \right)}{\left( \frac{s_i^{qt-1}}{s_j^{qt-1}} \right)} \right] = \sigma * \ln(P_j^{qt} / P_i^{qt}) + u_i, \quad \forall (i, j) \quad i, j = 1 \dots 1540 \quad (1)$$

Parameter  $\sigma_j^{LM}$  is classic elasticity coefficient known from economic literature. Looking at econometric estimation of  $\sigma_j^{LM}$  parameters, their significance and stability are of the crucial importance. Although data used in (1) are panel – especially on the right side of this equation, relative deflators are calculated for all 1540 among pairs of 56 NACE classes. The left side of (1) demonstrates panel data features. The data are among industries – cross section data - and in time – two consecutive years, which is characteristic of time series. Due to the time dimension of the data, the first passage through econometric software showed high positive autocorrelation demonstrated by very low Durbin-Watson (DW) statistics. In order to cure high positive autocorrelation, AR(1) has been applied differencing of the data. After the second passage through the econometric software the following  $\sigma_j^{LM}$  estimates, have been obtained:

Table 1. Estimates of 28 elasticities of substitution among 1539 industries' pairs by AR(1) transformation.

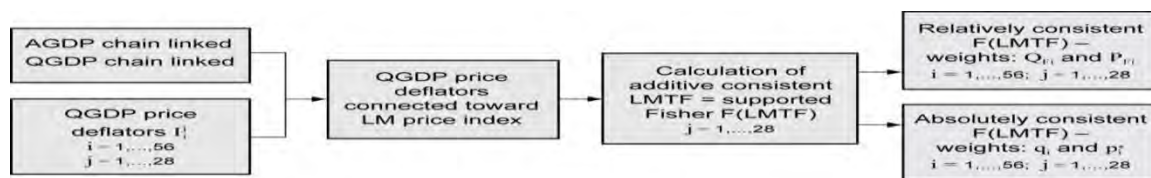


Quarter (2)	Elasticities of substitution estimates (2)	t – statistics (3)	p - values t-stat. (4)	F – statistics (5)	p - values F stat. (6)*	DW (7)
q1 -2001.	0,0100	0,4247	$0,6711 \times 10^0$	0,1804	$0,6711 \times 10^0$	2,2679
q2 -2001.	0,2539	11,3435	$1,0645 \times 10^{-28}$	131,4093	$1,0645 \times 10^{-28}$	2,4105
q3 -2001.	0,2271	11,7450	$1,4000 \times 10^{-30}$	137,9443	$1,4000 \times 10^{-30}$	2,4354
q4 -2001.	0,1672	9,8756	$2,4146 \times 10^{-22}$	117,9162	$2,4146 \times 10^{-22}$	2,2693
q1 -2002.	0,6926	25,1191	$5,8000 \times 10^{-117}$	630,9682	$5,8000 \times 10^{-117}$	2,0321
q2 -2002.	0,7026	38,9803	$9,7000 \times 10^{-232}$	1519,4632	$9,7000 \times 10^{-232}$	2,2398
q3 -2002.	0,6775	38,6095	$1,4013 \times 10^{-228}$	1490,6898	$1,4013 \times 10^{-228}$	2,5561
q4 -2002.	0,5165	26,0736	$2,0657 \times 10^{-124}$	679,8304	$2,0657 \times 10^{-124}$	2,4679
q1 -2003.	0,8069	18,2341	$2,0857 \times 10^{-67}$	332,4825	$2,0857 \times 10^{-67}$	1,9642
q2 -2003.	0,9085	27,3031	$3,600 \times 10^{-134}$	745,4601	$3,600 \times 10^{-134}$	1,9660
q3 -2003.	1,0955	44,1131	$2,2357 \times 10^{-235}$	1945,9640	$2,2357 \times 10^{-235}$	2,1551
q4 -2003.	0,4506	6,6662	$3,6470 \times 10^{-11}$	44,4387	$3,6470 \times 10^{-11}$	1,0799
q1 -2004.	-0,0266	-0,7740	$0,4390 \times 10^0$	0,5991	$0,4390 \times 10^0$	2,12025
q2 -2004.	0,5100	17,1599	$1,5832 \times 10^{-60}$	294,4616	$1,5832 \times 10^{-60}$	2,1185
q3 -2004.	0,5717	23,3486	$1,8647 \times 10^{-103}$	545,1588	$1,8647 \times 10^{-103}$	2,1544
q4 -2004.	0,5384	26,5078	$7,7496 \times 10^{-128}$	702,6641	$7,7496 \times 10^{-128}$	2,2991
q1 -2005.	0,0370	1,5858	$0,1130 \times 10^0$	2,5146	$0,1130 \times 10^0$	2,29333
q2 -2005.	0,0956	10,7141	$6,9736 \times 10^{-26}$	114,7921	$6,9736 \times 10^{-26}$	0,6148
q3 -2005.	-0,2595	-11,1775	$6,0709 \times 10^{-28}$	124,9358	$6,0709 \times 10^{-28}$	2,3925
q4 -2005.	-0,2718	-13,7523	$1,1321 \times 10^{-40}$	189,1249	$1,1321 \times 10^{-40}$	2,5001
q1 -2006.	0,2618	14,2912	$1,3477 \times 10^{-43}$	204,2378	$1,3477 \times 10^{-43}$	2,4414
q2 -2006.	-0,1507	-5,5595	$3,1844 \times 10^{-8}$	30,9082	$3,1844 \times 10^{-8}$	2,4864
q3 -2006.	-0,2553	-8,3247	$1,8435 \times 10^{-16}$	69,3012	$1,8435 \times 10^{-16}$	2,5127
q4 -2006.	0,0377	2,4136	$0,0493 \times 10^0$	204,2378	$0,0493 \times 10^0$	2,4136
q1 -2007.	0,1794	14,9943	$1,5405 \times 10^{-47}$	224,8279	$1,5405 \times 10^{-47}$	2,4752
q2 -2007.	-0,0958	-3,5603	$0,0004 \times 10^0$	12,6758	$0,0004 \times 10^0$	2,6488
q3 -2007.	0,0316	1,5769	$0,1150 \times 10^0$	2,4865	$0,1150 \times 100$	2,2983
q4 -2007.	-0,0586	-3,7962	$0,0002 \times 10^0$	14,4113	$0,0002 \times 10^0$	2,2954

## 2.2. Fisher index supported by Lloyd-Moulton-Törnqvist-Fisher counterpart

Beside the prime goal of the paper-improvement of GDP price-volume decomposition, the second not less important goal has been resolving of additivity problem (see Figure 2). This is not as important for the quality of GDP compilation as it is for the quality of GDP publication (dissemination). Namely, users like to see GDP components (in volume terms) additive into aggregate.

Figure 2. Construction scheme of Fisher index supported by Lloyd-Moulton- Törnqvist-Fisher counterpart.



Following procedure announced in Figure 2. Fisher index supported by Lloyd-Moulton relative additive consistent decomposition of quarterly GDP in volume terms is calculated.

Detailed calculation and comments are not shown here due to the shortage of this kind of paper. It would be presented to the reviewers if requested or in the full paper. The same comment is valid for 2.1.

### 3. CONCLUSIONS

By means of chain linking, index number drift has been resolved partially in the sense of the second best solution. Index number mathematics provides a better solution. By its theoretical considerations Törnqvist and Fisher indices have been chosen among so called “superlative indices” as superior ones for the GDP compilation. According econometric estimations Lloyd-Moulton index has been also calculated as the best estimator of elasticity of substitution. Putting together Lloyd-Moulton with Törnqvist and Fisher indices authors have constructed Lloyd-Moulton-Törnqvist-Fisher (LMTF) model. LMTF model improves GDP price-volume decomposition due to more precise substitution measurement. Fisher index supported by LMTF model has been also built and it resolves the problem of additive (absolute and relative) inconsistency in GDP data. The whole estimation procedure has been implemented on the case study of Croatia. The data base dealing with Croatian Quarterly GDP data has related to the period from q1 2000 to q4 2007. Thanks to the approach proposed in this paper, ex-post smoothing of the preliminary raw-data driven by original (price and volume) indicators preserves indicators content of GDP data but improve “mature” of GDP data. An integral part of the survey are testing results which prove that Fisher index supported by LMTF model can be considered as "ideal" in the practical applications. Namely, the new methodological approach proposed in this paper has at least three advantages: a) better decomposes “mature” GDP data on price and volume, b) assures additive consistent GDPs for publication and c) preserves (by means of F supported by LMTF) product test identity (value = volume times price). This is the reason to choose it.

### REFERENCES

- [1] Douglas W. Caves, Laurits R. Christensen and W. Erwin Diewert, *The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity*. *Econometrica*, Vol. 50 (November 1982), pp. 1393-1414.
- [2] Yuri Dikhanov, *The Sensitivity of PPP-Based Income Estimates to Choice of Aggregation Procedures*, Mimeo, International Economics Department, The World Bank (1997), Washington DC.
- [3] W.E. Diewert and A.O. Nakamura, *Essays in Index Number Theory, Volume 1*, North Holland, ELSEVIER SCINCE PUBLISHERS B.V.(1993). Amsterdam.
- [4] W.E. Diewert, *The Quadratic Approximation Lemma and Decompositions of Superlative Indexes*. *Journal of Economic and Social Measurement* 28 (2002). 63-88.
- [5] ILO, IMF, OECD, UN, Eurostat and The World Bank, *Consumer Price Index Manual: Theory and Practice (2004)*. Geneva, International Labour Office.
- [6] IMF, ILO, OECD, UN, Eurostat and The World Bank, *Producer Price Index Manual: Theory and Practice (2004)*. Washington, International Monetary Fund.
- [7] Ivan Šutalo, *Theoretical and Practical Implications of the Substitution Effect Impact onto Gross Domestic Product Decomposition*, doctoral dissertation (2012). Split. Croatia.

# BACKWARD RECALCULATION OF LABOR FORCE INDICATORS: A CASE OF TURKEY

Necmettin Alpay KOÇAK\*  
alpaykocak@tuik.gov.tr

Enes Ertad USLU†  
enesuslu@tuik.gov.tr

Özlem YİĞİT‡  
ozlemyigit@tuik.gov.tr

**Keywords:** backcasting, labor force, interpolation, unobserved components.

## 1. INTRODUCTION

Official statistics, especially economic time series (e.g. short-term business statistics, quarterly national accounts), are often subject to methodological and conceptual changes such as classification, definition and/or methodological changes. Although implementation of new definition to these statistics provides a better representative of economy, it creates an obstacle on sustainability of time series. To avoid the problem of unsustainable (incomparable) economic time series, statistical offices revise the old time series according to new definitions using by micro/macro approaches.

Household Labour Force Survey being implemented regularly since 1988 by Turkish Statistical Institute, is the main data source which provides information about those employed; economic activity, occupation, employment status and working hours of individuals in the labor market. In this paper, we attempt to implement various backcasting methods to time series are obtained from Household Labour Force Survey.

## 2. METHODS

Recently, Turkish labor force statistics are exposed to 3 different kinds of revisions, namely, 1.Changing the duration of job search used in unemployment criteria, 2.The use of new population projection estimates and amendment of new administrative division, 3. Other survey application changes (continues survey implementation, sampling design, etc.).

### 2.1. Duration of job search

The first revision is to changing the duration of job search used in unemployment criteria in the context of new regulations put into practice in Household Labour Force Survey within the framework of the European Union criteria. The unemployed comprises all persons (people) 15 years of age and over who were not employed during the reference period, had used at least one channel for seeking a job during the last 4 weeks and were available to start work within two weeks. Before 2014, the reference period of job search was "last three months".

The advantage of using micro method in backcasting is that the survey contains the questions which are asking the person's job search situation in both 4 weeks and 3 months for all surveys conducted from 2005.

---

\* Head of Data Analysis Techniques Unit, TURKSTAT

† Expert, Data Analysis Techniques Unit, TURKSTAT

‡ Expert, Data Analysis Techniques Unit, TURKSTAT

## 2.2. Population projection and administrative division

The second major revision is the use of new population projection estimates and recent amendment of new administrative division. According to the new concept based on recent amendment regarding administrative division, urban population is increased. So, this situation forced to recalculate the new population projection.

The difference between the new population projection and the old one has certain effects on all indicators related with HLFS. It was assumed that the difference has occurred spreading over time rather than immediately, since the projections provide true information at the time of the year projections were produced. In order to reflect this type of difference to the back, exponential approach was used for the back recalculation of the indicators.

In order to reflect this type of revision to the back, We start with current methodology to the linking of economic time series using interpolation (see: Fuente(2013)):

$$Z_t = X_t + D_t(\ln Y_T - \ln X_T)$$

$$D_t = \frac{t}{T}$$

$Z_t$ ; backcasted series with new definition for t'th period

$Y_T$ ; series with new definition for T'th period

$X_T$ ; series with old definition for T'th period

$T$ ; overlapping period

Here the term  $D_t$ , is used to reflect the difference of overlapping period to past by decreasing rate.

We extended current methodology by considering the seasonal effects as following:

$$Z_{t,s} = X_{t,s} + D_t(\ln Y_{T,s} - \ln X_{T,s})$$

$$D_t = \frac{(t-1)V + 1}{(T-1)V + 1}$$

$$s = 1, 2, \dots, 12$$

$$t = 1, 2, \dots, T$$

$Z_{t,s}$ ; backcasted series with new definition for t'th year and s'th month

$Y_T$ ; series with new definition for coincidence year

$X_T$ ; series with new definition for coincidence year

$T$ ; overlapping year

$V$ ; rate of convergence

### 2.3. Other survey application changes

With the new survey, the other methodological changes are incorporated. These can be classified as sampling design, changes in reference period, survey organization structure. These types of changes are supposed to lead to breaks in the trends of series examined. The level shift caused by the survey application change is estimated. For this, the series with old and new definitions are multiplicatively decomposed into their unobserved components with maximum observation number as possible as shown below in expressions (1) and (2);

$$Y_t^1 = TC_t^1 x S_t^1 x I_t^1; \quad t = T_1, \dots, T_M, \dots, T_N \quad (1)$$

$$Y_t^2 = TC_t^2 x S_t^2 x I_t^2; \quad t = T_{-K}, \dots, T_1, \dots, T_M \quad (2)$$

Expression (1) refers to the decomposition structure of the series with new definitions as Expression (2) refers to the decomposition structure of the series with old definitions.

$Y_t^1$ ; Series with new definition,

$TC_t^1$ ; Trend-cycle component of the series with new definition,

$S_t^1$ ; Seasonal component of the series with new definition,

$I_t^1$ ; Irregular component of the series with new definition.

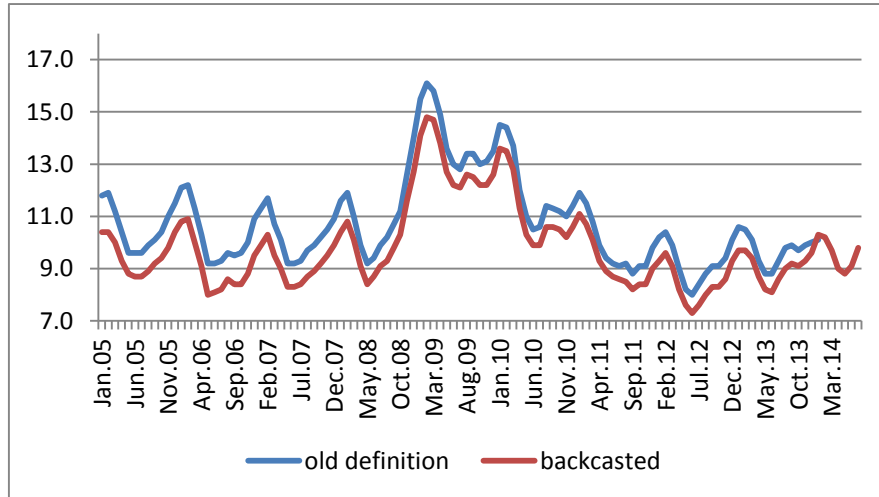
The Linking-Factor (LF) calculated by proportioning the means of trend-cycle components for matching periods (between  $T_1$  and  $T_M$ ) of series with both old and new definitions, is obtained as stated in the Equation (3).

$$LF = \frac{\overline{TC_t^2}}{\overline{TC_t^1}}; \quad t = T_1, \dots, T_M \quad (3)$$

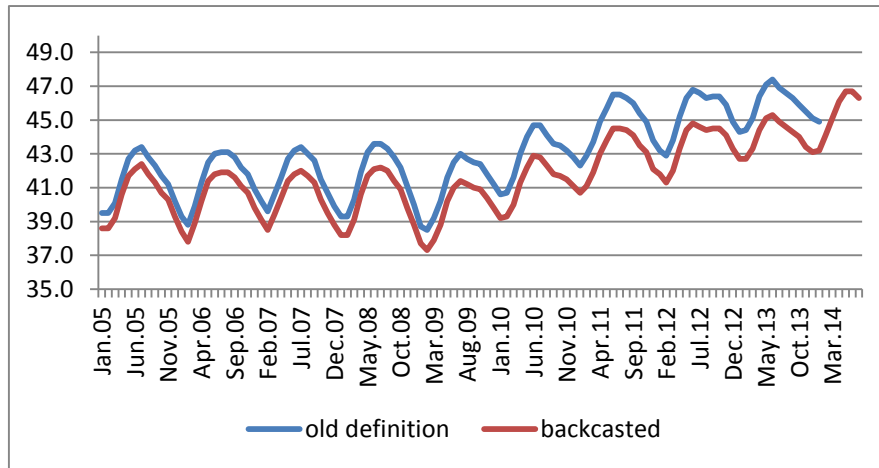
Then, the level of the series with old definition is adjusted to the level of the series with old definitions by using Expression (4). The term  $Y_t^{2*}$  in Expression (4) refers to the series, the level of which is adjusted to  $Y_t^1$ .

$$Y_t^{2*} = \frac{Y_t^2}{LF}; \quad t = T_{-K}, \dots, T_1 \quad (4)$$

### 3. RESULTS



**Figure 1.** Unemployment rate



**Figure 2.** Employment rate

Backcasted series and series with old definition for unemployment and employment rate are displayed in Figure 1 and Figure 2. The mean difference between the old definition and new definition is about 1 for unemployment rate series and 1.5 for employment rate series.

### 4. CONCLUSIONS

This paper proposed a couple of methods to meet the sustainability of labor force indicators in Turkey. The application contains the recalculation of the indicators covered time period January 2005 - January 2014 and the linking with the new indicators for 2014 and so on. The main contribution of this paper can be pointed out that different recalculation methods can be used to handle different type of revisions in an economic indicator.

### REFERENCES

- [1] A. de la Fuente, A mixed splicing procedure for economic time series, BBVA Working Papers 13/02, 2013.

# Backcalculating MIP (Macroeconomic Imbalances Procedure) indicators to improve data coverage: an empirical approach

Rosa Ruggeri Cannata ([rosa.ruggeri-cannata@ec.europa.eu](mailto:rosa.ruggeri-cannata@ec.europa.eu))<sup>1</sup>, Ferdinando Biscosi<sup>1</sup>,

Dario Buono<sup>2</sup>

**Keywords:** Indicators, time series, backcalculation, estimation

## 1. INTRODUCTION

The Macroeconomic Imbalances Procedure (MIP) is a system for monitoring economic developments and detecting potential harms to the proper functioning of the European Union (EU) and euro area economies in the form of internal and external imbalances, falling competitiveness, real estate bubbles or banking crises. It is part of a surveillance system for budgetary and economic policies, implemented via the European Semester, the EU's policy-making calendar.

Policy makers need an as complete as possible picture of the economy; to fulfill such requirement it might be necessary to apply statistical methodologies to ensure optimal data coverage, in particular in a context where statistics are subject to many events which could disrupt time series length, such as the adoption of new classifications or regulations, the availability of new primary data, a change in the production process, etc.

In order to fulfil the ten years timeframe required for MIP indicators and considering the definitions of some of them (e.g. averages or percentage changes over several years), needed data coverage can reach up to twenty years. In this context, statisticians have to ensure the availability of the required length for time series needed for MIP indicators and, where necessary, to apply statistical techniques, such as backcalculation, for this purpose.

In the last years, Eurostat tried to improve MIP data coverage, giving priority to information and sources coming from Member States. When those were not available, Eurostat proposed to concerned Member States an alternative solution based on modelling techniques whose results, if accepted, were then published as Eurostat estimates for MIP purposes.

This paper is structured as follows: section 2 briefly illustrates the methodological approach used for a backcalculation exercise in general; section 3 introduces two applications, the first one focuses on the House Price Index indicator and is a backcalculation exercise, while the second one covers the Unemployment rate and can be considered more as a disaggregation exercise; section 4 concludes.

## 2. METHODS

Backcalculation is the statistical process allowing to project back in time values of a given time series by using all relevant available information. It can be based on

---

<sup>1</sup> Eurostat, Macroeconomic Imbalances Procedure Task Force

<sup>2</sup> Eurostat, Unit B1, Methodology and corporate architecture

transparent and robust estimation techniques, such as retropolation, using static or dynamic modelling. When performing a backcalculation exercise, several aspects have to be analysed in order to define all issues related to the estimation of past values of a certain time series, in particular the following aspects will be of interest:

- The targeted horizon back in time
- The choice of the model
- The set of criteria used to validate obtained results.

The horizon will be a function of different factors: available information, historical context and availability of primary data in the past, and of existing data at national level. MIP scoreboard indicators should always be available for the analysed period, the ten years plus additional values required by data transformation, in order to look at the economic situation of Member States on the base of the same data coverage.

Concerning the choice of the model, for the sake of transparency, it is important for the model to be simple, robust, and parsimonious in the number of parameters. This would imply a high degree of automation in order to be able to run a backcalculation exercise whenever necessary and possibly with a very limited need for subjective judgement. Moreover, being MIP indicators annual, quarterly available information for the same indicator should be taken into account too in particular when directly related to the target variable, as can be the case of consolidated and non-consolidated data in the financial accounts. Several options are possible, varying from simply projecting back in time the growth rate of the series under consideration to the use of structural models with local trend or with seasonal component. Moreover, the backcalculation could be performed directly on the time series under consideration or after an opportune data transformation and using a univariate or a multivariate method.

Results of the backcalculation exercise need to be assessed on the base of clearly measurable criteria; in the context of the MIP indicators, it is important to get the involvement of national data producers in the validation process. For this reason when Eurostat backcalculates certain MIP indicators, the relevant Member State is informed via bilateral contacts in order to share all available information and to reach a common agreement on the extended time series.

### **3. RESULTS**

The MIP scoreboard uses annual data, however in some domains annual data are derived from monthly or quarterly ones; this is in particular the case of House Price Indexes which are collected and released on a quarterly basis, and of unemployment rates, which are mainly collected at quarterly frequency with final annual averages published along with the fourth quarter. As a consequence, backcalculation on those higher frequency data will have as a by-product the availability of past values for yearly MIP indicators.

#### **3.1. The House price index**

One of the headline indicators in the MIP is the House Price Index (HPI); since the beginning of 2013, with the entering into force of the reference legislation for this indicator, the monitoring of changes in house prices is based on data regularly compiled by Member States and transmitted to Eurostat.



However, HPIs have shorter historical coverage, for some Member States the series only start in 2009 and the regulation requirements do not cover preceding years.

The source data considered in this exercise to build up the regression variables were selected by the “Quarterly House Price Indices long time series Joint Residential Property Prices Indices Study Group” and then discussed at the Price Statistics Working Group in October 2013. The Study Group involved several international organisations and the following datasets were taken into account:

1. Eurostat HPI quarterly series
2. European Central Bank HPI quarterly series
3. Bank for International Settlements HPI quarterly series
4. National Central Bank quarterly series
5. OECD quarterly series

This exercise aimed at backcalculating the Eurostat (2010 = 100) series for the missing period by linking it with the available proxy variable provided by the above mentioned institutions. The correlation index, calculated over the common time span, was the key indicator for deciding which series had to be used as a proxy variable. Once the proxy has been chosen, both indexes are log transformed and first order differentiated. This transformation, i.e. delta log, can be interpreted as growth rate. OLS regression is run on the delta logs (given the loss of one observation at the beginning of the series) on the common time span and resulting parameters are used to back cast the Eurostat series. It is worth to notice that input data had a quarterly frequency and the overlapping period was covering twenty values at least. The arithmetic average of the HPI quarterly (backcasted and available) gives the respective annual value. The exercise was carried out for Member States where the proxy data available displayed a significant correlation level. These countries are Malta, for which the results were completed and published in 2013; Spain and Latvia whose backcalculated data were presented at Price Statistics Working Group of October 2014 and then used in the 2014 MIP exercise; and lastly Lithuania published in November 2014.

### **3.2. Croatian Unemployment data**

Croatia joined the EU on the 1<sup>st</sup> July 2013. Unemployment yearly data are derived from the monthly figures, which for Croatia are produced by Eurostat starting from the quarterly Labour Force Survey (LFS) and the monthly number of unemployed persons registered at the public unemployment office. From 2000 until 2006, the Croatian LFS was conducted twice a year, while in 2007 a continuous quarterly survey was started; Eurostat launched an empirical exercise with the aim to produce comparable figures for the period back to 2000 on the basis of the existing LFS data.

The exercise consisted of:

- outlining a pattern for the quarterly LFS, including seasonality
- estimate the values of monthly series for periods not available from the LFS

This exercise can be considered as a disaggregation from quarterly to monthly data; the adopted methodological approach was the proportional Denton procedure, which ensures that the resulting monthly figures on average equal the corresponding quarterly data. This method was applied at a disaggregated level, that is on the eight primary series for employment and unemployment levels, each broken down by sex and the two age groups, young (15-24 years) and adults (25-74 years). Aggregates (e.g. total employment

or total unemployment) were obtained by summing up the disaggregated series. The method was applied to the disaggregated series via the following steps:

- 1) For all years, half-yearly observations were allocated to the 2<sup>nd</sup> and 4<sup>th</sup> quarter respectively
- 2) Missing values for quarters 1 and 3 were estimated using a SARIMA model
- 3) Quarterly series were temporally disaggregated into monthly ones using the Denton\_modified method

#### **4. CONCLUSIONS**

This paper has illustrated the main issues related to the backcalculation of MIP indicators; the aim of the exercise is to enlarge data coverage in order to fulfil the needs of policy makers and other stakeholders. The paper includes also two case studies: the first one is a typical backcalculation exercise while the second one focuses on a disaggregation approach.

The backcalculation of time series for MIP indicators aims to ensuring required long series of data needed for the assessment of emerging or persistent macroeconomic imbalances in a country. It can become a systematic activity because statistics are subject to many events which could disrupt time series length. It is then particularly relevant to be able to manage this recurrent task; the adopted methodology for backcalculation uses a simple and robust approach, easily replicable and well documented.

Data used as input in this exercise are also subject to revisions, so that the model could be re-estimated giving new results. Past data however, should not depend too much on present data, especially for far away periods. When input data are revised, we can distinguish between major<sup>3</sup> and routine revisions. A major revision would require a through reconsideration of input datasets and of the models, with an assessment of the impact. When dealing with routine revisions, the stability of backdata should be favoured.

---

<sup>3</sup> For a definition of major and routine revision please see the [ESS guidelines on revision policy for PEEIs](#)

# **Use of register data for prior waves of EU-SILC in Austria: the case of “back-calculation”**

Thomas Glaser / Richard Heuberger

**Keywords:** EU-SILC, register data, income measurement, EUROPE 2020 indicators

## **1. INTRODUCTION**

In this paper methods and results of the integration of income register data for EU-SILC in Austria are presented. From EU-SILC 2012 onwards register data have been used on the basis of a national regulation. This inevitably resulted in a break in time series. In order to guarantee an unbroken time series in the course of the measurement of the Europa 2020 targets a revision of the EU-SILC results from 2008 onwards based on income register data was carried out.

The paper and the presentation will focus on the reasons for the back-calculation of EU-SILC data of 2008 – 2011, explain the main consequences of register data use on income variables as well as indicators and discuss further methodological issues that are of importance in the context of register data use.

### **1.1. EU-SILC in Austria**

EU-SILC in Austria is the national implementation of the EU-SILC survey and is conducted since 2004 as a rotational panel. The rotational design enables to follow households up to four years, once the panel component is fully established. Main aim of the survey is to provide data on income and living conditions of people in private households in a comparative manner. Since EU-SILC 2012 register data are used for the calculation of income variables.

### **1.2. Reasons for using register data**

Three reasons can be distinguished for using register data to calculate the income target variables. First, having the opportunity to gain almost all information on income from registers reduces the number of income related questions in EU-SILC to a minimum and therefore reduces response burden [1]. This is especially relevant for questions about income since these are very sensitive and can also be cognitively demanding when referring to the previous year. The second reason for using income registers lies in their objectivity [2]. Usually questionnaires about income are prone to socially desirable answers and suffer from imprecise answers caused by memory gaps of the respondents. Registers give full information on all incomes linked to a respondent, even extreme ones that are not desired to be reported or small and irregular income components that are easily forgotten. The third aspect in favour of register data is related to an easier data editing process. For most income register data all relevant information for gross and net values (if applicable) are included in the register and can be easily applied to the calculation of income target variables.

## **2. METHODS**

The usage of income register data influences the entire data collection, data editing and weighting procedure of EU-SILC. The following subsections give an overview of the change in methodology and subsequently in the changes of methods applied.

### **2.1. Record Linkage**

For the usage of register data linking survey and register data is essential since legal, practical and methodological limitations may affect this linkage. In Austria, a branch-specific personal identifier bPK (“bereichsspezifisches Personenkennzeichen”), allows for a practical, secure and “pseudonymised” linkage between register and survey data [3]. In principal this PIN is available for (almost) every person living in Austria, for all registers and therefore also in the sampling frame of EU-SILC. In practice not all persons can be assigned a PIN based on the gross sample, mainly because also persons not registered at the selected addresses as their main residence are included in the net sample. For these persons the PIN may be retrieved from the Ministry of the Interior. Assigning a PIN to all persons in the sample is only the first step in linking the EU-SILC sample with register data. The second one is obtaining data from all relevant income registers in due time. In most of the registers used about 5% of the records did not have a PIN, whereas for more than 96% of the persons in the surveys of EU-SILC 2012 onwards a PIN could be assigned.

### **2.2. Income Concept of EU-SILC in Austria**

Income in EU-SILC is collected in income target variables. These variables are defined by EUROSTAT and follow the structure of the ESSPROS classification (as far as pensions and benefits are concerned) and the general recommendations of the Canberra-Handbook. The household income, then, is the income of all household members within the income reference year, which is the year prior to the survey (for EU-SILC 2012 the year 2011). The following table provides an overview on the income components (income target variables) considered in EU-SILC.

## Model of the Household income of EU-SILC in Austria

Income on personal level		
	PY010	income from employment
+	PY050	income from self-employment
+	PY100	old-age benefits
+	PY090	unemployment benefits
+	PY110	survivor' benefits
+	PY120	sickness benefits
+	PY130	disability benefits
+	PY140	education related allowances
+	PY080	pensions from individual private plans
=	Sum of incomes on personal level	
Income on household level		
+	HY040	income from rental of property or land
+	HY050	family/child-related benefits
+	HY060	social exclusion benefits not elsewhere classified
+	HY070	housing allowances
+	HY080	regular inter-household cash transfers received
+	HY090	Interests, dividends, ...
+	HY110	income received by people aged under 16
=	Sum of incomes on household level	
Deductions		
-	hy130	regular inter-household transfers paid
-	hy145	repayments/receipts for tax adjustment
=	household income	

	Income information (mainly) from income register
	Income information (mainly) from income register from 2012 onwards
	Income information from survey

The table also provides the information for which income components register data for the calculation of incomes could be used. In EU-SILC 2012 the “full use” for the data production was implemented. For the years 2008 – 2011 this full implementation was not possible in every respect: either some register data were not “linkable” on a legal basis (since the national regulation allowing the link between survey and register data was only implemented in 2010) or some register data were technically not available for this record linkage procedure. However, the differences to the full implementation of EU-SILC 2012 are expected to be small.

### 2.3. Back-Calculation

Main aim for the back-calculation was the provision of a continuous time series for 2008 to 2011 and the provision of adequate micro-data for longitudinal analysis from 2008 onwards. Hence, all possible register data that were available (considering the limitations mentioned above) for EU-SILC 2008 – 2011 (income data for 2007 – 2010) were linked to the survey data of the respective years and the data editing process – as far as the income processing of the register income was concerned – was done anew for 2009 to 2011. In parallel, the weighting process was also repeated using information from register data.

### 2.4. Weighting

The last step of the EU-SILC weighting procedure consists of calibrating the household weights on known marginal distributions from reliable sources. Among these are the Austrian Microcensus and also register data from social security organisations. With the

usage of income register data additional marginal distributions from the wage tax register have been added. More importantly, the existence of register data in the EU-SILC data allows for calibrating register information in the sample on marginal distributions in the population based on the same source.

### **3. RESULTS**

The composition of register data as well as the adapted weighting procedure has a recognizable effect on the results, especially on income based indicators. Non-income based indicators are also affected by the change in weighting, but to a smaller extent. In order to evaluate the effect of using register data on the results income based on both sources has to be compared. From EU-SILC 2012 onwards income data are largely based on registers. Only the years 2008-2011 provide the opportunity to compare household income based primarily on the questionnaire or on registers. Among these EU-SILC 2011 is the most recent one and therefore also has the highest quality in terms of data linkage.

The results for most of the income target variables – the single income components as well as the total (equivalised) household income – showed a significant difference compared to the incomes derived from the questionnaires: incomes are more unequally distributed (and therefore a more complete coverage at the bottom and the top end of the income distribution), and for single income components a higher number of income recipients. Hence, register data use resulted in better coverage of the income distribution. Although, the median of the equivalised household income was not changed dramatically, the more unequal distribution led to higher at risk of poverty rate for all waves of EU-SILC from 2008 – 2011. Additionally, the use of register data for the weighting procedure also changed the outcomes for the non-income indicators of the EU-2020 indicators (severe material deprivation and jobless households).

### **4. CONCLUSIONS**

Using register data for income components had a significant impact on results of EU-SILC in Austria. The break in time series occurring in 2012 could be shifted back to 2008. Therefore, the observation period for the measurement of the Europe 2020 targets is not affected by the change in methodology.

Given that the back-calculation resulted in a continuous time-series for (up until now) EU-SILC 2008 – 2013, the exercise can be assessed as successful. Data checks have shown that the effects of register data use – analysed repeatedly during the process of introducing the use of register data for the data production of EU-SILC in Austria – are in principal consistent for the whole time series (i.e. the years for which income information based on questionnaire and register information exists in parallel).

In a technical perspective, further analysis can draw the attention in more detail to the record linkage, the weighting procedure, the effects of register data use on longitudinal analysis and other related topics. In a more political perspective, the effects of register data use for the time series of EU-SILC based indicators –for example relevant for the EU 2020 targets – are of importance.

### **REFERENCES**

- [1] Wallgren, A. and Wallgren, B. (2007), Register-based Statistics. Administrative Data for Statistical Purposes, Chichester, Wiley.

- [2] Rendtel, U. and Nordberg, L. and Jäntti, M. and Hanisch, M. and Basic, E. (2004), 'Report on quality of income data', Chintex Working Paper 21.
- [3] Hackl, P. (2009), Using Administrative Data at Statistics Austria: Legal Provisions. Paper presented at the 95<sup>th</sup> DGINS Conference, 1<sup>st</sup> October, Malta, MT

# Pitfalls of regression modelling with complex survey data

Florian Ertz ([ertz@uni-trier.de](mailto:ertz@uni-trier.de))<sup>1</sup>, Ralf Thomas Münnich ([muennich@uni-trier.de](mailto:muennich@uni-trier.de))<sup>1</sup>

**Keywords:** Complex surveys, HFCS, variance estimation, weighting

## 1. INTRODUCTION

The European Central Bank (ECB) recently initiated the Eurosystem Household Finance and Consumption Survey (HFCS) (see [1]). It collects a comprehensive set of micro data on private households' demographics, wealth, debt, and consumption in an ex ante-harmonised way in the Euro area member states. As the financial crisis once again highlighted, such detailed information (on specific subgroups) are relevant for economic and monetary policy-making. Multiply-imputed data for the first wave of the HFCS (covering 15 Euro area member states) were disseminated to researchers in 2013.

Micro data collected in wealth surveys like the HFCS are most prominently used in the field of household finance, where the interaction between private households and financial markets is investigated. Economic papers on the subject tend to ignore the nature of the data collection process, e.g. they treat the data as if being collected using simple random sampling (SRS). In light of the fact that surveys like the HFCS use complex survey designs (including multi-stage stratified cluster sampling), this might lead to incorrect variance estimation for regression model parameters causing erroneous inference. This problem may well be aggravated by special survey designs aimed at sampling a disproportionately high share of wealthy households (who are prone to non-response but of major interest for the data producer). Such oversampling was implemented in 9 out of the 15 country surveys of the HFCS.

Using the HFCS research data set, some typical household finance models and a simulation study, we illustrate the resulting problems and propose correction methods.

## 2. METHODS

Besides other things household finance tries to explain households' decision to participate in risky financial assets as well as the share of liquid wealth invested in such assets. Typically some forms of logit/probit and tobit models are estimated.

Due to confidentiality reasons the ECB does not provide any survey design information with the HFCS research data set. For the researcher the only possible way to incorporate design information into the model is to use the final household weights and (at least some of) the provided replicate weights which are based on a variant of the rescaling bootstrap of Rao and Wu (see [2] and [3]).

We estimate some typical household finance models (see, e.g., references in [4]). The estimation procedure is in line with the one used in the R package *survey* (see [5]). We compare the results reached when a) no weights are used, b) only final household weights are used, and c) sets of replicate weights are used. The differences between the used weighting schemes are illustrated by some summary measures. We document the (in)stability of regression parameter variance estimates when replicate weights are used.

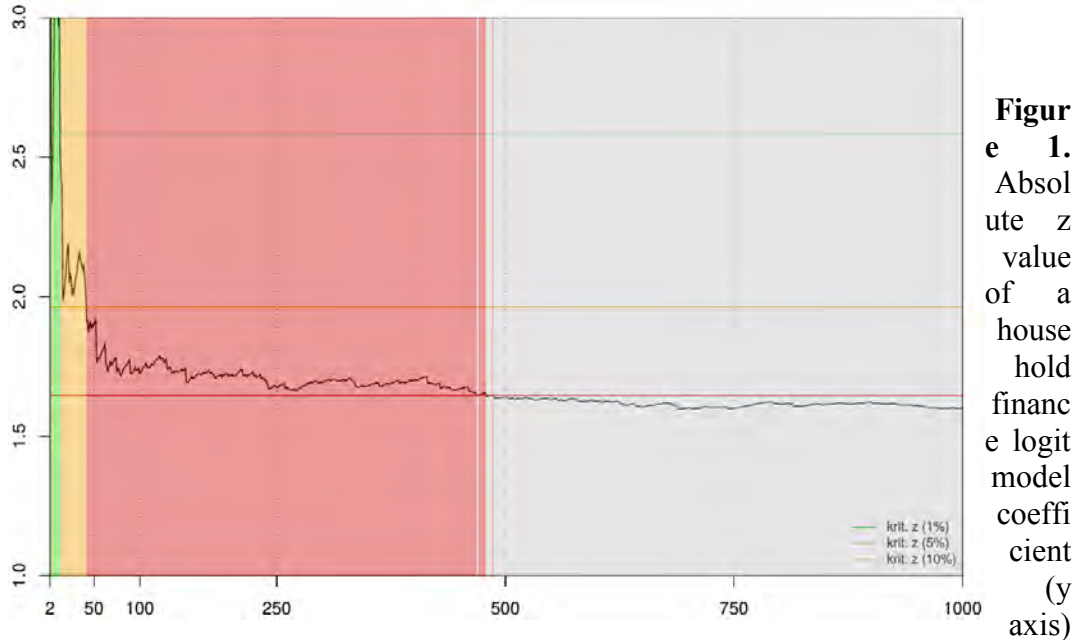
---

<sup>1</sup> Economic and Social Statistics Department, University of Trier.



Furthermore we use a synthetic household population in the context of a small simulation study to show the 'real' impact of ignoring design information on the estimation of basic household finance models. In the course of the study we analyse if the impact of different survey designs which try to mirror the designs used in the HFCS varies. Furthermore, we compare estimation procedures only using replicate weights with estimation procedures which directly take design information into account.

### 3. RESULTS



versus number of replicate weights used for model estimation (x axis)

For a given set of (country) data and a given model there is a considerable variation in the significance of regression model parameters when different numbers of replicate weights are used (see Figure 1). For some parameters each of the commonly used significance levels might be reached when an 'appropriate' set of replicate weights are used in the estimation process. Accordingly, there seems to be a lot of leeway to manipulate model results to the analyst's liking. It is likely that this problem will even be more severe when the data are collected using oversampling designs. We expect confirming results in our simulation study.

### 4. CONCLUSIONS

In light of the highly unstable variance estimates of many regression model parameters resulting from replication weight-based estimation, we suggest that the maximum number of replicate weights provided for each household (1,000) be used in empirical research. Of course this will come at the expense of computation time. This should insure against likely wrong conclusions reached when, e.g., only 100 replicate weights are used. It would be favourable if the ECB or national central banks would find a way to add some further design information to the HFCS research data set without compromising confidentiality requirements. Overall, researchers should be more cautious when using micro data gathered using complex surveys to estimate regression models.

### REFERENCES

- [1] European Central Bank, [https://www.ecb.europa.eu/home/html/researcher\\_hfcn.en.html](https://www.ecb.europa.eu/home/html/researcher_hfcn.en.html) .

- [2] J.N.K. Rao and C.F.J. Wu, Resampling Inference with Complex Survey Data, Journal of the American Statistical Association 83 (1988), 231-241.
- [3] J.N.K. Rao et al., Some Recent Work on Resampling Methods for Complex Surveys, Survey Methodology 18(2) (1992), 209-217.
- [4] L. Guiso and P. Sodini, Household Finance: An Emerging Field, Handbook of the Economics of Finance, Vol. 2B, Amsterdam, 2013.
- [5] T. Lumley, Complex Surveys: A Guide to Analysis using R, New York 2010.

# A Suggested Framework for National Statistical Offices for Assessing and Managing Privacy Risks Related to the Use of Big Data

Prepared by the Task Team on Big Data Privacy, within the Big Data project overseen by the High-Level Group for the Modernisation of Statistical Production and Services<sup>1</sup>

**Keywords:** big data, privacy, security, confidentiality, disclosure risk, risk management, data access, reputation

## 1. Background

In April 2013, the United Nations Economic Commission for Europe (UNECE) Expert Group on the Management of Statistical Information Systems (MSIS) identified Big Data as a key challenge for official statistics, and called for the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to focus on the topic in its plans for future work [1]. As a consequence, the project *The Role of Big Data in the Modernisation of Statistical Production* was undertaken in 2014. The project comprised four ‘task teams’, addressing different aspects of Big Data issues relevant for official statistics: Privacy, Partnerships, Sandbox, and Quality.

The Privacy Task Team was asked to give an overview of existing tools for risk management in view of privacy issues, to describe how risk of identification relates to Big Data characteristics, and to draft recommendations for National Statistical Offices (NSOs) on the management of privacy risks related to the use of Big Data. The Privacy Task Team was comprised of representatives from several NSOs all over the world.

The Task Team concluded that extensions to existing frameworks were needed in order to deal with privacy risks related to the use of Big Data. This abstract summarises the outcome of the Big Data Privacy Task Team.

The proposal from the Privacy Task Team is going to be presented at the HLG meeting in November 2014. Development projects could be launched by the UNECE HLG to finalise the work undertaken by the Task Teams.

---

<sup>1</sup> The task team was composed by the following members with affiliations: Josep Domingo-Ferrer, Universitat Rovira i Virgili, Italy; Jörg Drechsler, Institute for Employment Research, Germany; Luis Gonzalez and Shaswat Sapkota, United Nations Statistical Division (UNSD); Pascal Jacques, Eurostat; Jingchen Hu (Monika), Duke University, USA; Matjaz Jug, United Nations Economic Commission for Europe (UNECE); Anna Nowicka, Central Statistical Office of Poland; Peter Struijs, Statistics Netherlands; Vicenc Torra, Artificial Intelligence Research Institute, Spanish Council for Scientific Research; Shane Weir (chair) and James Chipperfield, Australian Bureau of Statistics (ABS).

## **2. Existing tools for privacy risk management**

NSOs try to manage two conflicting aims: providing access to its data for the benefit of society and to meet society's expectation that sensitive information about data providers will be kept private. For traditional data sources, in particular sample surveys and administrative data collected by governments, the Task Team has considered what types of risk exist. For estimate releases, disclosure can occur as for micro-data releases, except that information about a data provider must first be derived from one or more estimates. An extensive body of literature on this is available [2]. The risk of disclosure is also influenced by the type of access.

As to the management of disclosure risk, two broad areas are (1) reducing the risk that an attempt at disclosure will be made and (2) releasing data in a manner that it is not likely to enable disclosure. For each of these areas, a number of factors can be identified that affect the risk of disclosure.

Within this framework, the Task Team has specifically looked into the different ways in which statistical agencies allow analysts or researchers to access micro-data, managing the risk of disclosure associated with databases, and the advantages and disadvantages of the different approaches to managing privacy. Relevant software was also identified.

As to micro-data access, NSOs have developed different strategies to enable external researchers to analyse their data without violating confidentiality regulations. These strategies fall under three broad topics: micro-data dissemination, onsite analysis in research data centres, and remote access.

Concerning risks related to databases, it is useful to make a distinction between owner privacy, respondent privacy and user privacy. Owner privacy refers to situations where entities make queries across their databases in such a way that only the results of the query are revealed. Respondent privacy is about preventing re-identification of respondents, e.g. individuals or organisations. User privacy is about guaranteeing the privacy of queries to interactive databases, which is necessary to prevent user profiling and re-identification.

As to the advantages and disadvantages of approaches to managing privacy, one has to bear in mind that when an agency applies a statistical disclosure control (SDC) to micro-data, or estimates derived from them, there is an implicit trade-off between the disclosure risk and the utility. Disclosure risk depends upon the context in which the attack is assumed to occur. While a wide range of attacks are possible, a pragmatic approach for an agency to take is to focus on the scenarios that are likely to have the greatest disclosure risk. The Task Team has looked into disclosure risk and utility for queryable databases, micro-data SDC, and tabular data.

## **3. Big Data characteristics and privacy risk**

Big Data is often characterised by three or four Vs: velocity, variety, veracity, and sometimes also value. However, there are more aspects that may be relevant to privacy risk. In particular size, availability, aggregation, awareness of society, flexibility, provider infrastructure and geographical differences may be pertinent. One example discussed is GPS location data.

The existing tools for privacy risk management can be assessed for their application to Big Data by taking these characteristics into account. For instance, of the three micro-data access options (i.e. micro-data dissemination, onsite analysis in research data centres, and remote access), micro-data dissemination is no longer an option in most cases, since one feature of Big Data is that the size of the data implies that transferring the data is cumbersome. It would not be possible to send the data to the researcher on CDs or provide a link for a download as is the current practice with micro-data dissemination. Thus, onsite and remote access will be the only viable solutions.

Onsite access has the advantage of providing the agency with better control over who accesses the data and what can be done with the data. However, if the standard set of surveys and administrative data that is already offered by the NSO will be enriched by new data sources it is likely that offering manual output checking for all these databases will no longer be feasible.

As to the release of estimates, the common practice of re-identification experiments will probably not be useful in the Big Data context. Disclosure risks are typically assessed by assuming that the intruder has information on some of the variables in the dataset from another data source and then tries to use this information to identify somebody in the confidential dataset. Under this scenario record linkage experiments are run to see how many records would be matched correctly. Given the fact that such experiments usually have already long running times with survey data, their feasibility to Big Data is questionable. It is thus more relevant than ever that general strategies for ensuring the confidentiality of the generated outputs without manual interference are developed.

#### **4. Recommendations**

For a Big Data context, a number of guidelines for risk treatment can be given, some of which build on existing tools, whereas others are more novel. The recommendations given by the Task Team can be grouped into:

##### **(1) Information integration and governance:**

- a. Database activity monitoring, i.e., keeping track of who has access to your databases and what they are doing at any given time.
- b. Application of best practices for security of IT systems and business practices. As a baseline, four best practices should be applied: separation of duties, separation of concerns (a modular approach to functionality where possible), the principle of least privilege (no more access rights than needed), and defence in depth (multiple security mechanisms).
  - i. Separation of Duties;
  - ii. Separation of Concerns (a modular approach to functionality where possible);
  - iii. Principle of Least Privilege (no more access rights than needed); and
  - iv. Defence in Depth (multiple security mechanisms/layers).
- c. Application of best practices of security of transportation, such as Transport Layer Security (TLS).
- d. Data encryption. Examples are Full Disk Encryption (FDE) and File System-level Encryption (FSE).

- (2) Statistical disclosure limitation/control (identity disclosure, attribute disclosure, inferential disclosure and population/model disclosure):
- a. Preserving confidentiality by restricting data access and restricting data release.
  - b. Ensuring access to useful data. It is important to ensure that the released data, while not identifiable, fulfils a purpose and is still meaningful [3].
  - c. Balance data utility and disclosure risk. Apart from traditional approaches, modern techniques are proposed, such as the use of statistical models to simulate data records that emulate the main distributional characteristics of the original micro-data.
- (3) Managing potential risk to reputation (public image):
- a. Enforce ethical principles in the supply chain, including legal and administrative instruments of accountability and informed consent.
  - b. Establish strong compliance controls.
  - c. Develop a monitoring system to track reputational threats.
  - d. Ensure transparency and understanding through clear communication with stakeholders, for instance on the use of data on citizens, and organisation of a dialogue with the public.
  - e. Create a crisis communication plan.

## References

1. UNECE (2013a). Final project proposal: The Role of Big Data in the Modernisation of Statistical Production. UNECE, November 2013.  
<http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production>
2. Hundepool, A., Domingo-Ferrer, J. , Franconi, L. Giessing, S. Schulte Nordholt, E. Spicer, K., and de Wolf, P. P. (2012). Statistical Disclosure Control, Wiley.
3. Slavkovic, A. B. (2007). Overview of Statistical Disclosure Limitation: Statistical Models for Data Privacy, Confidentiality and Disclosure Limitation.

# Using Research Data Centres (RDCs) to access Big Data

David Schiller<sup>1</sup> ([david.schiller@iab.de](mailto:david.schiller@iab.de)) and Anja Burghardt

**Keywords:** Big Data, Data Access, Linkage, Confidentiality, Research Data Centre.

## 1. INTRODUCTION

Research Data Centres (RDCs) were made to support researchers by providing well documented research data. Documentations include information about sampling frames, quality of the information, and confidentiality issues. Research data are normally coming from surveys. The advantage of surveys lies in the more or less controlled environment. Sampling frames are known, questionnaires were build based on theories, and participants were asked to participate. At the same time surveys are cost and time intensive. Low case numbers often limit the possibilities to fit statistical models. One reaction on that issue was to make additional data sources available via RDCs. Administrative data derives from internal processes of institutions. The data collection process is most of the time not as good known as it is the case for survey data. On the other hand administrative data comes with huge case numbers and allows therefore analyses that were not possible by only using survey data. In the meantime survey and administrative data is used regularly to answer sophisticated research questions. Data documentation and data quality issues are solved and confidential issues are addressed by (among other things) secure data access ways. While RDCs have made survey and administrative data usable for research, the next data source has already appeared. Big Data offers a high potential to support research when used in a smart way. This text proposes and infrastructure to make Big Data usable for research.

First the RDC of the German Federal Employment Agency is introduced; afterwards a closer look at Big Data is taken, before an infrastructure to work with Big Data is outlined. The text closes with a conclusion.

## 2. THE RESEARCH DATA CENTRE (RDC) OF THE GERMAN FEDERAL EMPLOYMENT AGENCY (BA) AT THE INSTITUTE FOR EMPLOYMENT RESEARCH (IAB)

The RDC of BA at IAB was established in 2004 [1]. It was made to prepare and document data coming from the administrative processes of the German Federal Employment Agency for the scientific community. Those administrative data is linked to

---

<sup>1</sup> David Schiller and Anja Burghardt; (Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

survey data coming from the Institute for Employment Research. Both data sources are made available via secure data access ways [2]. The RDC of BA at IAB has therefore rich knowledge in the areas of data documentation, confidentiality, data linkage and data access. Coming from this background the RDC of BA at IAB is now looking at ways to make Big Data usable for research in order to enrich the already provided research data.

### **3. A CLOSER LOOK AT BIG DATA**

The term Big Data is not clearly defined. Sometimes it is used to refer to huge data files. This is a perspective that often comes from social scientists. Computer sciences have other kinds of data in mind when talking about Big Data. They refer to flow and process generated data. It is therefore of worth to have a closer look at the special character of Big Data.

One difference between survey and Big Data lies in the distinction between designed (survey) and organic data (Big Data) [3]. While survey data was created in a structured way to support scientific research, Big Data just grows for different kinds of reasons and the scientific community has to find out how to use those organic data for their research. Another characteristic of Big Data is that it is too big to be moved and too big to be used by the (in the social sciences) common statistical tools. Moving Big Data is nonsense because of network issues – it has to be kept in mind, that Big Data is sometimes not stored; it is a flow of data from which information can be extracted. Common tools for statistical analyses in the social science are not made to deal with such huge volumes of data nor are they made to deal with flow data. Another reason for not moving Big Data lies in the reason that Big Data may be confidential or under intellectual property and therefore protected by data security measures (which is also true for many standard social science data sets). In addition Big Data can be found as restricted or Open (Big) Data. Ownership is not always clear. While rich resources of Big Data are owned by the public sector the even more interesting ones are owned by private companies.

A recently published book “Privacy, Big Data, and the Public Good” [4] collects texts from experts from the social science that deal with different issues that arise when making Big Data usable for social science research. Kreuter/Peng suggest the combination of survey data with Big Data in order to be able to extract meaningful information out of Big Data [5]. Stodden introduces the idea of „Wallet Gardens“ to secure data (disclosure risk, business value) and syntax (intellectual property). The ideal scenario would be as followed: data is held by a trusted or trustworthy curator; the data



remain secret, the responses are published [6]. Karr and Reiter deal with the challenge of giving access to (confidential) Big Data. They propose an integrated system including: unrestricted access to synthetic data; a verification server that allows users to assess the quality of their inferences; and approved researchers to access the confidential data via remote access solutions [7]. Dwork finally outlines that cryptography will help to be close to the ideal scenario mentioned by Stodden [8].

The special characteristics of Big Data, the mentioned approaches to make Big Data usable and many more aspects have to be included in an infrastructure that enables the use of Big Data for research. Thereby the best mixture has to be found and the experiences and the knowledge of RDCs have to be used when developing such an infrastructure.

#### 4. AN INFRASTRUCTURE TO ACCESS BIG DATA

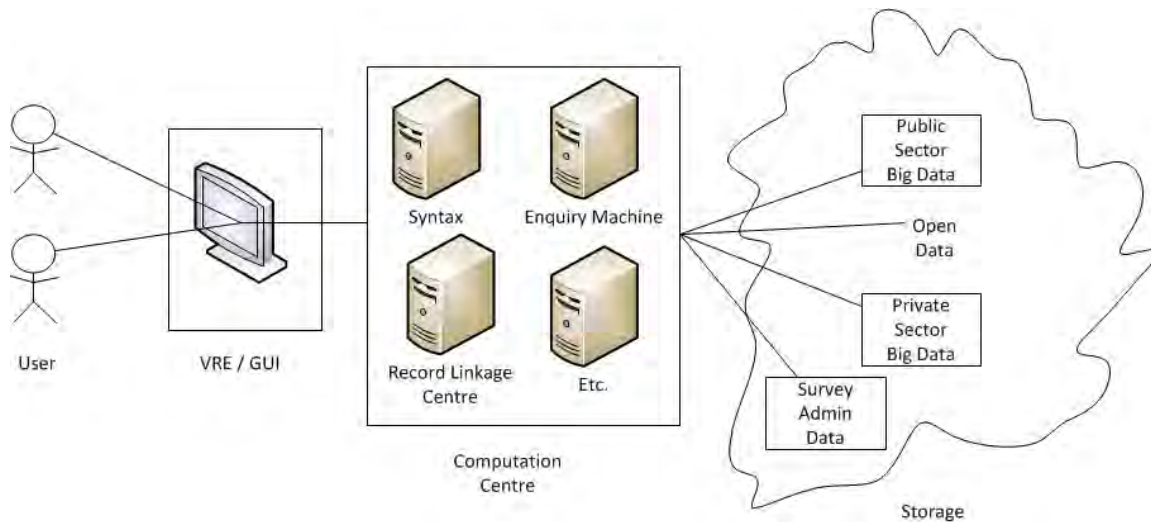


Figure 1. Infrastructure to access Big Data

Figure 1 shows an overview of an infrastructure that can enable the use of Big Data. User access the infrastructure from given access devices. The security level for access control depends on the confidentiality of the accesses information. An virtual research environment (VRE) works as graphical user interface (GUI) for the user. It is the working environment that is used to navigate through the offered infrastructure. The computation centre provides applications, and storage for different services. Only three should be mentioned; more can be added. Program syntax can be generated and securely stored. An enquiry machine searches for the needed data. A record linkage centre will support the linkage of different data sources. Via the computation centre different kinds of data sources can be accessed: namely survey/admin data, public sector Big Data, private sector Big Data and open data. It is important that the VRE, the computation

centre as well as public sector Big Data and Survey/Admin data can be hosted by the same RDC. Such an infrastructure would solve most of the challenges when working with Big Data. The main components are: the VRE as working environment for the users; the computation centre as knowledge, trust and service centre; and the storage area that can handle distributed data sources. Modern data access and data protection techniques have to be used.

## 5. CONCLUSIONS

It is still not clear what Big Data exactly is and what benefits social science research can extract out of it. At the same time it is clear that Big Data will reshape the way in which knowledge is generated in a modern world. Researchers have to make use of those new data sources; they have to create knowledge about Big Data; generate theories and methods in order to be able to work with Big Data in a useful way. Only if this happens, high quality scientific research can be done with Big Data and also this source of data can be used to answer the research questions that arise in a modern society. RDCs have to care about the needed infrastructures to make this possible.

## REFERENCES

- [1] J. Heining, The Research Data Centre of the German Federal Employment Agency: Data Supply and Demand between 2004 and 2009, RatSWD working paper, 129, (2009), 22p.
- [2] S. Bender and J. Heining, The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing, IASSIST Quarterly (IQ) Vol. 35 No. 3, (2011), 10-16.
- [3] R. Groves, "Designed Data" and "Organic Data," Director's Blog, <http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/> (accessed January 20, 2014).
- [4] J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Privacy, Big Data, and the Public Good: Frameworks for Engagement (2014).
- [5] F. Kreuter, R. Peng, Extracting Information from Big Data: Issues of Measurement, Inference, and Linkage, in: J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Privacy, Big Data, and the Public Good: Frameworks for Engagement (2014).
- [6] V. Stodden, Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency, in: J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Privacy, Big Data, and the Public Good: Frameworks for Engagement (2014).
- [7] A. Karr, J. Reiter, Using Statistics to Protect Privacy, in: J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Privacy, Big Data, and the Public Good: Frameworks for Engagement (2014).

- [8] C. Dwork, Differential Privacy: A Cryptographic Approach to Private Data Analyses, in: J. Lane, V. Stodden, S. Bender, H. Nissenbaum, Privacy, Big Data, and the Public Good: Frameworks for Engagement (2014).

# CASD-TERALAB, A SECURE REMOTE ACCESS SYSTEM TO CONFIDENTIAL BIG DATA: DESCRIPTION, DEMONSTRATION AND USE CASES

Alexandre Marty (alexandre.marty@casd.eu)<sup>1</sup>, Frank Cotton<sup>2</sup>, Kamel Gadouche<sup>1</sup>, Nawrès Guédria<sup>1</sup>

**Keywords:** Big Data, data science, remote access, security, privacy

## 1. INTRODUCTION

TeraLab is the name of a new Big Data platform made available in France a few months ago. It is built and operated by a consortium selected through a call for projects conducted by the Caisse des Dépôts, whose aim was to encourage industry players to develop and integrate technology and know-how for the emergence of a Big Data sector in France, and to foster the creation of innovative services exploiting these technologies.

Following this objective, the TeraLab project mainly aims at supporting R & D on these technologies. The project is scheduled for a five-year period, with an overall budget of €5.7 M.

The complete TeraLab Platform is composed of two compartments, and two organizations form the TeraLab consortium:

- IMT (Institut Mines-Télécom), a public institution dedicated to research and innovation in the engineering and digital fields, made up of ten top-ranking French higher education establishments. **IMT is in charge of the public compartment available from a web portal focusing on standard sets of bigdata datafiles.**
- GENES (Group of National Schools for Economics and Statistics, a national statistical authority), through its Secure Remote Access Center (CASD). The CASD has set up a very secure dedicated solution for remote access to confidential data by researchers. It provides a complete environment with secure data files and software in a closed, leak-proof, IT infrastructure with a strong authentication procedure based on at least two factors including biometrics. **GENES is in charge of the highly secured compartment called CASD-TeraLab, hosted in the secure area of the French RDC and mainly dedicated to confidential data.**

INSEE, the French national statistical institute, has formed a partnership with the GENES for this operation and thus also participates in the project.

This paper mainly focuses on the CASD-TeraLab platform. It gives a rapid overview of the CASD-TeraLab architecture, and presents some examples of use cases that it is designed to support.

## 2. PLATFORM DESCRIPTION

### *Platform architecture*

The objective of the platform is to facilitate the batch or real-time exploitation of large data sets, including confidential data, and to allow the integration of new data analytics algorithms

---

<sup>1</sup> GENES

<sup>2</sup> INSEE

developed by research laboratories or innovative start-ups. The platform architecture reflects this objective.

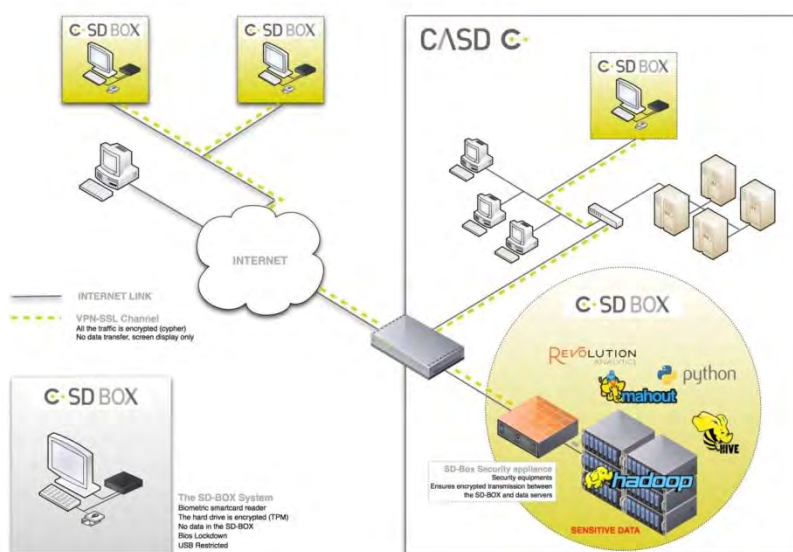
### *Technical infrastructure*

The platform is composed of several isolated clusters of Hadoop-type commodity servers (500 terabytes of physical storage). These clusters are included in the "bubble", a highly secure environment managed by the CASD. Access to those resources is possible through a remote access system connected to virtual client machines equipped with the necessary middleware, client software and configuration.

The remote access system requires the use of a specific remote hardware device (see image below). This system allows access from other countries.



This ultra-secure “bubble” is an important asset of the platform, especially in regard of the growing concerns for IT Security and data privacy.



### *Services and data*

A large catalog of software packages is available on the platform. All the common tools from the Hadoop world are present, such as MapReduce and Apache Pig for batch processing, Apache Hive and Impala for performing SQL queries on big data, or the more recent Apache Spark with MLlib for in-memory processing and machine learning. Connectors for the popular R and Python languages are provided, as well as data exploration and visualization tools such as Dataiku. Latest-generation tools are also used for the management and supervision of the platform.

TeraLab will progressively host a large catalog of data: open data like Common Crawl, OpenStreetMap... but also data that will be acquired with a specific budget from sources like Twitter, Facebook or Dun & Bradstreet. Furthermore, the data already proposed by the CASD will be available under the relevant legal procedures.

### 3. RESULTS

#### 3.1. Official statistics

Leveraging its involvement in the consortium and its mission of coordination of the French public statistical system, INSEE has proposed several projects for the TeraLab platform. Two data scientists have been recruited to help implement the selected projects.

More proposals should follow, for example along the following lines:

- Research of new statistical methods (models, algorithms, estimators...) adapted to the Big Data paradigm, or test of how existing ones perform in this case.
- Exploitation of new sources of data, or more efficient utilization of existing ones, in different domains like health, transports, social indicators, etc.
- Exploration of techniques enabling a combined use of Big Data and reference data.

As an example, for a few years INSEE has conducted experiments on the use of scanner data in the computation of the consumer price index. This has led to the identification of questions or problems of methodological or operational nature. The availability of the CASD-TeraLab platform offers totally new perspectives for dealing with these problems. The platform's key functionality will be to provide real time Consumer Price Index quotes computed based on exhaustive data instead of working only on a few data samples.

First experiences on scanner data, made for the moment on dummy data, are very positive and promising.

	Hadoop	RDBMS
Loading one week of data	10 min	1h30
Running time of an example SQL query	40 s	1h15

In light of those results, the efficiency of CASD-TeraLab platform seems to be obvious. For example, loading one week of data is about 10 times faster when using the Hadoop framework than with a traditional architecture.

The results are impressive not only for loading and running SQL queries but also for further processing. For example, running a complex program which processes the exhaustive data, makes calculations on each row and writes to disks takes only 3 hours.

The data used so far is collected from the top 5 French retail chains. Those first results will allow to acquire more data from more companies from other fields (minimarkets, hardware shops, etc...). A Big Data platform such as TeraLab will be able to process such a large amount of data unlike traditional databases with classical tools.

#### 3.2. Health data

The French health data system is very rich, but also very complicated. At its core is a database keeping an almost exhaustive record of all the health services provided to every citizen, integrating medical information, clinical data, data generated by hospitals, health care, etc. Its volume (more than 1.2 billion records with more than a thousand variables, about 250 terabytes of data each year) and complexity (sophisticated data structure, detailed codes) place it clearly in the Big Data domain. Exploiting it with the tools currently available is difficult and burdensome.

It should also be noted that everyone in France has a unique personal identifier (the NIR). This offers opportunities for data processing (matching by design, longitudinal studies), but also generates heavy security restrictions, since the usage of the NIR is very strictly constrained by law.

Still, there are strong needs regarding the health data system. An example would be the need for real-time analysis that could detect epidemic outbursts or public health problems. But also prescribing analytics will be possible.

The CASD-Teralab will also **host clinical data such as medical imaging and genomics data**. Discussions with several cohort projects have just started.

The CASD-TeraLab platform can meet both the volume and the security challenges that are needed to deal with such sensitive data.

### 3.3. Public challenges

It is important to look beyond the data and the technical implementation: the Big Data revolution introduces a new discipline called data science, which lies at the crossroads of computer science, mathematics and statistics, and which will be supported by a new community of students, consultants and researchers in those fields: the data scientists.

The DataScience.net web site, like the Kaggle web site, aims at gathering people involved in new developments on methodology concerning data: public data, confidential data and big data. The main idea is to allow data owners to meet a large community of data scientists.

The website, developed through a partnership between the GENES and Bluestone, a private company specialized in data science, allows data owners to issue public challenges based on their data. They have to provide a set of data with the description of the data, of the challenge and of the results evaluation criteria. Data scientists analyze the data, submit models and their results, and get evaluation scores.

The competitor with the highest score wins a prize. The data owner, on the other hand, will get the model from the winner and will be able to use it for his needs.

Many companies or public institutions are interested in this new way of collaborating with the data science community. However, most of them need to propose sensitive and/or very large data files for analysis in a private contest. Access to confidential, potentially big data is a strategic and high-stakes issue today for both data owners and data scientists.

During the first phase of its development, datascience.net offers only public contests based on small sets of public data for downloading. The second phase in the project roadmap, in 2015, will focus on offering a platform for private challenges based on sensitive data and/or massive data.

The main idea is to operate a first selection of candidates based on their profiles and their previous participations. Once the selection is made, only 3 to 5 teams, depending on the challenge, will get access to the data via the TeraLab platform. In case of confidential data, the secure CASD bubble will be used.

## 4. CONCLUSIONS

TeraLab will be an important platform in the development of the French Big Data ecosystem. It starts bringing together industrials, data providers and academics in the construction of real data value chains. We expect to learn a lot during the project, and we will seek to exchange our experience with other actors at the national or European level. In particular, we will seek

coordination with the Big Data project that UNECE decided to launch in 2014 and with the Big Data task force that has been recently created by Eurostat.



# A gateway to European Research Services

Anja Burghardt ([anja.burghardt2@iab.de](mailto:anja.burghardt2@iab.de))<sup>1</sup>; David Schiller ([david.schiller@iab.de](mailto:david.schiller@iab.de))<sup>2</sup>

**Keywords:** Data Access, Virtual Research Environment, Network, Service Hub.

## 1. INTRODUCTION

Europe is struggling with societal challenges in a number of fields like health, migration, demographic change and others. For evidence based approaches and the development of policy solutions to address these challenges on a European level innovative pan-European research is crucial. The upcoming challenges are thereby not limited to one discipline and not to special European countries. Therefore interdisciplinary research on a European Level is necessary.

The “Data without Boundaries” (DwB) project funded by the FP7 Framework Program of the European Commission aims to improve the access to confidential microdata. One work package is proposing a European Remote Access Network (EuRAN) to address such issues [1].

This paper is focusing on the usability of a harmonized data access system in the European Research Area. Such a system would benefit both: first researcher to find the right data for their research topic and support them during their complete research lifecycle and secondly data providers by making their data more visible and by harmonizing data access infrastructures. To fulfill this goal a Single Point of Access (SPA) as a Gateway to European Research Services is a step forward to interdisciplinary research on a European level.

## 2. GENERAL DESCRIPTION

Figure 1 shows the general setting for a Gateway to European Research Services. The proposed network consists of a Single Point of Access (SPA) that is equipped with a Service Hub. The main goal is to connect researchers with European Research Services, like data documentation, application procedures, wikis and research data access. The SPA helps the researcher to find all the relevant services for his research while the Service Hub works as a kind of umbrella for the distributed services. The virtual research environment assists the researcher, and research teams, to work in an European research

---

<sup>1</sup> Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute of Employment Research (IAB)

<sup>2</sup> Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute of Employment Research (IAB)

environment. The following text will concentrate on the description of the SPA and the Service Hub. Thereby first the infrastructure perspective and afterwards the user perspective will be shown.

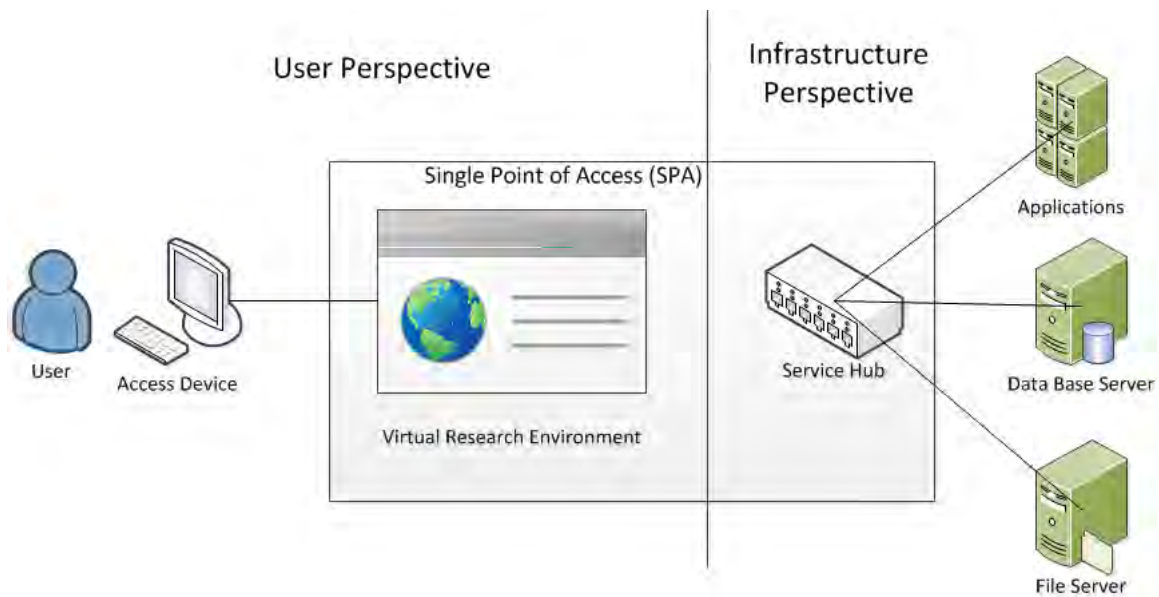


Figure 1. general description of the Gateway to European Research Services

### 3. INFRASTRUCTURE PERSPECTIVE

From an infrastructure point of view a Remote Access Network (RAN) as described in [1] is in place. Users connect from an access device via an internet connection and a web address. Afterwards they will be forwarded to a landing page, which is serving as a centralized Single Point of Access (SPA) where users have to log in with their credentials. By logging in at that Single Point of Access (SPA) they will be forwarded to a virtual research environment (VRE). The VRE itself is already one of the European Research Services provided via the Service Hub. Via menu buttons the user can access services provided via the Service Hub. These services are coordinated by a rights management system that allows the users to only access services they are accredited for. In general the Service Hub establishes connections to applications and data storage systems. Examples for applications are editor, an online submission tool, instant messenger, archiving tools, and statistical package. Examples for the content of the data storage systems are data documentation, research data, user data, publications, and text messages. It has to be kept in mind that data storage systems can be distributed all over Europe. The Service Hub and the SPA are responsible for putting them together in an organized way and provide them in an easy way the users. Data providers therefore do not have to build new infrastructures. They can bring in their existing solutions and benefit from the additional tools provided via the Service Hub. At the same time their

research data and services will become more visible and comparative analyses on a European level will be much easier.

#### **4. USER PERSPECTIVE**

For the researchers' the SPA with the Service Hub in the background deliver a one-stop shop feeling by offering a consistent and centralized access point to an adaptable virtual research environment containing different tools and data needed for the same or similar research projects; whereby the research data and services may be distributed all over Europe. For the user the most important service is the virtual research environment (VRE). For the user the VRE works as a graphical user interface (GUI) that allows interacting with the European Research Services); e.g. data documentations, online applications, messaging services, interfaces to research data, archiving tools, editors, and so on. The Service Hub is thereby build to be able to host additional services and tools if needed. But even if tools and data may be stored in different places, and the services accessible are depending on users digital rights the VRE will deliver a consistent working environment for the researchers always accessible through the SPA.

#### **5. RESEARCH WITH CONFIDENTIAL DATA**

When the mentioned infrastructure is used to work with confidential data a secure European Remote Access Network (EuRAN) becomes important [1]. This network will deliver secure connections and interfaces that are in compliance with the security restrictions of the participating data providers. Data provider can keep their data within their facilities and allow access only via secure remote access solutions as described in [2]. In addition to the VRE used to guide the users through the European research services, a SecureVRE to enable work with confidential data is needed. Such a SecureVRE will provide the same consistent but secure workspaces to the researchers and a sophisticated role management system to ensure security when working with confidential data. Also the access points need to be secured when working with confidential data. Users can choose one of the access points depending on the security level of the data they like to work with. Safe Rooms reflect the most secure access environment [3,4]; while data access from institutions is the most common solutions to work with confidential data.

#### **6. CONCLUSIONS**

A SPA and the Service Hub will harmonize the European research landscape and open up possibilities for real European research by providing the researcher with an access

location to access all the needed information – the Gateway to European Research Services. When adding the secure EuRAN and the SecureVRE also confidential data can be used within this infrastructure.

Building on the findings of the DwB project and other projects, a number of proposals are currently on the way to implement the proposed infrastructures and to create an environment for high level European research.

## **REFERENCES**

- [1] D. Schiller, Proposal for a European Remote Access Network – main components, Joint UNECE/Eurostat work session on statistical data confidentiality, (2013), 10p.
- [2] D. Schiller and R. Welpton, Distributing Access to Data, not Data – Providing Remote Access to European Microdata, IASSIST Quarterly (IQ), (forthcoming).
- [3] S. Bender and J. Heining, The Research-Data-Centre in Research-Data-Centre approach: A first step towards decentralised international data sharing, IASSIST Quarterly (IQ) Vol. 35 No. 3, (2011), 10-16.
- [4] M. Brandt and D. Schiller, Safe Centre Network – Need for a Safe Centre to Enrich European Research, Joint UNECE/Eurostat work session on statistical data confidentiality, (2013), 10p.

# An integrated system for euro area and member states turning points detection

Gian Luigi Mazzi ([gianluigi.mazzi@ec.europa.eu](mailto:gianluigi.mazzi@ec.europa.eu))<sup>1</sup>, Monica Billio<sup>2</sup> and Laurent Ferrara<sup>3</sup>

**Keywords:** Business cycle analysis, turning points detection, non-linear time series models, real-time evaluation.

## 1. INTRODUCTION

The global financial and economic crises of 2008-2009, as well as the sovereign debt crisis, started in 2011 for the European level, constitute the main two events of a new economic historical phase characterised by a high degree of uncertainty. This new phase, started between 2006 and 2007, is completely contradicting the paradigms of the so-called great moderation which characterised the 1990's and the beginning of the 21st century. In such a new economic situation, the occurrence of turning points is a more recurrent phenomenon which influences policy and decision makers' activities. It goes without saying that in this new scenario the importance of having a timely and reliable system for turning points detection has considerably grown its relevance and it has been considered as a priority for several institutions. In this paper, we are describing an integrated system for turning point detection developed by Eurostat during the last year, which is regularly used on monthly basis. The system is based on the assumption that the business cycle (as defined by Burns and Mitchell) and the growth cycle (expressed as a deviation from the trend) can be jointly monitored within the so-called ABCD approach developed by Anas and Ferrara (2005) and Anas, Billio, Ferrara and Mazzi (2008). In this approach, a complete cycle is decomposed into four distinguished phases delimited by four turning points. In particular, A and D are respectively the peak and the trough of the growth cycle, while B and C are the ones associated to the business cycle. The phases delimited by A-B and B-C, are usually called slowdown and recession, while those delimited by C-D and D-A, are respectively known as expansion and recovery. This approach provides a more complete picture of cyclical fluctuation than if we concentrate in a single cyclical definition, either the business or the growth cycle. Based on this approach we have then derived a pair of turning points coincident indicators, one for the business cycle and one for the growth cycle, which fully respect the ABCD sequence. This approach has been applied first to the euro area and then to its seven largest economies.

## 2. METHODS

In the context of turning point detection the main objective is to timely estimate discontinuity points or, in other words, regime changes for a given number of variables. For this purpose, thresholds non-linear models are considered the most appropriate to be used. In this paper, we decided to concentrate our attention on the Markov-Switching models family and in particular to the multivariate specification able to deal with multiple regimes in order to accommodate the various economic phases discussed in section 1.

---

<sup>1</sup> European Commission, DG Eurostat.

<sup>2</sup> University of Venice

<sup>3</sup> Banque de France.

Among the various kinds of switching mechanisms, we have privileged the models switching in mean but, due to the presence of heteroschedastic terms in several model specifications, at the end our models are allowed to switch both in mean and in variance. For the euro area and its largest economies we identified the best model specification in term of number of regimes and switching mechanisms and we fit it to a preselected set of statistical indicators including industrial production index, unemployment rate, and the confidence indicators respectively for industry, construction, retail trade and consumers. Table 1 presents the synthesis of model specifications of euro area and its largest economies.

**Table 1: model specifications of euro area and its largest economies**

Country	Model	Recessions	Slowdowns	Variables (differentiation order)					
				IPI	UR	BUIL	IND	CONS	RETA
EA	MSIH(4)-VAR(0)	R1	R1+R2	6	1	6	3	-	3
Belgium	MSI(4)-VAR(0)	R1	R1+R2	6	3	6	3	-	3
France	MSIH(4)-VAR(0)	R1	R1+R2	6	1	3	-	1	12
Germany	MSIH(4)-VAR(0)	R1	R1+R2	3	3	3	-	6	12
Italy	MSIH(5)-VAR(0)	R1	R1+R2	3	3	-	12	12	3
Netherlands	MSIH(4)-VAR(0)	R1	R1+R2	12	-	6	3	1	1
Portugal	MSI(5)-VAR(0)	R1+R2	R1+R2+R3	6	-	3	3	12	1
Spain	MSIH(4)-VAR(0)	R1	R1+R2	12	12	3	6	12	-

It is important to observe that among the chosen specifications, only two are characterised by five regimes instead of the usual four. This is the case of the model specification adopted for Italy and Portugal. Furthermore, two countries are not characterised by heteroschedastic behaviour, namely Belgium and Portugal.

### 3. RESULTS

The pair of coincident indicators respectively for the growth cycle and the business cycle is derived from the model describes in section 2. Those models are usually labelled as MS-VAR GCCI and MS-VAR BCCI respectively for the growth cycle and the business cycle. They measure the probability of being in slowdown or in recession respectively for the growth cycle and the business cycle. Since, according to the recommendation of Hamilton, we have decided to adopt a 0.5 threshold, an economy is supposed to be in slowdown/recession if the filtered probability returned by the MS-VAR GCCI/MS-VAR BCCI exceeds 0.5. Other ways the economy will be in a recovery/expansion phase. The behaviour of the two coincident indicators is regularly benchmarked with historical

dating chronologies obtained by means of non-parametric dating rules. The ability of indicators to replicate the turning points is regularly assessed by means of the Concordance Index and the Brier's Score. Furthermore, the type-1 error and type-2 error associated to our indicators are also regularly monitored. The type-1 error is defined as the inability of the model to signal an existing slowdown/recession, while the type-2 error is defined as the identification by the model of false slowdown/recession.

**Table 2: Growth Cycle Outcome Summary**

Country	Slowdown missed	False slowdown	Average delay in locating Slowdowns start (in months)	Accuracy in signalling slowdown	
				Brier's Score (QPS)	Concordance Index
Belgium	0	1 (2005)	0.7	0.16	0.83
France	1 (1998)	0	3.2	0.16	0.82
Germany	1 (1998)	0	2.5	0.20	0.79
Italy	0	0	4.3	0.22	0.77
Netherlands	1 (1995 – 1997)	0	2.0	0.18	0.80
Portugal	0	3	0.6	0.18	0.80
Spain	1 (1997-1998)	0	4.3	0.27	0.72
EA direct	0	1 (2004-2005)	2.0	0.15	0.83
EA indirect	1 (1998)	0	3.0	0.10	0.87

**Table 3: Business Cycle Outcome Summary**

Country	Recessions missed	False recessions	Average delay in locating peaks (in months)	Accuracy in signalling recessions	
				Brier's Score (QPS)	Concordance Index
Belgium	2 (1998 - 2000)	0	6	0.15	0.84
France	1 (2012)	0	2.5	0.06	0.94
Germany	0	1 (2001 - 2002)	3.4	0.08	0.92
Italy	1 (2001)	0	2.8	0.12	0.87
Netherlands	0	0	3.5	0.12	0.88
Portugal	0	0	4.3	0.17	0.82
Spain	0	0	1.3	0.05	0.94
EA direct	0	0	2.3	0.06	0.94
EA indirect	1 (2011-2013)	0	2.5	0.06	0.90

#### **4. CONCLUSIONS**

The proposed MS-VAR specification for the turning point indicators has proven over a quite long period to provide reliable turning point signals both for the growth and the business cycles. The pair of indicators derived for each geographical entity is fully consistent with the ABCD approach and they are characterised, generally, by very small delays in detecting turning points. Furthermore, the number of false signals and missed cycles observed for each pair of indicators is very limited with a few exceptions showing the need of improving the indicators. The elaboration of turning points indicators for the remaining euro area countries is ongoing and first results are expected by middle 2015. This will be also the opportunity for a global revision of national turning point indicators in order to achieve a comparable degree of reliability among countries. The full coverage of the euro area countries will constitute a very important contribution towards a reliable monitoring of the cyclical situation for the euro area and its member countries.



# A new graphical tool for business cycle monitoring

Jacques Anas<sup>1</sup> (jacques.anas@free.fr), Ludovic Calès<sup>2</sup> (ludovic.cales@hendyplan.com) and Gian Luigi Mazzi<sup>3</sup> (Gianluigi.Mazzi@ec.europa.eu)

**Keywords:** business cycle monitoring, graphical representation, coincident indicators, non-linear models, temporal and spatial comparability.

## 1. INTRODUCTION

Complementing and/or integrating traditional dissemination tools with more advanced graphical and interactive ones can facilitate data understanding and interpretations. It can also facilitate the identification of signals which are usually hidden when looking in a traditional way to statistics. In the recent years, several institutions including Eurostat have invested a lot in such new dissemination tools. More specifically, in the field of the business cycle analysis, several versions of business cycle monitoring tools have been implemented, for example by Eurostat, the CBS, DESTATIS, the IFO institute and the OECD. In designing a new version of the business cycle clock (BCC) we have carefully investigated and critically reviewed the above mentioned business cycle monitoring tools. With reference to the Eurostat business cycle clock, we have identified the following weaknesses:

1. It presents the evolution of a number of basic indicators which may at times provide diverging signals on the economy. The presentation of few synthetic cyclical indicators would provide clearer messages on the economic cyclical situation, as the OECD's BCC.
2. The "economic recessions" as defined by Burns and Mitchell (1946) cannot be identified with the BCC as it is. Indeed, the descendent phase of the cycle is divided into a "downturn" step when the deviation to trend (or output gap) remains positive and a "slowdown" (sometimes misleadingly worded as "recession") step when the deviation to trend becomes negative. The term "recession" often used in the business clock is confusing because it does not match the classical definition of a recession.

The new version of the business cycle clock presented below, aims to overcome such weaknesses and to provide a comprehensive and statistically sound picture of the cyclical situation at euro area and member countries levels.

## 2. METHODS

In this section we are proposing a new representation of the cyclical situation on a clock basis where the hand of the clock moves clockwise, according to the values returned by the cyclical probabilistic indicators of turning points produced by Eurostat. Those indicators refer respectively to acceleration cycle (ACCI), growth cycle (GCC) and business cycle (BCCI)<sup>4</sup>, according to the  $\alpha\text{AB}\beta\text{CD}$  approach. The indicators for growth

---

<sup>1</sup> ACE (Analysing Cycles in Economies)

<sup>2</sup> Hendyplan

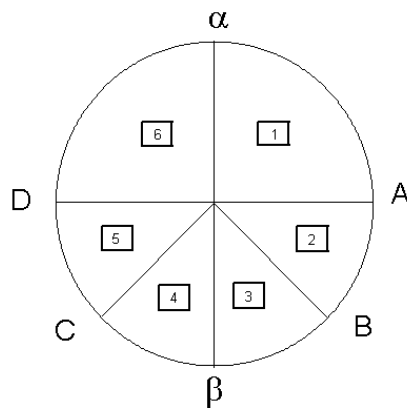
<sup>3</sup> European Commission, DG Eurostat.

<sup>4</sup> For details on the indicators, we refer to Billio M., Ferrara L. Mazzi G.L., Moauro F. (2011) 'A multivariate system for turning point detection in the euro area', [www.oecd.org/dataoecd/47/58/49067716.pdf](http://www.oecd.org/dataoecd/47/58/49067716.pdf);

and business cycle are based on a MS-VAR representation which jointly returns growth cycle and business cycle recession probabilities. The indicator for the acceleration cycle is based on a univariate Markov-switching model.

The clock is designed so that:

- Noon is  $\alpha$ , peak of the growth rate cycle
- 3 is A, peak of the growth cycle
- 4.30 is B, peak of the business cycle
- 6 is  $\beta$ , trough of the growth rate cycle
- 7.30 is C, trough of the business cycle
- 9 is D, trough of the growth cycle



**Figure 1: Clock based on the  $\alpha AB\beta CD$  approach<sup>5</sup>.**

Thus, the quadrants can be read as follows:

- The upper and lower right quadrants (sections 1, 2 and 3) show a decrease in the growth rate. In the first quadrant, the growth rate is still above the trend growth rate. In A the growth rate slips below the trend growth rate. In the second quadrant, the growth rate is below the trend growth rate. At point B, it becomes negative, and at  $\beta$  the growth rate reaches a minimum.
- The lower and upper left quadrants (sections 4, 5 and 6) show an increase in the growth rate. In the third quadrant, the growth rate is still below the trend. At point C, it becomes positive and at point D it overpasses the trend growth rate.

Visually, the clock can be read as follows:

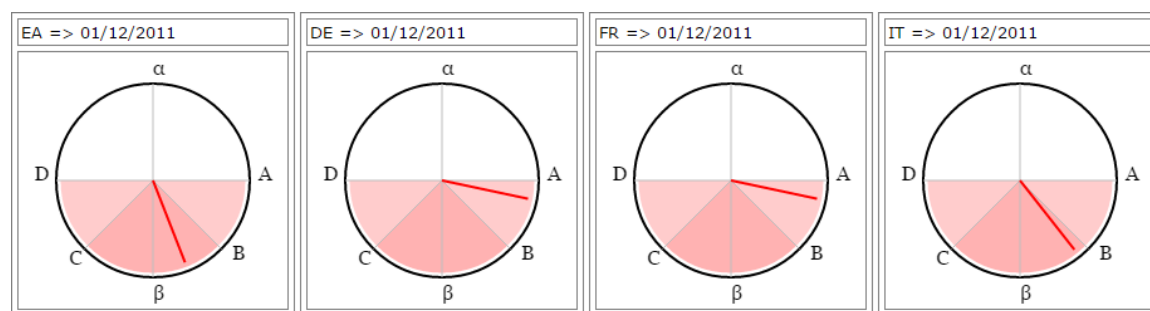
- The lower half of the clock (sections 2, 3, 4 and 5) represents below-the-trend growth rates. The bottom region between B and C, i.e. between 4.30 and 7.30 would depict the recession. It corresponds to one quarter of the graph. The two half quarters of the clock between A and B (3 and 4.30) and between C and D (7.30 and 9) depict the slowdown. They surround the recession area.
- The upper half of the clock (sections 6 and 1) represents the above-the-trend growth rates (expansion growth cycle phase).

<sup>5</sup> For a pure  $\alpha AB\beta CD$  business cycle.

With this representation, the arrow moves forward and backward. It also permits the visualization of pure acceleration cycles (e.g. a jump from section 1 to section 6) and pure growth cycles (e.g. a jump from section 2 to section 5).

This clock representation of the cyclical situation allows for a static and dynamic analysis of the euro area and each member country, as well as for cross-country comparisons.

The four clocks in figure 2 refer to the Euro area, Germany, France and Italy.



**Figure 2: 4 clock representations**

### 3. RESULTS

In this section we show how the clock representation can be used to analyse in real time the exit from the recession in 2013.



**Figure 3: Illustration of the  $\alpha AB\beta CD$  clock with the 2013 recession exit**

Each clock shows the cyclical situation at a given month, based on the smoothed probabilities of the cyclical indicators computed in February 2014. Within the recession, from July to October 2012, the hand already passed the peak of the business cycle (B) and did not reach the trough of the growth rate cycle ( $\beta$ ). It corresponds to a worsening of the recession as the economy decelerates. It is characterized by  $ACCI > 0.5$ ,  $BCCI > 0.5$  and  $GCCI > 0.5$ , i.e. section 3 in Figure 1. The hand passed the trough of the growth rate cycle ( $\beta$ ) in November. The economy re-accelerates while still being in recession. Thus it corresponds to the ACCI passing below 0.5 (section 4). Finally, in June 2013, the economy exits the recession, passing point C and entering in section 5 characterized by  $ACCI < 0.5$ ,  $BCCI < 0.5$  and  $GCCI > 0.5$ .

#### **4. CONCLUSIONS**

The new business cycle clock proposed in this paper constitutes, in our view, a relevant step forward with respect to the previous one. It is a real time monitoring tool for the cyclical situation of euro area and its member countries allowing for both temporal and cross- country analysis and comparison. It also permits to go deeper in the details of all phases of the cycles due to the joint analysis based on acceleration, growth and business cycles. The default version presented in this paper is based on the probabilistic turning point indicators elaborated each month by Eurostat. The new business cycle clock could be subject to further extension allowing for a higher degree of interactivity for the users. The first extension is constituted by the possibility offered to the users of recalculating in real time the probabilistic indicators based on the most recent information. In this case, users should be warned that the recalculated indicators are automatically computed and not subject to the detailed control quality which is performed on a monthly basis by Eurostat on the indicators used for the default version of the tool. A second extension consists in offering the users the possibility of specifying the probabilistic models by modifying the thresholds, the number of regimes and lags of the autoregressive part. In this case, users should be aware that using unrealistic hypothesis on the specifications could lead to very misleading exercise. Nevertheless, this simulation exercise could have a high didactic content showing how probabilistic turning point indicators are sensitive to model specifications.

# Towards a daily indicator of economic conditions

Massiliano Marcellino ([massiliano.marcellino@unibocconi.it](mailto:massiliano.marcellino@unibocconi.it))<sup>1</sup>, Claudia Foroni<sup>2</sup>, Gian Luigi Mazzi<sup>3</sup> and Fabrizio Venditti<sup>4</sup>

**Keywords:** Composite indicators, mixed-frequency models, daily measure of economic conditions, smoothing, interaction between financial and macroeconomic variables, moderating volatility.

## 1. INTRODUCTION

In the recent years, the need for macroeconomic information at very high frequency has considerably increased. The strong interaction between macroeconomic activity and financial markets and the ability of the latter to influence almost on daily basis macroeconomic conditions, has generated a need for a daily updated macroeconomic indicator reflecting such conditions. Ideally, such kind of indicator should be represented by a daily updated version of GDP or by a robust proxy of it. The aim of this paper is to derive an innovative macroeconomic indicator, timely available, updated on a daily basis, attractive, easy to read and to communicate, with strong methodological foundations and having a high macroeconomic content.

### 1.1. Monthly indicator daily updated versus daily indicator

PEEIs are available at monthly or quarterly frequency but some of them, especially those referring to financial markets, can also be available at daily frequency. Furthermore, additional variables, non-included in the PEEIs, such as the oil price and the commodity prices index can be considered in the compilation of the indicators, due to their ability to impact the economic activity. In constructing a composite indicator, an important point to be addressed is which frequency we would like to construct the indicator at. Alternatives are: monthly and daily.

A related issue is the frequency of the updating of the indicator. Since in a month there are a number of releases of PEEIs covering almost all working days, it is clear that even an indicator available at monthly frequency can be updated almost daily. Furthermore, if we include daily financial variables it is obvious that such indicator will change every day. The main difference between the two cases is that the monthly indicator will remain stable in terms of number of observations for a month (even if the most recent observations will be revised) while the daily one will show a new observation each working day. Taking into account the fact that both the monthly daily updated and the daily indicator will require a daily computation and considering the higher content of actual information, the daily indicator appears to be the best and most appealing choice.

### 1.2. What the indicators should measure

There are three kinds of indicators that could be constructed:

- A) GDP based indicator
- B) Indicator of the general economic situation, such as the Conference Board type
- C) Economic perception based indicator, such as the economic sentiment indicator based on tendency surveys

---

<sup>1</sup> Bocconi University

<sup>2</sup> Norges Bank

<sup>3</sup> European Commission, DG Eurostat.

<sup>4</sup> Bank of Italy

The first possibility is the most appealing one also because it has a direct interpretation. The construction of a daily indicator of GDP based on the national accounts framework is obviously unfeasible at the moment. By contrast, constructing a daily proxy of GDP reflecting the daily perception of economic agents on the total production of the current months is absolutely feasible.

## **2. METHODS**

Firstly we have analysed from the theoretically point of view a number of mixed frequency models which have suitable characteristics for dealing with daily information too. Then we have presented an evaluation of the relative performance of mixed frequency models applied to a large daily/monthly dataset for constructing a daily indicator of economic conditions for the euro area. We have focused on single indicator bridge/MIDAS/UMIDAS models and we have identified the most promising indicators for each class, which only partly overlap.

We have also compared the results of single indicator bridge/MIDAS/UMIDAS models with pooling, to have a first evaluation of the usefulness of a larger information set. It turns out that pooling is not so useful in this context, and therefore we will not further consider it.

We have then constructed a Conference Board type of indicator and compared its performance with that of the best daily and monthly single indicators. It turned out that the latter are generally better, which highlights the usefulness of a more model based approach to the construction of the daily indicator of economic activity.

We have then extended the analysis as follows:

- Assess the robustness for the single indicator MIDAS-UMIDAS models to the use of real time (rather than final) data.
- Implement multi indicator UMIDAS and bridge models, based on the best single indicators resulting from the real time single indicator analysis.
- Evaluate multi indicator UMIDAS and bridge models for the euro area from a statistical and economic point of view.
- Compare them with pooling of selected single indicator UMIDAS and bridge models.
- Implement the preferred multi indicator based procedure for the euro area also for the largest euro area countries and assess the performance.

## **3. RESULTS**

Based on the detailed investigation described in section 4, we have in a first step identified a limited number of variables which have shown to be the best performing ones. This set of variables is composed of two monthly ones, the industrial production index and the economic sentiment indicator, and two daily ones, the nominal exchange rate and the 10 years bond yield. Those variables have then been modelled using a UMIDAS approach which has shown to be computational efficient and performing not worse than more sophisticated models. Finally, we have compared four different specifications of the UMIDAS model in forecasting GDP: with and without AR error structure and with and without a step dummy to accounting for the recessionary period. The retained specification providing the most satisfactory results in forecasting the current quarter GDP has been the one without AR component and with step dummy variable.

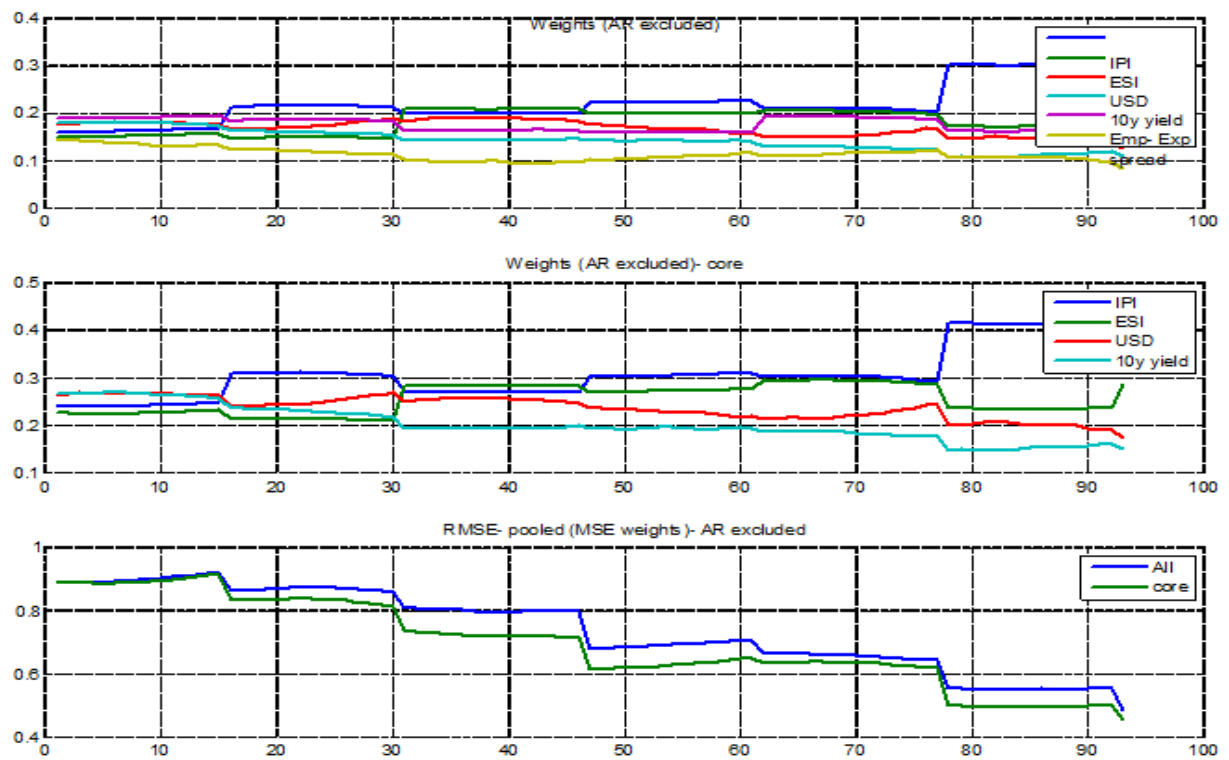


Figure 1

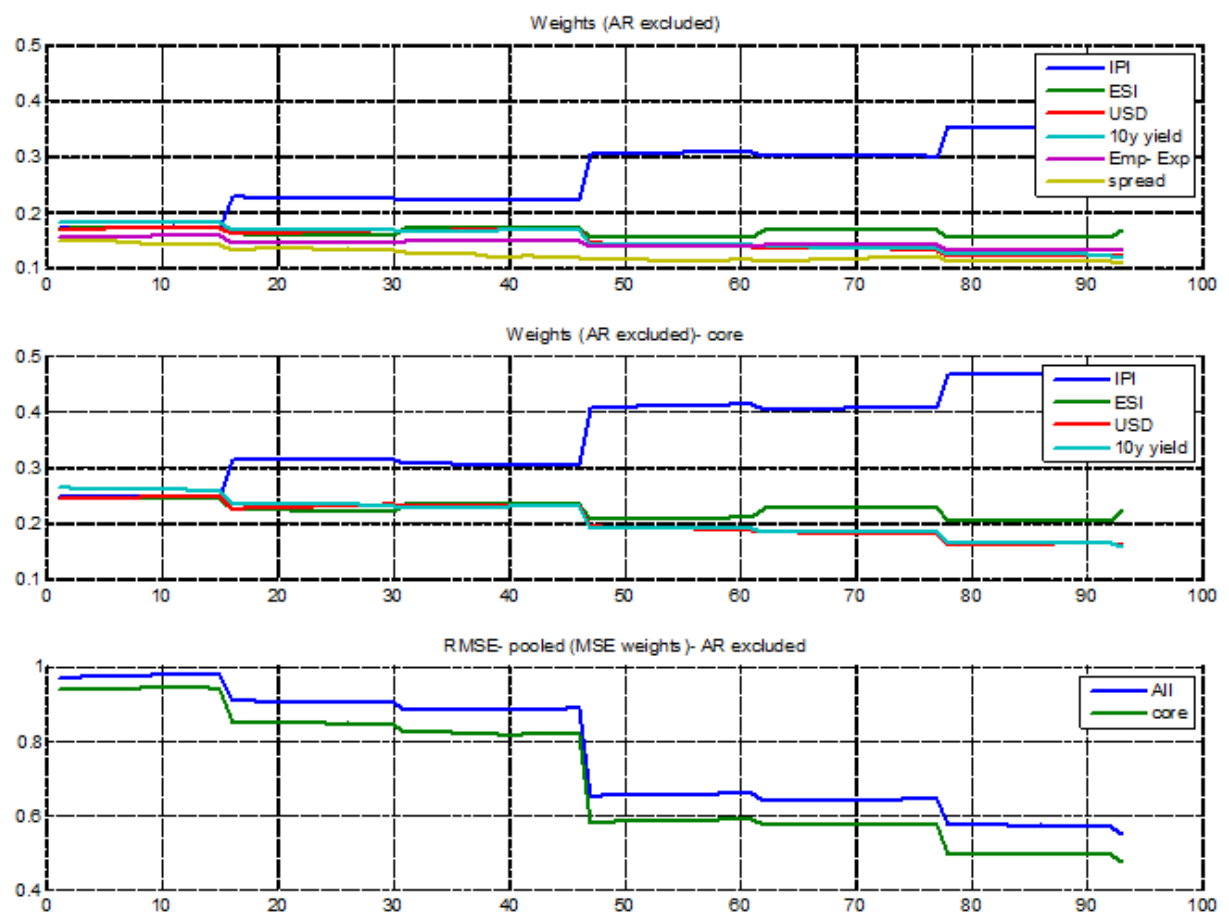


Figure 2

The upper and medium panels of Figures 1 and 2 report the (day by day evolution of the rescaled inverse MSE) weights of each single, respectively, UMIDAS and Bridge model. We see that the weights are similar across indicators and stable over time, with a slight dominance for IP, whose weight also increases a bit over time. The lower panels of Figures 1 and 2 presents the daily evolution of the RMSE of the 6- and 4-combined GDP nowcasts from the pooled UMIDAS and bridge approaches, respectively. We see that the RMSE values are very similar across methods, and the core 4-variable nowcast is slightly better than the 6-variable combination for both methods. Even if the RMSEs are a bit large mainly due to the inclusion of some quarters of the crisis in the simulation period, they nicely decline over the quarter, as more and more information is available.

In summary, the analysis conducted in this section supports the use of an inverted MSE combination of four single UMIDAS models (without AR) based on monthly IP and ESI and daily 10y and USD as the econometric tool to produce a daily economic indicator for the euro area.

#### **4. CONCLUSIONS**

The identified procedure shows the following interesting characteristics which suggest it like the most appealing for regular production.

1. Is easy to implement (based on OLS estimation) but still slightly more sophisticated than standard bridge models.
2. Is based on well-known and generally accepted indicators, available and comparable also across countries.
3. Produces results that have an economic interpretation (all the indicators have the proper sign).
4. Gives increasing weight to the monthly indicators as time passes within the quarter but takes into account daily information.
5. Produces good nowcasts of GDP growth.



# “Remote Access to European Microdata”

Maurice Brandt ([maurice.brandt@destatis.de](mailto:maurice.brandt@destatis.de))

**Keywords:** Remote Access, Confidential Microdata, Improvement of Access, Science

## 1. INTRODUCTION

To improve the access to European microdata for scientific purposes the ESSnet-project “Decentralised and Remote Access to Confidential Data in the ESS” (DARA) was conducted. The aim of the ESSnet-project DARA was to establish a secure channel from a safe centre within a National Statistical Institute (NSI) to the safe server at Eurostat, so that researchers can use confidential EU microdata in their own Member States without travelling to Luxembourg.

The ESSnet DARA-project team has defined a concept of technical implementation and safety requirements for a European remote access system. The concrete task of participating NSIs was to provide a secure channel to guarantee access for data users to the central node and also to provide service and IT-support for the researchers on the local national level.

The project team has drafted a handbook with descriptions and guidelines for NSI staff and researchers and an accreditation system for access facilities. For a proof of the concept and feasibility, the project team has implemented a remote access pilot with 6 access points in 5 countries in Europe.

## 2. METHODS AND WORK COMPLETION

One main task was the preparation of the implementation of a pilot infrastructure from a technical point of view. At the beginning of the ESSnet DARA project, the team developed security requirements and user demands for a European Remote Access System.

One crucial topic that was investigated was the detail of the connection to the central system. The safe connection between the European Commission (EC) and the NSIs in the Member States is guaranteed by a safe network named “Secure Trans European Services for Telematics between Administrations” (sTesta) based on a private network. Only institutions that are part of the private network are able to connect to the CITRIX server within the network of the EC. This means that the NSIs have to join the sTesta network before they are able to establish a connection. There is also a login and password required to access the working platform. On the one hand it is a higher burden for the NSIs. On the other hand the whole system itself is more secure because it is not possible to connect to the CITRIX server from any other location outside the network. During the preparation phase of the pilot, the method for connecting to the Citrix server at the EC was investigated. Until midterm of the project there were two NSIs connected via sTesta network. Other NSIs had difficulties to connect via sTesta because this part of the IT infrastructure was outsourced to an external company. This also made the connection method much more expensive as the external companies were charging monthly fees for maintaining the connection. For various reasons it is not widely used by the NSIs in the

MS. The method for a secure connection was not very promising during the first half of the project. This is why further investigation on a secure channel via VPN has been conducted. The aim was to find a secure mode of connection between MS and Eurostat that can be used by all MS.

To test and evaluate the usability of the pilot, a testing plan was drafted. The aim of the proof of concept was to set up the testing phase of the pilot of a trans-border Remote Access for each country involved in the project.

The pilot was designed for a European system, but it was built from an existing solution in France. The project has set up a new infrastructure based on the requirements defined at the beginning, like user needs, security requirements, roles, workflow specifications, etc.. The conclusions are that the proof of concept demonstrates that all requirements defined are relevant and a solution that fulfils all these requirements can be implemented considering the security of the system, the usability for the users, and the decentralised management in each country.

The pilot shows that a European Remote Access System can be implemented in safe conditions. To establish a connection, only a broadband internet connection is needed. First the IP settings and the Proxy-server need to be configured in the so called DARA Box. A registered person like a Support Officer or researcher can login with a smartcard and fingerprint reader. Then they can access the central server via an encrypted and secure connection to a virtual machine using a familiar windows desktop that has statistical editing and analysis software with MS Office programs.

### **3. RESULTS**

After the experience of the ESSnet DARA, the project team can give a clear recommendation on a European microdata access infrastructure. After first tests the alternative pilot looks very promising in terms of security, costs, management, deployment and delegation in a European context and user friendliness. Therefore the recommendation is the implementation of the DARA pilot with Eurostat as the central node. It can be also a system that fulfills all defined user and security requirements with Eurostat as centre, where the DARA pilot is an example that has been tested successfully according to the specifications. The first accreditations of Access Facilities, starting with NSIs, will show how the realisation of the Regulation (EU) No 557/2013 works in practice. For the near future it is advisable to build up a circle of trust and competences for microdata access for scientific purposes in Europe.

There are also lessons learned that lead to recommendations which cannot be given for a production phase of a Remote Access System. Such a system has very specific requirements and should be implemented in a dedicated environment and not on a shared platform with other services. The sTesta connection is appealing, but for practical reasons it should be avoided for a real implementation because it is not widely used and an extension to other access points for a long term vision is doubtful.

During the project there have been requests from researchers in Europe, especially from the OECD, Portugal, Hungary and the UK for detailed microdata on the Labour Force Survey and other datasets. The interest was huge to use those data from location near their own research institutions. This shows that there is demand and utilisation for this system and it should be also expanded to other EU statistics and countries.

The complete final project report is accessible on cros portal under the following link:  
[http://www.cros-portal.eu/sites/default/files//final\\_report\\_ESSnet\\_DARA\\_20140321\\_publishable.pdf](http://www.cros-portal.eu/sites/default/files//final_report_ESSnet_DARA_20140321_publishable.pdf)

And a summary paper of the ESSnet DARA project is available under:  
[http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic\\_3\\_Brandt.pdf](http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_3_Brandt.pdf)

#### **4. CONCLUSIONS**

In conclusion, the DARA pilot shows, as a proof of concept, that it is possible to set up a European decentralised Remote Access to confidential microdata in excellent conditions considering:

- A high level of security with a strong authentication method and a leak-proof infrastructure prohibiting data files extraction.
- A good usability for researchers.
- A flexible system management that allows delegation to local Safe Centre for enrolment, project creation, output checking, etc.
- An easy method for the deployment and the installation.
- A cost-effective solution for both central node and remote sites (no strong local security IT requirements).

All this was possible due to the specifications produced, like user requirement, handbooks, workflows, IT security requirements and user needs. The proof of concept validates all the concepts defined in this way, and furthermore helps to produce the cost analysis study.

The extended study, produced by the project, would be very useful for the design of the real implementation as well as for the production phase (handbook, workflows, requirements, etc.). The benefits for this European microdata access system are that secure data server and devices for thin clients in the Safe Centres could be provided by the central node so that there are no investments for IT equipment necessary for the MS planning to join this system. No microdata will be transferred to another MS, only the access will be granted from another accredited Access Facility whereas the microdata itself will remain at the secure servers inside Eurostat. This system needs to be affordable to maintain it over the years and to build a sustainable solution which can be also used in the future.

Only if the mode of microdata access is secure and user friendly the data can be used by European researchers for their analysis. This will contribute to a better understanding of processes and developments in Europe and can help to find best practise examples that can improve conditions in all Member States. In a context of a “European society” only an evidence based assessment of the situation with empirical microdata can lead to realistic measures for Europe. Furthermore, actions which have been implemented in Europe can be evaluated and adjusted if high quality EU microdata is available for researchers who are working and experienced in the field of empirical European studies. All this can contribute to the improvement of social and living conditions in Europe.

## REFERENCES

- [1] Bond, Stephen, and Maurice Brandt, Tony Chapple, Anja Crössmann, Eric Debonnel, Philippe, Donnay, Ana Dulce Duarte Pinto, Kamel Gadouche, Anja Hlawatsch, Julia Hoeninger, György Káplán, Joaquim Machado, Anja Malchin and Zoltán Vereczkei (2013): ESSnet DARA - FINAL REPORT, Luxembourg.
- [2] Brandt, Maurice (2013): Improvement of access to European microdata, Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa.
- [3] Brandt, Maurice und David Schiller (2013): Safe Centre Network - Need for Safe Centre to enrich European research, Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa.
- [4] Brandt, Maurice (2013): Decentralised and Remote Access to Confidential Data in the ESS, New Techniques and Technologies for Statistics, Brussels.

# Variance Estimation in Complex Sampling Designs: The Finite Population Bootstrap Using Pseudo-Populations

Andreas Quatember  
Johannes Kepler University Linz (Austria)

**Keywords:** Variance Estimation, Resampling, Complex Surveys, Bootstrap, Finite Populations

## 1. INTRODUCTION

When no explicit variance formula is available and the calculations for Taylor linearization are too cumbersome, so-called “computer-intensive methods” that use computer power instead of heavy calculations, can be applied alternatively. One such procedure is the bootstrap method. This strategy falls under the resampling methods. Lahiri (2003) described the bootstrap as “probably the most flexible and efficient method of analyzing survey data since it can be used to solve a variety of challenging statistical problems (e.g., variance estimation, imputation, small-area estimation, etc.) for complex surveys involving both smooth and non-smooth statistics” (p.199).

The technique was originally developed by Efron (1979) for the estimation of the sampling distribution of a statistic under i.i.d. conditions. The empirical distribution of a random variable  $y$  as observed in the i.i.d. sample can be interpreted as the ML estimator of the true probability distribution of  $y$ . Drawing i.i.d. “resamples” of the same size as the original sample from this empirical distribution, the true sample distribution of the statistic under study is approximated by the theoretical distribution of the estimator calculated in all possible resamples. This bootstrap distribution in turn can be approximated by the Monte Carlo approximation, for which  $B$  resamples are drawn. Within each of these bootstrap samples, the statistic is calculated in the same way as in the original i.i.d. sample. For a large  $B$ , the distribution of the statistic in the bootstrap samples is interpreted as an estimation of its sample distribution.

With increasing computer power, this technique has also become attractive for complex finite populations surveys. For this purpose, different approaches are available in the relevant literature (cf., for instance, Shao and Tu (1995), p.247ff, or Wolter (2007), p.200ff). However, for a direct extension of the i.i.d. bootstrap to finite population sampling, the population  $U$  of  $N$  elements takes over the role of the unknown probability distribution in the i.i.d. context. Gross (1980) proposed in this context for a sample  $s$  drawn by simple random sampling without replacement (SI) and integer design weights  $N/n$  to generate a set-valued estimator  $U_{SI}^*$  of the true population  $U$  of size  $N$  by replicating each element of  $s$   $N/n$  times. The bootstrap population  $U_{SI}^*$  can be seen as the finite population of size  $N$  with the maximum likelihood regarding the sample drawn. In the next step,  $B$  bootstrap samples of size  $n$  are drawn from the bootstrap population  $U_{SI}^*$  by applying the original sampling method. These resamples form the basis for estimating the SI sampling distribution of the estimator for the interesting parameter based on simulations.

Obviously, for general applicability in survey sampling, the idea of Gross (1980) had to be extended to non-integer design weights, and general probability sampling with arbitrary first-order inclusion probabilities. In fact, the key to an efficient direct application of the bootstrap method in finite population sampling is the generation of a pseudo-population  $U^*$  suitable as the basis for drawing the bootstrap samples with respect to the estimation problem to be solved.

## 2. METHODS

The bootstrap methods using pseudo-populations try to establish bootstrap populations, which include integer numbers of replications of the original sample values. For example, Booth et al. (1994) present an approach for SI sampling, in which each sampling unit  $k$  is replicated  $i$  times, where  $i$  is the integer part of  $N/n = i + r$  and  $r$  is the non-integer “rest”. For the completion of the bootstrap population, an additional SI sample of  $N - n \cdot i$  of the units is drawn from  $s$ . Addressing the random process of this last step of the generation process, this step is repeated  $C$  times providing  $C$  different bootstrap populations. In each of them, the interesting estimator is calculated in  $B$  bootstrap samples and the distribution of these estimates estimates the interesting SI sampling distribution.

In this way, such “solutions affect the characteristics of the resulting bootstrap population which might differ from the *nominal*  $U^*$  to an uncontrollably large extent, thus violating in the same measure the mimicking principle and the plug-in approach” (Ranalli and Mecatti 2012, p.4095) of Efron’s original bootstrap.

Further, according to Barbiero and Mecatti (2010), the following understandable properties should apply to a bootstrap algorithm with respect to the estimation of a total  $t$  of variable  $y$  (cf. p.60ff):

1. Given the original sample  $s$ , in a bootstrap population  $U^*$ , the total of an auxiliary variable  $x$  should be equal to the total of  $x$  in  $U$ .
2. The total of  $y$  in  $U^*$ , should be equal to the Horvitz-Thompson (HT)  $t_{HT}$  estimator of  $t$ .
3. Over the resampling process, for given  $s$ , the  $B$  HT estimators of  $t$  calculated in the  $B$  resamples should have an expectation of  $t_{HT}$ .

In the talk, a procedure is presented, which complements the proposals of Holmberg (1998) and Barbiero and Mecatti (2010) for general probability proportional to size sampling ( $\pi ps$ ) with design weights defined as usual as the reciprocals of the first-order inclusion probabilities in the most natural way and does not violate the mimicking principle of the original approach anymore. This HT based bootstrap approach allows also non-integer numbers of replications of the sample values of variables  $y$  and  $x$  to generate the bootstrap population  $U_{HT}^*$ . The replication factor of each unit  $k$  of  $s$  corresponds exactly to their design weights  $d_k = i_k + r_k$  with  $i_k$  being the integer part of  $d_k$  and  $r_k$  being the “rest”. In this way, a bootstrap population is generated, which contains not only whole units with values  $y_k$  and  $x_k$ , but also parts of whole units with these values as it is the case also in the common HT estimator  $t_{HT}$  for the total of variable  $y$ :

$$t_{HT} = \sum_s y_k \cdot d_k$$

### 3. RESULTS

The pseudo-population  $U_{HT}^*$  has an expected size of  $N$ . For SI sampling, this means that a bootstrap population with size  $N$  is guaranteed. In the resampling process, based on this bootstrap population  $U_{HT}^*$ , a whole unit  $k$  belonging to this population has a resample inclusion probability proportional to its original  $x$  value. But, for the  $r_k$ -piece of unit  $k$ , this probability is proportional to  $r_k$  times  $x_k$ .

For the proposed HT based technique, all three desirable properties for efficient variance estimation mentioned in Section 2 (cf. Barbiero and Mecatti 2010, pp.60ff) apply. Hence, after the generation of  $U_{HT}^*$  as a set-valued estimator of  $U$ , the design weights  $d_k$  of the elements in  $U_{HT}^*$  will not have to be recalculated.

In the talk, results from a simulation study are presented that show the effect of the HT based bootstrap method on the sizes of the generated pseudo-populations, the quality of the variance estimation and compare these results to other methods from the literature.

### 4. CONCLUSIONS

Besides being more understandable because it follows the same idea as the one behind the widely used HT estimator when it comes to the composition of the bootstrap population, the presented method is also more efficient than other approaches because properties, important for the performance of the method, apply to this bootstrap algorithm. Additionally, the HT based bootstrap can still be used in  $\pi$ PS situations, where other methods fail for first-order sample inclusion probabilities of the population units which are close to one, because with the HT based method, these probabilities need not be recalculated before the resampling process.

### REFERENCES

- Barbiero, A. and F. Mecatti (2010). Bootstrap algorithms for variance estimation in  $\pi$ PS sampling. In P. Mantovan and P. Secchi (Eds.). *Complex Data Modeling and Computationally Intensive Statistical Methods*. 57-69, Springer, Milan.
- Booth, J. G., Butler, R. W. and P. Hall (1994). Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association*. 89(428), 1282-1289.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*. 7, 1-26.
- Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 181-184.
- Holmberg, A. (1998). A Bootstrap Approach to Probability Proportional-to-Size Sampling. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 378-383.
- Lahiri, P. (2003). On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation. *Statistical Science*. 18(2), 199-210.
- Ranalli, M. G. and F. Mecatti (2012). Comparing Recent Approaches For Bootstrapping Sample Survey Data: A First Step Towards A Unified Approach. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 4088-4099.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer, New York.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer, New York.



# An R Library to construct empirical likelihood confidence intervals for complex estimators

Yves G. Berger, University of Southampton, UK

**Keywords:** Calibration, Design-based approach, Estimating equations, Finite population corrections, Hajek estimator, Horvitz-Thompson estimator, Regression estimator, Stratification, Unequal inclusion probabilities.

We developed an R library which can be used to compute empirical likelihood point estimates and confidence intervals. After explaining the empirical likelihood theory, we show how to use this library and an example based on the 2009 EU-SILC survey.

## 1. INTRODUCTION

Under complex sampling designs, point estimators may not have a normal sampling distribution and linearised variance estimators may be biased. Hence standard confidence intervals based upon the central limit theorem may have poor coverages. We propose an empirical likelihood approach which gives design based confidence intervals. The proposed approach does not rely on the normality of the point estimator, variance estimates, design-effects, re-sampling, joint- inclusion probabilities and linearisation, even when the estimator of interest is not linear. It can be used to construct confidence intervals for a large class of complex sampling designs and complex estimators which are solution of an estimating equation [4]. It can be used for means, regressions coefficients, quantiles, totals or counts even when the population size is unknown. It can be used with large and negligible sampling fractions. It also provides asymptotically optimal point estimators, and naturally includes calibration constraints [2]. The proposed approach is computationally simpler than the pseudo empirical likelihood [9] and the bootstrap approaches [8]. Berger and De La Riva Torres [1] show that the empirical likelihood confidence interval may give better coverages than the approaches based on linearisation [3], bootstrap [8] and pseudo empirical likelihood [9].

## 2. EMPIRICAL LIKELIHOOD APPROACH

Let  $U$  be a finite population of  $N$  units; where  $N$  is a fixed quantity which is not necessarily known. Suppose that the population parameter of interest  $\theta_0$  is the unique solution of the following estimating equation [4].

$$G(\theta) = 0, \quad \text{with} \quad G(\theta) = \sum_{i \in U} g_i(\theta);$$

where  $g_i(\theta)$  is a function of  $\theta$  and of the characteristics of the unit  $i$ , such as the variables of interests and the auxiliary variables.

We propose to use the following *empirical log-likelihood function* [e.g. 1, 7].

$$\ell(m) = \sum_{i=1}^n \log(m_i),$$

where  $\sum_{i=1}^n$  denotes the sum over the sampled units. The quantities  $m_i$  are unknown positive scale loads. The maximum likelihood estimators of  $m_i$  are the values  $\hat{m}_i$  which maximise  $\ell(m)$  subject to the constraints  $m_i \geq 0$  and

$$\sum_{i=1}^n m_i \mathbf{c}_i = \mathbf{C};$$

where  $\mathbf{c}_i$  is  $Q \times 1$  vector associated with the  $i$ -th sampled unit and  $\mathbf{C}$  is  $Q \times 1$  vector (see Section 2.1). The values  $\hat{m}_i$  are survey weights.

## 2.1. Maximum empirical likelihood estimator

Suppose that the finite population  $U$  is stratified into  $H$  strata denoted by  $U_1, \dots, U_h, \dots, U_H$ ; where  $\cup_{h=1}^H U_h = U$ . Suppose that a sample  $s_h$  of fixed size  $n_h$  is selected with replacement with unequal probabilities from  $U_h$ . Let  $\mathbf{c}_i = \mathbf{z}_i$  and  $\mathbf{C} = \mathbf{n}$ ; where  $\mathbf{z}_i$  are the values of the design (or stratification) variables defined by  $\mathbf{z}_i = (z_{i1}, \dots, z_{iH})^\top$ , where  $\mathbf{n} = (n_1, \dots, n_H)^\top$  denotes the vector of the strata sample sizes, with  $z_{ih} = \pi_i$  when  $i \in U_h$  and  $z_{ih} = 0$  otherwise. It can be shown that  $\hat{m}_i = \pi_i^{-1}$ . Let  $\ell(\hat{m}) = \sum_{i=1}^n \log(\hat{m}_i)$  be the maximum value of the empirical log-likelihood function.

Let  $\hat{m}_i^*(\theta)$  be the values which maximise  $\ell(m)$  subject to the constraints  $m_i \geq 0$  and  $\sum_{i=1}^n m_i \mathbf{c}_i^* = \mathbf{C}^*$  with  $\mathbf{c}_i^* = (\mathbf{c}_i^\top, g_i(\theta))^\top$  and  $\mathbf{C}^* = (\mathbf{C}^\top, 0)^\top$ , for a given  $\theta$ . Let  $\ell(\hat{m}^*, \theta) = \sum_{i=1}^n \log(\hat{m}_i^*(\theta))$ . The *empirical log-likelihood ratio function* (or deviance) is defined by the following function of  $\theta$ .

$$\hat{r}(\theta) = 2 \{ \ell(\hat{m}) - \ell(\hat{m}^*, \theta) \}.$$

The *maximum empirical likelihood estimate*  $\hat{\theta}$  of  $\theta_0$  is defined by the value of  $\theta$  which minimises the function  $\hat{r}(\theta)$ . As the minimum value of  $\hat{r}(\theta)$  is zero,  $\hat{\theta}$  is the solution of  $\hat{r}(\theta) = 0$ . It can be easily shown that this implies that  $\hat{\theta}$  is the solution of the following estimating equation.

$$\hat{G}(\theta) = 0, \quad \text{with} \quad \hat{G}(\theta) = \sum_{i=1}^n \hat{m}_i g_i(\theta);$$

when  $g_i(\theta) = y_i - n^{-1}\theta\pi_i$ , we have  $\hat{G}(\theta) = \sum_{i=1}^n g_i(\theta)\pi_i^{-1}$  and  $\hat{\theta}$  is the Horvitz-Thompson estimator [6] given by  $\hat{Y}_\pi = \sum_{i=1}^n y_i \pi_i^{-1}$ . When  $g_i = y_i - \theta N^{-1}$ ,  $\hat{\theta}$  is the Hajek [5] ratio estimator  $\hat{Y}_H = N \hat{N}_\pi^{-1} \hat{Y}_\pi$ , where  $\hat{N}_\pi = \sum_{i=1}^n \pi_i^{-1}$ .

## 2.2. Empirical likelihood confidence intervals

Berger and De La Riva Torres [1] show that the random variable  $\hat{r}(\theta_0)$  follows asymptotically a chi-squared distribution with one degree of freedom. Thus, the  $\alpha$  level empirical likelihood confidence interval for the population parameter  $\theta_0$  is given by

$$\{ \theta : \hat{r}(\theta) \leq \chi_1^2(\alpha) \}.$$

Note that  $\hat{r}(\theta)$  is a convex non-symmetric function with a minimum when  $\theta$  is the maximum empirical likelihood estimator. This interval can be found using a bisection search method. This involves calculating  $\hat{r}(\theta)$  for several values of  $\theta$ . Berger and De La Riva Torres [1] showed how this approach can be used to accommodate large sampling fractions and non-response.

It is also possible to calibrate towards parameters more complex than totals. For example, we may want to calibrate with respect to population means, quantiles or variances. In this case, the calibration constraint is specified by the estimating equations  $\sum_{i=1}^n m_i \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0) = \mathbf{0}$ ; where  $\mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0)$  is a vector function of the auxiliary variables and of a known parameter  $\boldsymbol{\vartheta}_0$  which is the solution of the following estimating equation

$$\sum_{i \in U} \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0) = \mathbf{0}.$$

Calibration constraints are taken into account by including the auxiliary variables within the  $\mathbf{c}_i$ . In this case, we use  $\mathbf{c}_i = (\mathbf{z}_i^\top, \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0)^\top)^\top$  and  $\mathbf{C} = (\mathbf{n}^\top, \mathbf{0}^\top)^\top$ . For example, if we want to calibrate towards known population means  $\boldsymbol{\vartheta}_0 = \mathbf{X}N^{-1}$ , we need to use  $\mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0) = \mathbf{x}_i - \boldsymbol{\vartheta}_0$ ; where  $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$  is a vector of known population totals. Simultaneous calibration on totals, means, proportion or any known parameter is also feasible.

### 3. AN R LIBRARY

In order to implement the approach describes in Section 2, we developed a library in R called *emplikfpop*. First, we need to specify the design and calibration variables. Secondly, we need to specify the parameter of interest (i.e. the definition of  $g_i(\theta)$ ) and the finite population correction. This information is needed for point estimation and for confidence intervals.

- i. *Design and auxiliary information:* This information is contained in the matrix `MatDA` and the vector `Var.Labels`. `MatDA` is a  $n \times p$  matrix containing the stratification labels, the  $\pi_i$  and  $\mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\vartheta}_0)$  (depending on  $\boldsymbol{\vartheta}_0$ ). The vector `Var.Labels` gives the columns (in `MatDA`) of stratification and  $\pi_i$ .
- ii. *Definition of the parameter of interest:* A function object `FunctG` defining the function  $g_i(\theta)$ . This function depends on a matrix `Data` and a vector `Vect.Const` which specify the data needed in the definition of  $g_i(\theta)$ . This function also depends on `Theta`, a given value for  $\theta$ . This function has the following format:  
`FunctG = function(Data, Vect.Const, Theta){...}`
- iii. *Finite population corrections:* The  $n \times 1$  vector `fpc` contains the  $1 - \pi_i$  or  $1 - nN^{-1}$  or 1 (if the finite population correction is ignored).
- iv. *Survey weights:* The survey weight  $\hat{m}_i$  are computed using the function  
`> Mi = Vect.Mi(MatDA, Var.Labels)`
- v. *Point estimate:* The point estimate is computed using the following function  
`> Theta.Hat = SolEE(FunctG, Data, Vect.Const, Mi, Min, Max)`  
 where `Min` and `Max` specify the range of  $\hat{\theta}$ .
- vi. *Confidence interval:* The confidence interval is computed using the following function  
`> Bounds = ELBound(MatDA, Var.Labels, FunctG, Level, Data, Vect.Const, Theta.Hat, fpc)`  
 where `Level` is the level of the confidence interval. For example, for the 95% confidence interval, we use `Level = 0.95`. The object `Bounds` contains the bounds of the confidence interval. We have also predefined function for means, totals and quantiles: `ELBoundMean()`, `ELBoundTotal()` and `ELBoundQuantile()`. For the Rao-Hartley-Cochran sampling design, we have the function `ELBound.RHC()`.

#### 4. AN APPLICATION TO THE EU-SILC HOUSEHOLD SURVEY

We use the 2009 EU-SILC user database to estimate the *persistent at-risk-of-poverty rate*. we adopted an ultimate cluster approach, where the units are the primary sampling units. In the table below, we have the point estimate and several confidence intervals for a couple of countries: the empirical likelihood confidence intervals, the standard confidence intervals based on variance estimates and the rescaled bootstrap confidences intervals [8]. Note that the bounds of the standard intervals are negative for Ireland, Austria, Malta and Denmark. The bootstrap bounds and the empirical likelihood bounds are larger than the bounds of the standard intervals. These differences are more pronounced for Austria, Malta, Denmark, the Netherlands, Estonia, Latvia and Greece. This is due to the skewness of the sampling distribution.

**Table 1: Persistent at-risk-of-poverty rate & confidence intervals. 2009 EU-SILC.**

Country	Rate (%)	Emp. Likelihood		Standard		Rescaled Bootstrap	
		Lower	Upper	Lower	Upper	Lower	Upper
Ireland	0.53	0.08	1.76	-0.26	1.31	0.00	1.58
Austria	2.14	0.53	6.50	-0.52	4.80	0.14	5.26
Malta	2.90	0.97	7.75	-0.10	5.89	0.62	6.09
Denmark	3.46	1.09	8.95	-0.06	6.98	0.67	7.76
France	4.50	3.33	5.99	3.21	5.8	3.23	6.04
UK	5.18	2.56	9.90	1.78	8.57	2.15	8.85
Netherlands	5.22	1.88	11.66	0.69	9.75	1.31	10.25
Estonia	7.45	4.07	14.69	2.87	12.03	3.47	13.11
Poland	8.58	5.89	12.49	6.32	10.85	5.32	12.13
Latvia	10.34	6.09	17.36	5.05	15.63	5.36	15.27
Greece	11.34	7.51	18.32	6.72	15.96	7.03	16.95

#### REFERENCES

- [1] Berger, Y. G. and De La Riva Torres, O. (2012) Empirical likelihood confidence intervals for complex sampling designs. S3RI Working paper, <http://eprints.soton.ac.uk/337688/>
- [2] Deville, J. C. and Sarndal (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.
- [3] Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology, 25} 193-203.
- [4] Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics, 31, 1208-1211.
- [5] Hajek, J. (1971) Comment on a paper by D. Basu. in Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston.
- [6] Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.
- [7] Owen, A. B. (2001) Empirical Likelihood. New York: Chapman & Hall.
- [8] Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys. Survey Methodology, 18, 209--217.

[9] Wu, C. and Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34, 359-375.

# Design-based confidence intervals and significance test for regression parameters using an empirical likelihood approach

Melike Oguz-Alper\* ([M.OguzAlper@soton.ac.uk](mailto:M.OguzAlper@soton.ac.uk))<sup>1</sup>, Yves G. Berger ([Y.G.Berger@soton.ac.uk](mailto:Y.G.Berger@soton.ac.uk))<sup>1</sup>

**Keywords:** Design-based inference, estimating equations, nuisance parameter, unequal inclusion probabilities.

## 1. INTRODUCTION

Confidence intervals based on least squares may have poor coverages for regression parameters when the effect of sampling design is ignored. In addition, confidence intervals obtained from the standard design-based approaches [e.g. 1, 2, 3, 4] may not have the right coverages when the sampling distribution is skewed.

We propose to use an empirical likelihood approach to construct design-based confidence intervals and to test hypotheses for regression parameters under unequal probability sampling. Berger and De La Riva Torres [5] proposed an empirical likelihood approach which can be used for point estimation and to construct confidence intervals under complex sampling designs for a single parameter. We show that this approach can be extended to the multidimensional parameter case, in the sense that we can derive confidence intervals and test the significance of a subset of model parameters while taking the sampling design into account. This requires profiling which is not covered by Berger and De La Riva Torres [5].

The proposed approach intrinsically incorporates sampling weights, design variables, and auxiliary information. It may yield to more accurate confidence intervals when the sampling distribution of the regression parameters is not normal, the point estimator is biased, or the regression model is not linear. The proposed approach is simple to implement and less computer intensive than bootstrap. It does not rely on re-sampling, linearisation, variance estimation, or design-effect.

### 1.1. Parameter of interest and estimating equations

Let  $s$  be a random sample of size  $n$  which is selected from the finite population  $U$  of size  $N$  with respect to a probability sampling  $p(s)$ . Let  $y_i$  and  $\mathbf{x}_i$  be some variables of interest. Suppose that  $\psi_N$  is an unknown finite population parameter, which is the solution of the following population estimating equation.

$$G(\psi) = \sum_{i \in U} g_i(y_i, \mathbf{x}_i, \psi) = \mathbf{0},$$

where  $\mathbf{g}_i(y_i, \mathbf{x}_i, \psi)$  is a vector of estimating functions [e.g. 1, 2, 4, 6]. For example, for a simple linear regression, we have  $\mathbf{g}_i(y_i, \mathbf{x}_i, \psi) = \mathbf{x}_i(y_i - \mathbf{x}_i^\top \beta)$ .

We assume that the finite population parameter  $\psi_N$  converges to the model parameter  $\psi_0$ . If  $\hat{\psi}$  is a design-consistent estimator of  $\psi_N$  based on a sample data (see Section 2.1), the estimator  $\hat{\psi}$  is also an estimator of  $\psi_0$ . Assuming that the sampling fraction is negligible, the variability of  $\hat{\psi}$  is driven by the sampling design. Hence, design-based

---

<sup>1</sup> University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, United Kingdom. \* Funded by the Economic and Social Research Council (ESRC), United Kingdom.

confidence intervals proposed in this paper can be viewed as confidence intervals of  $\psi_N$  or  $\psi_0$ .

## 2. EMPIRICAL LIKELIHOOD INFERENCE

We use the *empirical log-likelihood function* given by Berger and De La Riva Torres [5]. It is defined as follows.

$$\ell(m) = \sum_{i \in s} \log(m_i), \quad (1)$$

where the  $m_i$  are unknown scale loads. The empirical log-likelihood function in (1) can be used for the sampling with replacement with unequal probability designs as shown by Hartley and Rao [7]. In this paper, we assume that the sampling fraction is negligible. Hence, the proposed approach is valid under the  $\pi$ ps sampling as  $n/N \rightarrow 0$ .

The *maximum empirical likelihood estimators*  $\hat{m}_i$  maximise the empirical log-likelihood in (1) with respect to the constraints  $m_i \geq 0$  and

$$\sum_{i \in s} m_i \mathbf{c}_i = \mathbf{C}, \quad (2)$$

where the  $\mathbf{c}_i$  and  $\mathbf{C}$  are vectors defined in Section 2.1. We assume that  $\mathbf{c}_i$  and  $\mathbf{C}$  satisfy with a set of regularity conditions given by Berger and De La Riva Torres [5] and the condition  $\|\partial \mathbf{c}_i / \partial \boldsymbol{\lambda}\| = O(1)$ , for all  $i \in s$  and  $\boldsymbol{\lambda} \in \mathbf{\Lambda}$ , where  $\|\cdot\|$  denotes the Euclidean norm,  $O(\cdot)$  defines the order of convergence, and  $\mathbf{\Lambda}$  is a neighbourhood around the true population value  $\boldsymbol{\lambda}_N$ . This condition implicitly implies that the  $\mathbf{c}_i$  are differentiable with respect to  $\boldsymbol{\lambda}$  in a neighbourhood of  $\boldsymbol{\lambda}_N$  [e.g. 1, 6, 8].

Berger and De La Riva Torres [5] showed that the maximum empirical likelihood estimators  $\hat{m}_i$  are given by  $\hat{m}_i = (\pi_i + \boldsymbol{\eta}^\top \mathbf{c}_i)^{-1}$ , where  $\boldsymbol{\eta}$  is such that the constraint (2) is satisfied.

### 2.1. Point estimation

Let  $\ell(\hat{m})$  be the maximum value of the empirical log-likelihood function  $\ell(m)$  under the constraints  $m_i \geq 0$  and (2) with  $c_i = \pi_i$  and  $C = n$ . This implies that  $\hat{m}_i = \pi_i^{-1}$ . Assume that  $\hat{m}_i^*$  maximises  $\ell(m)$  subject to the constraints  $m_i \geq 0$  and  $\sum_{i \in s} m_i \mathbf{c}_i^* = \mathbf{C}^*$  with  $\mathbf{c}_i^* = (c_i, \mathbf{g}_i(y_i, \mathbf{x}_i, \boldsymbol{\psi})^\top)^\top$  and  $\mathbf{C}^* = (C, \mathbf{0}^\top)^\top$ , for a given vector  $\boldsymbol{\psi}$ . The *empirical log-likelihood ratio function* is defined by

$$\hat{r}(\boldsymbol{\psi}) = 2\{\ell(\hat{m}) - \ell(\hat{m}^*(\boldsymbol{\psi}))\}. \quad (3)$$

The *maximum empirical likelihood estimate*  $\hat{\boldsymbol{\psi}}$  of the population parameter  $\boldsymbol{\psi}_N$  is defined by the vector which minimises (3). The minimum value of (3) is obtained when  $\hat{r}(\boldsymbol{\psi}) = 0$ ; that is, when  $\hat{m}_i^* = \hat{m}_i = \pi_i^{-1}$ . Thus, the maximum empirical likelihood estimator of  $\boldsymbol{\psi}_N$  is the solution of the following sample estimating equation.

$$\hat{\mathbf{G}}(\boldsymbol{\psi}) = \sum_{i \in s} \mathbf{g}_i(y_i, \mathbf{x}_i, \boldsymbol{\psi}) \pi_i^{-1} = \mathbf{0}.$$

### 2.2. Hypothesis testing

Let  $\boldsymbol{\psi}_N = (\boldsymbol{\theta}_N^\top, \boldsymbol{\lambda}_N^\top)^\top$  where  $\boldsymbol{\theta}_N$  is a  $p \times 1$  vector of parameters of interest and  $\boldsymbol{\lambda}_N$  is a

$q \times 1$  vector of parameters which are not of primary interest. Suppose we wish to test  $H_0 : \theta_N = \theta_N^0$ . Consider the *profile empirical log-likelihood ratio function* defined by

$$\hat{r}(\theta_N^0) = 2\{\ell(\hat{m}) - \max_{\lambda} \ell(\hat{m}^*(\theta_N^0, \lambda))\}, \quad (4)$$

where the set of  $\hat{m}_i^*$  maximises  $\ell(m)$  subject to the constraints  $m_i \geq 0$  and  $\sum_{i \in s} m_i \mathbf{c}_i^* = \mathbf{C}^*$  with  $\mathbf{c}_i^* = (\pi_i, \mathbf{g}_i(y_i, \mathbf{x}_i, \theta_N^0, \lambda)^\top)^\top$  and  $\mathbf{C}^* = (n, \mathbf{0}^\top)^\top$ . Note that in (4), we maximise  $\ell(\hat{m}^*(\theta_N^0, \lambda))$  over the parameter  $\lambda$  for a given value of  $\theta_N = \theta_N^0$ .

Under  $H_0$ , it can be shown that the profile empirical log-likelihood ratio function  $\hat{r}(\theta_N^0)$  given by (4) follows asymptotically a *chi-squared distribution* with a  $p$  degree of freedom. Based on this, we can compute the *p-value*. Note that lack of fit would not affect the performance of the proposed empirical likelihood test [e.g. 8].

### 2.3. Confidence region

We can obtain confidence region for each parameter individually profiling out over the other parameters. In this case,  $p = 1$  and we have the scalar  $\theta_N$  instead of the vector  $\theta_N$ . Then, based on the asymptotic chi-squared distribution of  $\hat{r}(\theta_N^0)$  under the null hypothesis  $H_0 : \theta_N = \theta_N^0$ , the  $(1 - \alpha)\%$  empirical likelihood confidence region for  $\theta_N$  is given by the set  $\{\theta : \hat{r}(\theta) \leq \chi_{df=1}^2(\alpha)\}$ , where  $\chi_{df=1}^2(\alpha)$  is the upper  $\alpha$  - *quantile* of the chi-squared distribution with one degree of freedom.

## 3. RESULTS

We present some numerical results for a linear regression model with one intercept and one slope. We generated the Hansen, Madow and Tepping (HMT) population [see 9]. The population size is  $N = 10\,000$ . We selected 1000 random samples of size  $n = 500$  from this population using the randomised systematic sampling with unequal probabilities.

The linear regression model of interest is defined by  $y_i = \lambda + \theta x_i + u_i$ , where the  $u_i$  are independent random variables with  $\text{var}(u_i | x_i) \propto x_i^{3/2}$ . The parameter of interest is the slope  $\theta$ . We profile out over the intercept  $\lambda$  when minimising (4).

Table 1 gives the observed coverages of the 95% confidence intervals constructed based on several methods. We considered two Pseudo likelihood approaches which are given by Binder and Patak [2] and Godambe and Thompson [4] [see also 1].

Standard confidence intervals are based on the normality of the point estimator. Note that, when the sampling distribution is skewed, the normality assumption may not hold. This explains the poor coverages of the Wald and the pseudo likelihood 1 approaches (see Table 1). The poor coverage of the Wald type of confidence intervals is also due to the fact that this method ignores the sampling design. We have an overcoverage with the rescaled bootstrap [e.g. 10]. Moreover, it has the largest confidence intervals on average compared to the other methods (see the ratio of average lengths in Table 1).

The coverage probabilities of the empirical likelihood and the pseudo likelihood 2 confidence intervals are not significantly different from the nominal level (i.e. 95%). However, the former is more reliable than the latter with regards to the standard deviation of length (see the last column of Table 1).



#### 4. CONCLUSIONS

We proposed an empirical likelihood approach which can be used to make inferences for regression parameters incorporating the sampling design. The proposed approach can be used for generalised linear models.

It can be easily shown that the population level information can be taken into account with the proposed approach. Unlike the usual calibration approach [11], the proposed approach can be used for testing and constructing confidence intervals. Moreover, the auxiliary information does not have to be in the form of totals or means [5].

The proposed approach can be easily extended to stratified sampling designs by incorporating the strata information into the  $c_i$ .

**Table 1. Observed coverages of the 95% confidence intervals for the slope  $\theta_N$ .**

N=10 000, n=500	Coverage probability	Lower error	Upper error	Ratio average length	Ratio SD length
Wald	76.6*	23.8*	0.1*	0.96	0.53
Empirical likelihood	94.8	3.1*	2.1*	1.00	1.00
Pseudo likelihood 1	94.0*	3.5*	2.5	0.97	1.07
Pseudo likelihood 2	94.8	3.3*	1.9*	0.99	1.09
Rescaled bootstrap	96.5*	2.4	1.1*	1.05	0.91

\* Significantly different from the nominal levels (95% and 2.5% for coverage probability and tail errors respectively) at the 5% significance level (i.e.  $p - value \leq 0.05$ ).

#### REFERENCES

- [1] D. A. Binder, On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, (1983), 279–292.
- [2] D. A. Binder and Z. Patak, Use of estimating functions for estimation from complex surveys, *Journal of the American Statisticsl Association*, 89, (1994), 1035–1043.
- [3] J. C. Deville, Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, 25, (1999), 193–203.
- [4] V. P. Godambe and M. E. Thompson, Estimating functions and survey sampling, *Handbook of Statistics: Design, Method and Applications*: D. Pfeiffermann and C.R. Rao.(editors), Elsevier, 29B, (2009), 83–101.
- [5] Y. G. Berger and O. De La Riva Torres, *Empirical likelihood confidence intervals for complex sampling designs*. S3RI, <http://eprints.soton.ac.uk/337688>, (2012).
- [6] J. Qin and J. Lawless, Empirical likelihood and general estimating equations, *The Annals of Statistics*, 22, (1994), 300–325.
- [7] H. O. Hartley and J. N. K. Rao, *A new estimation theory for sample surveys*, II. Wiley-Interscience, New York, 1969.
- [8] A. B. Owen, *Empirical Likelihood*, Chapman & Hall, New York, 2001.
- [9] M. H. Hansen, W. G. Madow, and B. J. Tepping, An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, 78, (1983), 776–793.

- [10] J. N. K. Rao, C. F. J. Wu, and K. Yue, Some recent work on resampling methods for complex surveys, *Survey Methodology*, 18, (1992), 209–217.
- [11] J. C. Deville and C. E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, (1992), 376–382.

# A Multivariate Regression Estimator for Rotating Sampling Surveys

Karen Caruana ([kc12g13@soton.ac.uk](mailto:kc12g13@soton.ac.uk))<sup>1</sup> and Yves G. Berger ([y.g.berger@soton.ac.uk](mailto:y.g.berger@soton.ac.uk))<sup>1</sup>

**Keywords:** Design-based approach, multivariate regression, regression estimator, calibration, correlation

## 1. INTRODUCTION

Longitudinal surveys collect information on several occasions, or time points [1], [2]. Consider that we have two occasions or waves labelled 1 and 2. The samples selected on occasions 1 and 2 are rarely completely overlapping samples, as not all the units are selected on both occasions. It is common practice to have a large fraction of units sampled at both occasions. Surveys which have this feature are called rotating sampling surveys.

The customary point estimators are the Horvitz Thompson [3] and generalised regression [4] estimators of a total or a mean. We propose a new regression estimator for cross-sectional totals and change between totals. This estimator uses the information from both occasions simultaneously instead of each occasion separately. This estimator incorporates the auxiliary variables similar to the general regression estimator and the sample design variables specifying the rotating sampling design. The proposed estimator is multivariate because it combines the auxiliary information from the first and second occasion.

Longitudinal surveys are used to monitor change between population target parameters. For social policy makers, the estimation of change over time of social indicators as such youth employment rate, literacy rate and social deprivation indicators may be as important as cross-sectional indicators. The variance of change, for rotating sampling surveys, is a challenging subject since it requires to estimate correlations. Several authors proposed different estimators for correlations [5], [6], [7], [8] and [9]. A variance of change is proposed by extending the estimator proposed by [9] where besides the design variables, the auxiliary variables are included.

In the simulation study, the proposed estimator is compared with the Horvitz Thompson (HT) and generalised regression estimators. The relative bias and ratio of relative mean square errors are computed for the estimator of totals. We consider different correlations between the response variables and the auxiliary variables.

## 2. METHODS

In rotating sampling designs, a fixed proportion of sample units are replaced by new units at each wave. Each unit remains in the sample for the same number of waves [2].

Let  $s_1$  and  $s_2$  be the probability samples for the first occasion (selected from population  $U_1$ ) and for the second occasion (selected from population  $U_2$ ) respectively. Let  $s_{12}$  be

---

<sup>1</sup> University of Southampton, UK

the sample of units that are both in  $s_1$  and  $s_2$ . Suppose that the sample size is fixed for both occasions. We consider that  $s_1$  is composed of  $n_1$  units with first-order inclusion unequal probabilities  $\pi_{1,i} = pr\{i \in s_1\}$ , where  $pr\{\cdot\}$  denotes the probability with respect to the design. Similarly,  $s_2$  is composed of  $n_2$  units. The  $n_2$  units are selected with conditional inclusion unequal probabilities  $\pi_{2,i}(s_1) = pr\{i \in s_2 | s_1\}$  which are such that  $n_c$  units are contained in  $s_c$ ; where  $s_c = s_1 \cap s_2$ . Thus, the second wave inclusion probabilities are given by  $\pi_{2,i} = E_1[\pi_{2,i}(s_1)]$ ; where  $E_1[\cdot]$  denotes the design expectation with respect to the first wave design. Finally, assume that for both waves, the sampling fractions are negligible; that is,  $1 - \pi_{l,i} \approx 1$ .

Let  $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)^T$ ; where  $\mathbf{y}_l = (y_{l,1}, y_{l,2}, \dots, y_{l,n_l})^T$ , be the responses for the variable of interest for wave  $l = 1, 2$ . Define  $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2)^T$  where  $\tilde{\mathbf{y}}_l = (y_{l,1} \pi_{l,1}^{-1}, y_{l,2} \pi_{l,2}^{-1}, \dots, y_{l,n_l} \pi_{l,n_l}^{-1})^T$ . Let  $\hat{\mathbf{y}}$  be the vector of the HT estimators of the response variables  $\mathbf{y}_1$  and  $\mathbf{y}_2$ :

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2)^T,$$

where  $\hat{y}_l = \sum_{i=1}^{n_l} y_{l,i} \pi_{l,i}^{-1}$ .

Assume that  $J$  auxiliary variables are available for both waves. The vector of auxiliary variables of the  $k^{th}$  element of wave  $l$  is defined as:  $\mathbf{x}_{l,k} = (x_{l,1;k}, x_{l,2;k}, \dots, x_{l,J;k})^T$ . Let  $\tilde{\mathbf{X}}_l = [\tilde{\mathbf{x}}_{l,1}, \tilde{\mathbf{x}}_{l,2}, \dots, \tilde{\mathbf{x}}_{l,n}]$  where  $\tilde{\mathbf{x}}_{l,i} = (x_{l,1;k} \pi_{l,1}^{-1}, x_{l,2;k} \pi_{l,2}^{-1}, \dots, x_{l,J;k} \pi_{l,J}^{-1})^T$ . Let  $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$  be the  $(2J \times 1)$  vector of population totals of the auxiliary variables, where  $\mathbf{x}_l = (\sum_{i \in U_l} x_{l,1;i}, \sum_{i \in U_l} x_{l,2;i}, \dots, \sum_{i \in U_l} x_{l,J;i})^T$  and the corresponding HT estimator vector is  $\hat{\mathbf{x}} = (\hat{\mathbf{x}}_1^T, \hat{\mathbf{x}}_2^T)^T$ ;  $\hat{\mathbf{x}}_l = (\sum_{i \in s_l} x_{l,1;i} \pi_{l,1}^{-1}, \sum_{i \in s_l} x_{l,2;i} \pi_{l,2}^{-1}, \dots, \sum_{i \in s_l} x_{l,J;i} \pi_{l,J}^{-1})^T$ .

Let the design variables be  $z_{1,i} = \delta\{i \in s_1\}$  and  $z_{2,i} = \delta\{i \in s_2\}$ , where  $\delta\{A\}$  is one when  $A$  is true and zero otherwise. Let define the matrix of the design variables as

$$\mathbf{Z}_s = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_c)^T$$

where  $(z_{l,1}, z_{l,2}, \dots, z_{l,n_{12}})^T$  and  $\mathbf{z}_c = (z_{1,1}z_{2,1}, z_{1,2}z_{2,2}, \dots, z_{1,n_{12}}z_{2,n_{12}})^T$ ;  $n_{12} = n_1 + n_2 - n_c$ . Define  $\hat{\mathbf{y}}_s = (\hat{\mathbf{x}}^T, \hat{\mathbf{z}}_s^T)^T$  and  $\mathbf{y}_U = (\mathbf{x}^T, \mathbf{z}_U^T)^T$  as two  $(2J + 3 \times 1)$  vectors; where  $\hat{\mathbf{z}}_s = (\sum_{i \in s_1} z_{1,i}, \sum_{i \in s_2} z_{2,i}, \sum_{i \in s_c} z_{c,i})^T = (n_1, n_2, n_c)^T = \mathbf{z}_U$

The proposed multivariate generalised regression estimator is:

$$\hat{\mathbf{y}}_s^{(PROP)} = \hat{\mathbf{y}} + (\mathbf{y}_U - \hat{\mathbf{y}}_s)^T \hat{\boldsymbol{\beta}}_{XZ},$$

where  $\hat{\boldsymbol{\beta}}_{XZ} = (\tilde{\mathbf{C}}_s^T \tilde{\mathbf{C}}_s)^{-1} (\tilde{\mathbf{C}}_s)^T \tilde{\mathbf{y}}$ ;  $\tilde{\mathbf{C}}_s = (\tilde{\mathbf{X}}, \mathbf{Z}_s)$ ,  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ .

The multivariate regression estimator of change  $\Delta = \sum_{i \in U_1} y_{1,i} - \sum_{i \in U_2} y_{2,i}$  is given by

$$\hat{\Delta} = (1, -1) \hat{\mathbf{y}}_s^{(PROP)}.$$

The proposed variance of change is based upon [9] where the design variables and the auxiliary variables are included.

### 3. SIMULATION RESULTS

For the simulation study, we consider a population of  $N = 20,000$  units. The sample size is the same for both waves,  $n_1 = n_2 = 200$  and the number of sampling units common for both waves is  $n_c = 120$ . The population is generated from a multivariate (i) normal distribution and (ii) lognormal distribution. 1,000 samples are selected using a random systematic sampling where the probabilities are unequal without replacement. We consider several correlations between the response variables and the auxiliary variables.

The proposed (PROP) estimator is compared with the HT and generalised regression (GREG) estimators. The relative bias (RB) and ratio of relative mean square error (RRMSE) are computed for the point estimators, cross-sectional variance and the variance of change. Tables 1 and 2 below shows the results for the data generated from the two different distributions considered.

The RB and RRMSE of the proposed point estimator is always smaller than the HT and GREG estimator. The RRMSE of the variance estimators are of a comparable order for normal distributions (see Table 1). With a log-normal distribution (see Table 2), the standard GREG estimator has the smallest RRMSE for the variance. We observe a small RB for the variance estimator for change of the proposed estimator.

Table 1: Results from data generated from a multivariate normal distribution

Correlation		RB			RRMSE		
		HT	GREG	PROP	HT	GREG	PROP
$\delta_{YY} = 0.2$ ,	$\hat{y}_1$	0.03	-0.01	-0.01	1.77	1.59	1.59
$\delta_{YX} = 0.2$ ,	$\hat{y}_2$	0.09	0.05	0.05	1.70	1.52	1.51
$\delta_{XX} = 0.2$ .	$\widehat{var}(\hat{y}_1)$	-2.61	3.27	-3.47	11.08	11.02	10.69
	$\widehat{var}(\hat{y}_2)$	3.61	5.60	3.61	11.40	12.23	11.64
	$\widehat{var}(\hat{\Delta})$	11.55	0.99	-2.21	15.56	9.05	9.21
$\delta_{YY} = 0.8$ ,	$\hat{y}_1$	0.09	0.02	0.00	1.77	1.07	1.00
$\delta_{YX} = 0.8$ ,	$\hat{y}_2$	0.09	0.05	0.04	1.73	1.17	1.01
$\delta_{YX} = 0.8$ .	$\widehat{var}(\hat{y}_1)$	-3.28	-1.14	-1.82	11.33	10.63	10.80
	$\widehat{var}(\hat{y}_2)$	0.23	3.47	-1.25	10.22	10.74	10.28
	$\widehat{var}(\hat{\Delta})$	39.45	11.76	0.76	42.32	15.68	9.09

Table 2: Results from data generated from a multivariate lognormal distribution

Correlation		RB			RRMSE		
		HT	GREG	PROP	HT	GREG	PROP
$\delta_{YY} = 0.2$ ,	$\hat{y}_1$	-0.20	0.04	-0.14	3.92	4.53	3.84
$\delta_{YX} = 0.2$ ,	$\hat{y}_2$	0.05	0.18	0.16	3.67	4.41	3.68
$\delta_{XX} = 0.2$ .	$\widehat{var}(\hat{y}_1)$	-4.46	-4.55	-6.88	20.93	17.62	20.90
	$\widehat{var}(\hat{y}_2)$	10.24	1.54	3.00	23.50	17.00	20.64
	$\widehat{var}(\hat{\Delta})$	8.15	-9.58	-0.54	25.57	15.59	15.63
$\delta_{YY} = 0.8$ ,	$\hat{y}_1$	-0.20	0.05	0.00	3.92	2.44	2.36
$\delta_{YX} = 0.8$ ,	$\hat{y}_2$	-0.07	0.36	0.05	3.68	2.39	2.28
$\delta_{YX} = 0.8$ .	$\widehat{var}(\hat{y}_1)$	-4.46	-0.26	-5.97	20.93	16.73	20.08
	$\widehat{var}(\hat{y}_2)$	8.53	3.39	0.43	22.04	16.94	18.78
	$\widehat{var}(\hat{\Delta})$	79.57	-6.58	-3.07	86.55	14.37	15.79

#### 4. CONCLUSIONS

We proposed a multivariate regression estimator that exploits the information from both waves simultaneously instead of each wave separately. This estimator besides using the auxiliary variables, also incorporates the sample design variables.

The simulation study shows that the RRMSE of the proposed point estimator is always smaller than the classical Horvitz-Thompson and generalised regression estimator. With respect to the RRMSE of the variance and variance of the change of the proposed estimator is similar to the other estimator. The variance of change of the proposed estimator has a small relative bias.

#### REFERENCES

- [1] Kalton, G. and Citro, C. F. (1995) Panel surveys: adding the fourth dimension. *Innovation: The European Journal of Social Science Research*, 8, 25 -39.
- [2] Lynn, P. (2009) *Methodology of longitudinal surveys*. John Wiley & Sons.
- [3] Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- [4] Sarndal, C., Swensson, B. and Wretman, J. (1992) *Model assisted survey sampling*. Springer-Verlag.
- [5] Kish, L. (1965) *Survey sampling*.
- [6] Tam, S. (1984) On covariances from overlapping samples. *The American Statistician*, 38,
- [7] Qualite, L. and Tille, Y. (2008) Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 173 - 181.
- [8] Wood, J. (2008) On the covariance between related Horvitz Thompson estimators. *Journal of Office Statistics*, 24, 53.
- [9] Berger, Y. G. and Priam, R. (2015) A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys. To appear in the *Journal of Royal Statistical Society, Series A*.

# Estimation from Contaminated Multi-Source Data Based on Latent Class Models

Ugo Guarnera ([guarnera@istat.it](mailto:guarnera@istat.it))<sup>1</sup>, Roberta Varriale ([varriale@istat.it](mailto:varriale@istat.it))<sup>1</sup>

**Keywords:** multi-source data, data integration, contamination models

## 1. INTRODUCTION

In recent years, statistical analysis based on different data-sources has become an active area of research in both theoretical and applied statistics. In particular, due to the increasing availability of administrative data, problems concerning the use of multiple sources for estimation purposes have been receiving an increasing attention in Official Statistics. Frequently National Statistical Institutes (NSIs) try to combine data from available sources in order to build “statistical” archives to be used in different phases of the statistical production process. Massive use of “external” data is being considered by NSIs as an important alternative to the traditional approaches based on survey data. In fact, this approach allows NSIs to move resources previously allocated in conducting surveys to other activities, reducing at the same time the response burden on respondents. Moreover, statistical analysis based on large datasets may result in more accurate estimates than the ones that can be obtained through sample surveys. On the other hand, combining data to build a statistical information system is a complex task. In fact, administrative data are typically collected by different institutions for specific purposes (for instance, data on enterprises provided by the tax agency have “fiscal nature”) and may not be usable in their original form for statistical purposes. Thus, a lot of “pre-processing” work has to be done in activities (such as harmonization of definitions, variable standardization, etc.) aiming at providing users with data that satisfy their informative requirements. Another important issue is related to the possibility of partial (or total) overlapping among informative contents from different sources. This is of course of no concern when differences among values of corresponding items are negligible, but problems can arise when, due to “measurement errors”, strong discrepancies are observed. In the latter case some decision strategy is necessary. A possible approach is to rely on a “hierarchy” based on preliminary analyses of the data quality of each source. In presence of discordant values, the source with the “highest” score according to the established hierarchy is chosen. Getting information from a single source for each statistical unit has the advantage of preserving coherence among different items. This approach has been used for instance at the Italian Institute of Statistics (ISTAT) to build a statistical information system for annual Structural Business Statistics on small and medium enterprises [1]. The problem with the hierarchical approach is that it is not obvious how to define the hierarchy among sources. Moreover, in some situations, information from sources with low score in the hierarchy could be used when not plausible values or missing values are observed in the highest quality source. These observations suggest an alternative approach where one takes advantage of the simultaneous availability of information from different sources. According to this approach, all the available information is used and “weighted” according to its reliability. In this paper, a model for the prediction of “true” values of some numeric variable of interest conditional on *all* the available information is presented. The true values of the target variable are viewed as realizations from a latent (unobserved) variable and the

---

<sup>1</sup> Istat – Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Rome

distinct (possibly coinciding) observed values from different sources are considered as imperfect measurements of this latent variable. Given a model for the true data and a measurement error model for each available source (through the specification of a conditional distribution of the data observed in the source given the true unobserved data), one can easily derive, via Bayes formula, the distribution of the true data given the observed data. In the next section, details are provided.

## 2. THE MODEL

### 2.1. The true data model

Let us assume that the true unobserved data are realizations from  $n$  iid random Gaussian variables  $Y_i^*$ , with mean  $\mu_i$  and common variance  $\sigma^2$  ( $i=1,..n$ ). We also allow for the possibility of a linear dependence of the means  $\mu_i$  on some set of  $q$  covariates  $x_i = (x_{i0}, x_{i1}, ..., x_{iq})'$  observed without error, i.e., we assume the relation  $\mu_i = \beta' x_i = \beta_0 x_{i0} + \beta_1 x_{i1} + ... + \beta_q x_{iq}$ , where  $\beta_j$  ( $j=0,..q$ ) are unknown coefficients to be estimated and, as usually, we put  $x_{i0} \equiv 1$ . Thus, true data are modeled via the ordinary linear regression model:

$$1) \quad Y_i^* = \beta' x_i + U_i, \quad i=1,..n,$$

where  $U_i$  are iid Gaussian variables with zero mean and variance  $\sigma^2$ . In real applications on economic data, logarithms of data instead of data in their original scale are often assumed to be normally distributed. This does not imply substantial changes in the proposed methodology.

### 2.2. The error model

Assume that the variable of interest is observed with error in  $G$  sources (for instance administrative archives) and let  $Y_i^g$  be the variable corresponding to the value observed in the source  $g$  for the unit  $i$  ( $i=1,..n$ ;  $g=1,..G$ ). In order to complete the modeling, we have to specify the measurement error model for each source, that is, the conditional distribution of  $Y_i^g$  given the true value  $y_i^*$ . We assume an “intermittent” error mechanism reflecting the fact that only a fraction of the available data are affected by errors, or, in other words, that data are only partially contaminated. This assumption naturally leads to the adoption of contamination models for the observed data. Contamination models have been largely used to detect outliers or influential errors in statistical data [2],[3] available from a single data source. In detail, we model the intermittent nature of the error on the different data sources via Bernoullian variables  $Z_i^g$  with parameters  $\pi_g$ , i.e.,  $Z_i^g = 1$  if an error occurs for the unit  $i$  in the source  $g$ , or in other words, if  $Y_i^g \neq Y_i^*$  and zero otherwise. Also, given the event  $\{Z_i^g = 1\}$  (presence of error in source  $g$ ), we assume that  $Y_i^g = Y_i^* + \varepsilon_i^g$  where  $\varepsilon_i^g$  is supposed to follow a Gaussian distribution with zero mean and variance  $\alpha_g \sigma^2$  where  $\alpha_g$  is a scalar constant ( $g=1,..,G$ ). In short, the measurement error model can be described through the equation:

$$2) \quad Y_i^g = Y_i^* + Z_i^g \varepsilon_i^g \quad g=1,..,G; \quad i=1,..n.$$



Equations 1) and 2) completely specify the assumed model. We note that the parameters  $(\pi_g, \alpha_g)$  can be thought of as quality indicators for the source  $g$ , representing, respectively, the (a priori) error probability and the effect of the error (variance inflation).

### 2.3. Estimation

From the above model assumptions it follows that the distribution  $f(y_i) = f(y_i^1, \dots, y_i^G)$  of the random vector  $Y_i = (Y_i^1, \dots, Y_i^G)$  associated with the measures of the target variables from the different sources is a mixture of probability distributions corresponding to the different errors patterns across the sources. Formally:

$$3) \quad f(y_i) = \sum_{k=1}^{2^G} w_k h_k(y_i; \beta, \sigma^2, \alpha), \quad \alpha \equiv \alpha_1, \dots, \alpha_G, \quad \beta \equiv \beta_0, \dots, \beta_q,$$

where the sum is over the  $2^G$  error patterns, and for the  $k$ th pattern the “mixing weight”  $w_k$  is the product of  $G$  factors of the form  $\pi_g$  or  $1 - \pi_g$  depending on whether the pattern  $k$  corresponds to an erroneous or correct value in the source  $g$ . The densities  $h_k$  in 3) are suitable products of Gaussian distributions possibly degenerated in mass points. It is important to note that the problem of associating the observations with the different error patterns is partially super-visioned, in that assignment is without uncertainty whenever at least two values across the different data-sources coincide. Based on the observed-data distribution, we can estimate the model parameters  $\theta \equiv \beta_j, \sigma^2, \pi_g, \alpha_g$  ( $j=0, \dots, q$ ;  $g=1, \dots, G$ ) to be used for assessing the quality of the different sources and making predictions on the true values  $y_i^*$ , conditional on a set of simultaneous observations of the target variable on the available sources. We have implemented an appropriate Expectation Maximization (EM) algorithm for the maximum likelihood estimation (MLE) of  $\theta$ . The estimated parameters are plugged in the distribution of the true data conditional on the observed data to obtain predictions of true values in all cases where no equalities are observed among the different measures. Experiments on both simulated and real data show good performances of the method. In particular, the estimates of the measurement error model agree with the quality assessments of the subject matter experts. In order to make the methodology usable also in the (frequent) situations where data are not always available from every source, the algorithm has been extended to incorporate also the incomplete observations in the estimation process. Although this requires working with combinations of error and missing patterns, the extension is quite straightforward and does not imply excessive increase of computation time.

## 3. RESULTS

In this section, an evaluation of the proposed methodology based on a Monte Carlo (MC) study is presented. At each MC iteration, a sample of  $n=1000$  “true data”  $y_i^*$  ( $i=1, \dots, n$ ) has been generated in logarithmic scale according to the regression model 1) with two  $x$  variates. The corresponding regression parameters are:  $\beta_0 = \beta_1 = 1, \beta_2 = 2, \sigma^2 = 0.25$ . Three measurement processes (i.e., three data sources) have also been simulated according to 2) where  $G=3$ ,  $(\pi_1, \pi_2, \pi_3) = (0.2, 0.3, 0.4)$ , and  $(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 8)$ . This implies that the best data source is the first one, while the third source provides the least reliable information. Finally, missing values have been randomly introduced in the three sources according to the missing rates 0.50, 0.10, 0.20 respectively. At each MC run, the estimates of the model parameters have been obtained and used to estimate the posterior

probabilities corresponding to the different error patterns on each unit. Moreover, predictions of true values conditional on the available information have also been computed for each unit. Three methods have been compared for building a single set of micro-data. For all the methods, the “obvious” cases corresponding to at least two coinciding sources have been preliminary treated considering the repeated values as true. The remaining micro-data have been determined as follows. With the first method (*hierarchical*) the values are chosen based on the source hierarchy: thus, the first source is always chosen whenever it is available, while, in cases where  $Y_i^1$  is missing,  $Y_i^2$  is always preferred to  $Y_i^3$ , and the third source is used only if it is the only available source. With the second method the source is chosen where the reported value has the “highest” (estimated) posterior probability of being error-free. Finally, the third method is based on the model prediction, i.e., the micro-data are the expectation of true data conditional on data observed from the different sources. In all approaches, predictions for units where no source is available have been obtained by simply regressing the  $Y$  variable on the (always observed)  $X$  variates. We assume that the target quantity is the population mean. For each method, the relative root mean square error has been estimated by averaging over 500 MC iterations. Results are reported in Table 1.

**Table 1. RSSME for methods based on source hierarchy (*hier*), posterior probabilities (*pp*), and model predictions (*pred*)**

<i>hier</i>	<i>pp</i>	<i>pred</i>
0.30	0.24	0.07

As expected the best performances are provided by the method that uses the model predictions to impute micro-data, while the worst method is the one based on the fixed source hierarchy.

At the moment, methodology and its software implementation have been fully developed in the univariate case. Multivariate extensions imply more complex procedures for the likelihood maximization and are under investigation.

## REFERENCES

- [1] O. Luzi, M. Di Zio, F. Oropallo, A. Puggioni, R. Sanzo, Integrating administrative and survey data in the new Italian system for SBS: quality issues, Paper presented at *The 3rd European Establishment Statistics Workshop*, 9-11 Sep, 2013. Nuremberg, Germany,
- [2] B. Ghosh-Dastidar, J.L. Shafer, outlier Detection and Editing Procedures for Continuous Multivariate Data, *Journal of Official Statistics* (2006) Vol. 22, No. 3, 487-506.
- [3] M. Di Zio, U. Guarnera, A contamination model for Selective Editing, *Journal of Official Statistics* (2013) Vol. 29, No. 4, 539-555.

# A web-semantic data-warehouse approach to the compilation of national accounts: a test case on European National Accounts

Francois Libeau ([francois.libeau@hendiplan.com](mailto:francois.libeau@hendiplan.com))<sup>1</sup>  
Roberto Barcellan ([roberto.barcellan@ec.europa.eu](mailto:roberto.barcellan@ec.europa.eu))<sup>2</sup>  
Bo Sundgren<sup>3</sup> Dominique Ladiray<sup>4</sup>; Boris Motik<sup>5 6</sup>

**Keywords:** Holistic Integration of Information Systems, National Accounts in the ESS, Semantic Web, Production of Statistics, Service Oriented Architecture (SOA).

## INTRODUCTION

In the aftermath of the financial and economic crisis, most national statistical offices and international statistical institutions started a modernisation process to address the gaps highlighted by the crisis and to be ready to match the new evolving users' requirements. The modernisation of the European Statistical System (ESS) has been focussing, in a first phase, on the production method of European statistics, as described in [1]. Key elements of the modernisation of production methods are:

- Get quick and ad hoc information, often across domains, to face major events (e.g. crises);
- Decrease response burden by making data widely available in a reliable way;
- Implement ICT methods and tools to increase efficiency.

These goals have been identified, among others, to gradually replace the existing inefficient stovepipe model for the production of statistics (thematic oriented production process). The stovepipe model proved lacking standardisation.

In the context of the production of European statistics, the modernisation of production processes has therefore consequences

- at the level of Member States, where a holistic integration of information systems is becoming necessary, throughout the entire production chain, from sources (with the inclusion of, for example, administrative data) till dissemination;
- at the EU level, where the need to foster the implementation of collaborative networks in the community of official statistics producers is becoming also necessary.

This trend is expected to have deep consequences on the technical infrastructure harbouring the statistical production processes of European statistics.

---

<sup>1</sup> Hendyplan

<sup>2</sup> European Commission - Eurostat

<sup>3</sup> Stockholm University

<sup>4</sup> INSEE

<sup>5</sup> Oxford University

<sup>6</sup> The views expressed are the authors' alone and do not necessarily correspond to those of the corresponding organisations of affiliation.

In 2010, Eurostat launched a project to explore the methodological and technical aspects of a concrete application of the modernisation approach philosophy, as outlined in the Vision [1], to a statistical domain: the production of European National Accounts.

## **METHODS**

The compilation of European National Accounts in a holistic set of integrated Information Systems requires a good understanding of the key conditions and components underlying it: wide-scales exchanges and flexible technological infrastructure. The corresponding network of information systems has to host a powerful information retrieval mechanism to collect and combine the information necessary to compile the European National Accounts. Therefore, standards play a major role in such network as they frame the architecture, foster harmonisation and facilitate communication. A metadata driven approach relying on common agreed standards is then the starting point for the integrated production method explored in the study.

The project has been articulated in three phases: (i) definition of the standards and the architecture; (ii) assessment of the degree of maturity of existing national accounts systems with respect to the standards; (iii) proof of concept and design of architecture.

The first step in the project corresponded to the identification of the appropriate standards underlining the architecture. Three national accounts production systems have been chosen as case-studies, being sufficiently representative of the diversity in the ESS:

- The Swedish national accounts system, as representative of a quite advanced statistical production system;
- The French national accounts system, because of its many specificities and consolidated tradition;
- Eurostat main aggregates, as it represents the typical aggregation exercise EU-wide.

The experimentation phase assessed the three production processes against the standard (GSBPM – generic statistical business process model, see [2]), with the aim of translating the production processes in GSBPM phases and sub-processes - proof of concept for the description of the process and identification of common parts in the process.

In the light of the results of the experimentation phase, the Team developed a proposal for a solution: the design of the architecture.

### **Experimentation phase**

The authors have defined and implemented a protocol of experimentation starting from the ESS/EA (European Statistical System / Enterprise Architecture) model – (as available in 2011), considering the three layers governance, information, solution. In practice, only the last two layers were relevant in the framework of this study: the governance layer was considered as acquired.

The pluri-disciplinary team (national accountant, IT specialist, web-semantic expert) mapped the National Accounts production process according to different conceptual words and paradigms (see [3]): in this area, an actor (an object), belonging to a certain sector, performs an activity (consumption/investment/production). Activities and

purposes are then classified according to the international standards (e.g. NACE, COICOP, COFOG, ...).

Regarding the processes, the study demonstrated that the GSBPM is a powerful standard to outline the communication between holistic information systems throughout the ESS. It realises a strong and deep harmonization of production systems.

### Design of the architecture phase

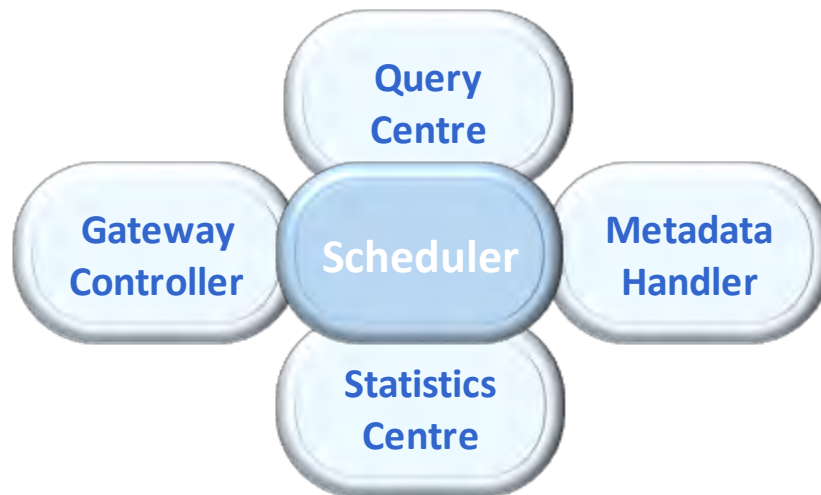
The design of the architecture has been conceived on the basis of the experience of the authors or the results of the experimentation. The use of other technical standards greatly facilitated this phase:

- Service Oriented Architecture (SOA)
- Standard Data and Metadata eXchange (SDMX).

The team also kept in mind the usage of standard components which already exist in the field, promoted by Eurostat and many ESS members: eDAMIS, building blocks such as JDemetra (tool for seasonal adjustment), etc. The experience gained also via the reengineering of the Eurostat national accounts production system (NAPS phase 1 project) (see [4]) was also a particularly good substratum for this study.

### RESULTS

The authors proposed a combination of modules interconnected as illustrated in Fig. 1.



**Figure 1. Interconnected modules**

The solution is widely SOA-Based with the capacity to be replicated in as many production sites as necessary, fostering the cloud computation of the information.

The proposal is based on a collection of modules under the control of a Scheduler, the director of the system, responsible for the automation. The Scheduler is a state machine from the IT point of view, used in the control of industrial processes, dispatching flows.

The Scheduler is making the requests circulating across agents, activating them according to topics and stage of computation. These agents are:

This is the first attempt to do this type of analysis by using the linkage. It is important to develop a timelier framework – optionally a quarterly one – which would provide timely this type of distributional analysis. The full version of this paper presents time series for these indicators. This attempt shows that this framework can potentially provide interesting results. However, the performed sensitivity analysis showed that more data and additional analysis of this framework is needed.

The fundamental challenge of this framework is the assumption that at the instrument level different income quintiles follow the average investment and price development. The sensitivity analysis showed that especially during the crisis in a country which is in deep financial problems the assumption is not always correct. In reality, the bottom income quintile households were either amortising debt quicker than expected or their financial assets stock was rather stable comparing to the other quintiles. This caused particularly large errors in the estimation. However, due to lack of data this issue cannot be enough deeply investigated and in the future when more of these survey results are available this issue should also further be investigated. Additionally, it should be further investigated whether this framework could be used in some other contexts, i.e. for instance for completing the estimates for years in which the survey has not been run or estimating risk adjusted balance sheets for households by different household types.

## REFERENCES

- [1] O. Castrén and I. K. Kavonius (2013): “Sector-Level Financial Networks and Macroprudential Risk Analysis in the Euro Area” in Handbook of Systematic Risk edited by J-P Fouque and J Langsam,, Cambridge University Press, Cambridge, pp. 775–790.
- [2] J. Honkkila. and I. K. Kavonius (2013): “Micro and Macro Analysis on Household Income, Wealth and Saving in the Euro Area”, Working Papers 1619, European Central Bank, Frankfurt am Main.
- [3] I. K. Kavonius and J. Honkkila (2013): “Reconciling Micro and Macro Data on Household Wealth: A Test Based on Three Euro Area Countries”, Journal of Economic and Social Policy: Vol. 15: Iss. 2, Article 3.
- [4] J. E. Stiglitz & A. Sen & J-P Fitoussi (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress, [www.stiglitz-sen-fitoussi.fr](http://www.stiglitz-sen-fitoussi.fr).
- [5] M. Fleurbaey (2009): “Beyond GDP: The Quest for a Measure of Social Welfare”, Journal of Economic Literature 47: 1029–1975.
- [6] ECB (2013). The Eurosystem Household Finance and Consumption Survey: Methodological Report for the First Wave. European Central Bank Statistics Paper Series, No. 1.
- [7] D. A. Love, P. A. Smith and L. C. McNair (2008): A New Look at the Wealth Adequacy of Older U.S. Households. Review of Income and Wealth, December 2008, 54 (4), pp. 616-42.

# An Analysis of Household Debt using the Linkage between Micro and Macro Balance Sheet data

Juha Honkkila (Juha.Honkkila@stat.fi)<sup>1</sup> and Ilja Kristian Kavonius (Ilja\_Kristian.Kavonius@ecb.europa.eu)<sup>2</sup>

**Keywords:** wealth distribution, wealth survey, national accounts, micro-macro link, indebtedness

## 1. INTRODUCTION

Since 2008 when the U.S. subprime mortgage crisis triggered the financial crisis, financial stability analysis has been increasingly interested in how leveraged the households are and how the potential risks and imbalances related to these are transmitted to the other sectors (see for instance: Castrén and Kavonius 2013). Hitherto, the lack of timely, comparable household balance sheet data has limited this type of analysis. This paper uses the micro-macro linkage of wealth and income accounts and thus, creates a set of macroeconomic wealth accounts broken down by household groups, using national level micro data. From the methodological point of view, this paper is a continuation of our previous work where first macroeconomic accounts broken down by household type were created for Finland, Italy and the Netherlands.<sup>3</sup>

## 2. APPLIED METHODOLOGY

Our previous paper mentioned above presented a linkage between the Household Finance and Consumption Survey (HFCS) and Euro Area Accounts (EAA). In this paper, we use, to a large extent, the same linkage as in our previous paper. Concerning the comparability of the data sets, we also made two adjustments to the financial accounts estimates which are repeated in this paper. First of all, many countries cover also non-profit institutions serving households. In case there are household sector accounts without non-profit institutions (NPISH), these are naturally used but in case these are not available, the series without NPISH are estimated. Additionally, there are pure population differences in the figures.

This paper focuses on household indebtedness using some of the most common indicators in order to analyse different aspects of indebtedness. The indicators used measure household debt divided by income, liquid assets and total assets. This approach, where several aspects of wellbeing are covered simultaneously, is emphasised in the current welfare analysis, and is one of the key recommendations in the Stiglitz, Sen and Fitoussi (2009) report. Marc Fleurbaey (2009) identifies four approaches of the measurement welfare. The approach applied in this paper is near to the capability approach which emphasise the need of looking several aspects of the welfare even though the capability approach goes even further by covering material and immaterial aspects of wellbeing.

In order to make theoretically solid household group breakdowns and time series analyses the wealth concepts used have to consist of items that are comparable across the

---

<sup>1</sup> Statistics Finland

<sup>2</sup> European Central Bank

<sup>3</sup> Honkkila and Kavonius 2013. Kavonius and Honkkila 2013.

two sources as presented Honkkila and Kavonius (2013). This is why, for example, currency, unquoted shares and other equity, financial derivatives and other accounts payable/receivable are excluded from our concept of liquid wealth. Unquoted shares are included in the wealth concept of both HFCS and National accounts. However, the sector delineation of unquoted shares, especially concerning self-employment businesses, is different in the two sources (ECB 2013, p.93) and it is not possible to combine the survey data on unquoted shares as such with the National Accounts data.

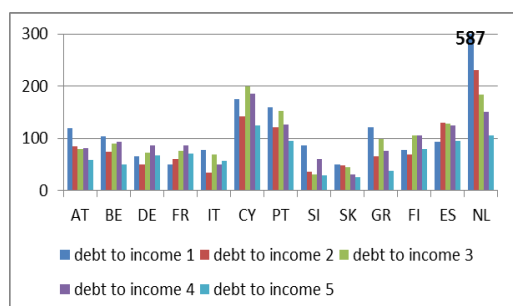
Comparing to the concept used in Honkkila and Kavonius (2013), net equity of pension funds and life insurance is included in liquid assets. Several papers suggest that there are conceptual and coverage problems in the comparison of this item in micro and macro statistics (Love, Smith and McNair 2008). Given that pensions and life insurances are an important part of household wealth, we consider the items to be comparable enough so that the breakdowns by household groups can be applied to the National accounts framework.

The concept of total assets used in this paper includes liquid assets, real assets, as well as unquoted shares. As the availability of these data at macro level is not particularly good, this indicator has required more estimation than the other indicators in this paper and thus, the quality of these estimates are worse than the other estimates presented in this paper. The detailed estimation of the account is presented in the full version of the paper.

### 3. RESULTS

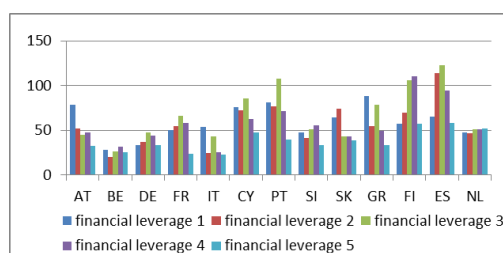
The full version of the paper reports time series from 2008 to 2011 for these indicators but in this short summary we focus only on 2010 results. Figures 1 and 2 show the debt-to-income ratios and financial leverage ratios by country and by income quintile. These figures illustrate the risk of financial insolvency in the short and medium term. As can be seen, cross-country differences in the levels of these indicators can be observed. The debt to income ratio of the entire household sector is more than 100% in Spain, Cyprus, the Netherlands and Portugal. Highest financial leverage ratios are observed in Spain, Cyprus and Finland. Slovenia and Slovakia are countries with relatively low debt levels in comparison with income levels, while the debt levels of Belgian and Italian households are low compared to their financial wealth stock. In the Netherlands, debt to income ratios are extremely high, but financial leverage ratios reasonable. This can be partially explained by the specific mortgage contracts in the Netherlands, leading to unusually large debt stocks that are offset by financial assets held for the purpose of repaying the entire mortgage after a longer period.

With regard to the differences between income groups, several patterns can be distinguished. First of all, high-income households have relatively low debt-to-income ratios as well as financial leverage ratios in all countries. This means that in spite of the high share in the debt stock, high income households seem rarely to be the most vulnerable ones.

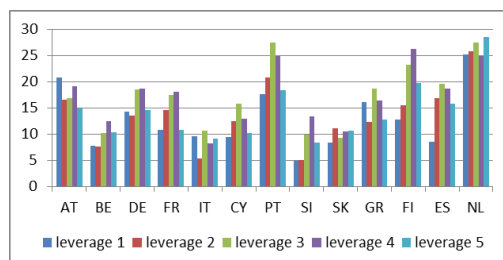




**Figure 1. Debt to income ratios by country and income quintile**



**Figure 2. Financial leverage ratios by country and income quintile**



**Figure 3. Leverage ratios by country and income quintile**

Depending on the country, the most vulnerable income groups with regard to short-and medium term financial insolvency are the bottom quintile or the third and fourth income quintiles. Not surprisingly, the bottom income quintiles have highest debt to income ratios in most countries. In Cyprus, Greece and Portugal the debt ratio in the first income quintile is particularly high. However, the risk of financial insolvency seems to be a problem for the middle income class households, with the exception of Greece, Austria and Slovakia where low income households also have relatively high financial leverage ratios in addition to high debt to income ratios.

The ratio of the debt stock to total assets, including real wealth, varies between nine and 26% in the euro area countries. The causality between income and leverage ratio is not very systematic (see figure 3). Overall, the indebtedness in relation to the total assets seems to be highest in the third and fourth quintiles. The reason for this might be that most households in the first quintile do not have sufficient income to be able to participate in the debt markets and several households in the fifth quintile have enough wealth and do not need to borrow money. The exceptions are Austria, Italy and the Netherlands. In Austria and Italy, the first income quintile is particularly indebted in relation to their assets and in the Netherlands all the income quintiles are highly indebted. However, overall, i.e. in relation to income as well as in total assets, the Dutch households are the most indebted ones in the euro area. The more detailed results and reliability analysis of the results is presented in the full version of the paper.

#### **4. CONCLUSIONS**

The discussion on the micro-macro linkage is old and so far, the discussion has mostly focused on the theoretical micro-macro linkage rather than on practical applications. In this paper, we have made an attempt to create a practical application of a micro-macro linkage and estimated debt-to income, financial leverage and leverage ratios by income quintiles. From the financial stability point as well as from the household insolvency point of view, these are important indicators and we use them in this paper to analyse the risk of financial insolvency in different countries.

- the Statistics Centre, containing all references to toolboxes solving computational aspects of the daily statistical work - a sort of typical library for National Accounts;
- The Metadata Handler, embedding all mechanisms handling the reference metadata repository; SDMX concepts and principles should be implemented here; it may address databases with local information (private Data Source Definitions, confidential categories of reference metadata, ...) beside global ones which may be shared over the web for better integration of distributed information systems, better computation of sophisticated queries; the Eurostat metadata handler is already very advanced in this direction;
- The Query Centre where algorithms for a new generation of queries will be settled; the central organ of this agent is a mediator which will use semantic web technologies to accommodate different formats of databases and different languages (see [5]); it will amplify immediately the effort over the last decade to increase the documentation of statistics through structured and less structured metadata;
- The gateway controller where the flows coming in and out of the local information system are secured, checked early enough before entering in the production systems, etc; eDamis, as designed by Eurostat, is particularly ready to render services in direction.

## CONCLUSIONS

This study proposes an architecture, SOA oriented, for the compilation of European National Accounts based on a (European) data warehouse system and proves how it can be realistic to implement. It highlights how the fundamental aspect of such architecture is harmonisation of data models across official statistics producers; at the moment of the study, a quite ideal stage of development, difficult to achieve in practice. The interesting elements developed in the study are the technologies introduced in the Query centre, especially conceived for accommodating different formats and languages – semantic web technologies.

What emerged from the study is that the semantic web technologies in semi structured information systems are suffering different speed of convergence on the data models: at national level because of legacy issues and different historical backgrounds that lead to non-homogeneous evolution of production processes; at sectoral level because source databases (often administrative databases) play a major role in the National Account production processes and are often far from a SDMX normalisation.

Since the study in 2011, the ESS has evolved towards a more integrated production system for European statistics and launched a coordinated set of modernisation activities (ESS Vision 2020, see [6]) that will set up the IT technical basis for starting making the proposed integrated solution effective. The ESS Vision 2020 initiative should be seen as the leading initiative towards the implementation of a metadata driven production system.

## REFERENCES

- [1] European Commission, Communication on the production method of European statistics: a vision for the next decade”, COM (2009)404, August 2009.
- [2] UNECE Secretariat, Generic Statistical Business Process Model, Version 4.0 – April 2009.

- [3] Michael Bruneforth, UNESCO Institute for Statistics, Bo Sundgren, Statistics Sweden, Conceptual analysis of UOE and UNESCO education surveys - June 2007
- [4] Emmanuel Libeau, François Libeau, Laurent Molini, National Accounts Production System (Naps) Design Report, Sponsored by Eurostat, Final report contract 20102.2008.001-2010.144, July 2011.
- [5] Héctor Pérez-Urbina, Boris Motik, and Ian Horrocks. Tractable Query Answering and Rewriting under Description Logic Constraints. *Journal of Applied Logic*, 8(2):151–232, 2009.
- [6] European Statistical System, ESS Vision 2020.

# Households in Europe in years of economic crisis

**Leonidas Akritidis**, Eurostat – European Commission

**Filippo Gregorini**, Eurostat – European Commission

**Keywords:** Household sector, saving rate, consumption theories, economic convergence.

## 1. INTRODUCTION

Analysis of saving rates has become increasingly accepted in Europe as a useful tool for macro-economic monitoring and decision-making. More in general, household decisions on consumption and saving paths play a very important role in the outlook for aggregate demand. A range of factors are supposed to help explaining the fluctuations of the household saving ratio over the period 1999 to 2013. These factors are based upon widely accepted economic theories on consumption and saving decisions.

The aim of the paper is to illustrate the fit of these theories to the evolution of saving paths in the last years, with particular attention to the changes occurred because of the financial and economic turbulence since 2007.

Can theories of consumption and saving help us to understand?

Is Europe still moving towards economic convergence as it was up to 2007?

## 2. METHODS

In The General Theory of Employment, Interest and Money, J.M Keynes (1936) argued that people mainly save money "to build up a reserve against unforeseen contingencies", the so-called "precautionary motive". Following the same arguments, the seminal papers by M. Friedman (1957) and H. Leland (1968) put emphasis on the fact that in case of bad expectations on future income and / or increasing economic uncertainty savings should increase. Friedman recognizes that households are to some degree forward looking and that they would prefer a smoother consumption path to a more volatile one, i.e. consumption smoothing – as explained to 1st year students of economics. Leland focused mainly on insurance and credit market to point out that the presence of imperfect insurance markets and risk-averse consumers imply that they will save as a precaution against unexpected and unpredictable falls in income.

Other macroeconomics determinants of variations in consumption and savings are interest rates, credit conditions and wealth.

The Fisher's model of inter-temporal consumption shows us that if interest rate increases the effects on consumption and saving depends upon the relative position of the consumer in the market: namely, if net saver or net borrower. In general, consumers are supposed to postpone consumption in case of higher interest rates because it increases the real return to saving.

Credit constraints also play an important role in determining saving behaviour of households. Beyond the spread between rates on borrowing charged and rate on deposit charged paid by banks, some households that are estimated to be less creditworthy could face a much higher cost of borrowing or be denied access to credit, thus they could not be able to borrow as much as they want to finance their desired consumption. If credit becomes more expensive or more difficult to obtain, then borrowing and spending will be lower and saving higher as a consequence.

The net effect of wealth on saving is more uncertain. Households are not expected to react to every movement of asset prices since they can be highly volatile. Moreover, a high share of household financial assets is in the forms of funds, pension funds and life insurance, whose variations in market prices are less visible. With respect to housing

asset values, the effect depends on the relative position of each single household in the market (an owner, a renter, a potential owner?). But also indirect effects have to be carefully considered. If, for example, house values increase, households have more collaterals against which to borrow.

The main problem of Friedman's theoretical model of permanent income is a pure empirical one: how is it possible to identify situations where income changes in a predictable way to properly "test" the validity of the model? Moreover, liquidity and credit constraints play a crucial role in the critique of Friedman's model, since their effects were not deeply analysed and discussed. In the presence of liquidity and credit constraint consumers cannot borrow in anticipation of an income increase, thus consumption will change at the time of income increase only! To overtake this point from a theoretical perspective the model recently developed by T. Jappelli and L. Pistaferri (2010) shows that consumption should react to unexpected income shocks conditionally on the characteristics of the shocks themselves and on the previously discussed constraints. Consumer should not respond therefore to anticipated income changes simply because in most cases this is not a viable option.

Basic accounting identities for households sector:

$$(1) B8G = B6G - P31,$$

where B8G is savings, B6G gross disposable income and P31 final consumption expenditure.

$$(2) \text{ Saving rate} = B8G / (B6G + D8net),$$

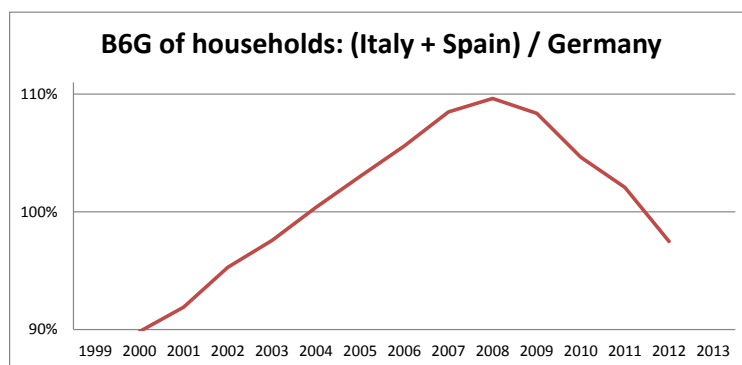
where D8net is the change in the net equity of households in pension funds reserves.

### 3. RESULTS

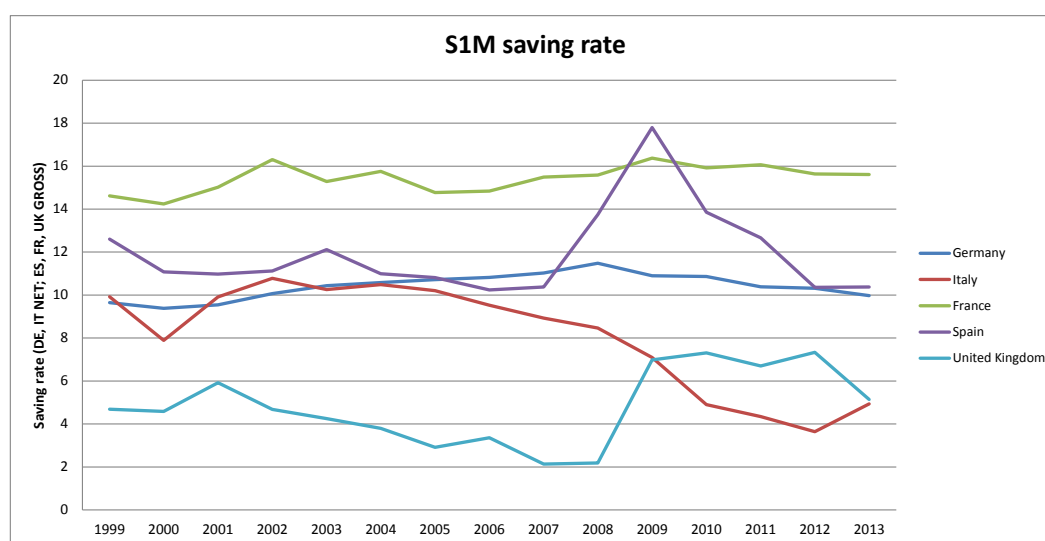
The importance of households sector in Sector Accounts: In terms of gross disposable income, the households and non-profit institutions serving households sectors (S14 + S15) share of domestic sectors (S1) is similar almost everywhere and in all time span 1999-2013. Also in "non-western" countries like China and in Russia it is constantly around 60%.

Comparing 2007-2012 to 1999-2007 in terms of nominal earning and spending growth in big EU countries and also Australia, China, Japan, Russia and the US, only Germany and Sweden are performing better after 2007 in terms of both the variables.

Focusing on Europe, the difference between "core" and "periphery" Euro Area countries is evident. Pre-crisis, the economies (and the household sector) of Spain and Italy grew faster than the one of Germany, but they show no growth from 2007 onwards, whereas Germany kept on increasing since 2009, after a stable path in 2007-2009.



If we now focus on saving rate of households, we observe flat paths in Germany and France, and a modest level shift in the UK. On the contrary, something changed for Spain and Italy after 2007. For Spain and Italy, classical theories can explain only Spain between 2007 and 2009 – precautionary savings because of bad expectations on future income.



If we also control for the effect of durable goods, we still observe constant paths for Germany and France. In UK lower spending on durables is "compensated" by higher savings after 2007. For Spain and Italy, we observe a constant decline of both savings and consumption on durable goods (except Spain between 2007 and 2009, only because of the saving boom).

Germany increased financial assets and decreases financial liabilities from 2007 onwards. Italy did the opposite, whereas Spain liabilities increased a lot after 2007 and assets were almost stable.

If we consider real assets and housing in particular, we observe that homeowners share is lower in Germany (44.2%) compared to the one of Italy (68.7%) and Spain (82.7%). Since housing play a decisive role in determining the net wealth of households, it is apparent that the relative position of Spain and Italy compared to Germany has changed a lot since 2007 because of the housing bubble that occurred in Spain and Italy, and not in Germany.

Up to 2007, the correlation in the EU between earning (B6G) and spending (P31) was above 0.96 everywhere. During and after the financial and economic crisis, in some countries it has fallen below 0.9. In Italy it is 0.56 in 2007-2012, and in Spain it is -0.02 in the same period. Low correlation between B6G and P31 means no stable behaviour over time between consumption and saving.

The same countries where correlation has fallen also show high variance in saving rate after 2007. It is apparent that both these two phenomena are related with uncertainty.

#### 4. CONCLUSIONS

The link between lack of correlation between B6G and P31 on the one hand and higher variance of saving rates on the other hand displays a picture where uncertainty on the future has increased to unpredictable levels in some EU countries from 2007 onwards. No "classical" theories seem to be able to explain this.

A detailed comparison between "core" and "periphery" EU countries seems to suggest that Europe going from pre-crisis convergence to long term divergence, as argued by Wunsch (2013).

*The final version of the paper will include insights on the effects of the implementation of ESA2010 on saving rate of households. Preliminary results are available but are confidential till the deadline for NTTS Conference 2015 abstracts presentation.*

## **REFERENCES**

- [1] ECB (2013), The Eurosystem Household Finance and Consumption Survey.
- [2] Eurostat Sector Account database.
- [3] M. Friedman (1957), The Permanent Income Hypothesis. Princeton University Press.
- [4] T. Jappelli and L. Pistaferri (2010), The Consumption Response to Income Changes. NBER Working Paper No. 15739.
- [5] J.M. Keynes (1936), The General Theory of Employment, Interest and Money. Palgrave Macmillan.
- [6] H. Leland (1968), Saving and uncertainty: the precautionary demand for saving. The Quarterly Journal of Economics, Vol. 82, 465-473.
- [7] OECD Sector Account database.
- [8] P. Wunsch (2013), Is the European Integration Machine Broken?, InterEconomics, Vol. 48, No. 2.

# Summarizing Data using Partially Ordered Set Theory: An Application to Fiscal Frameworks in 97 Countries

Julia Bachtrögl<sup>1</sup>, Harald Badinger<sup>1,2</sup>, Aurélien Fichet de Clairfontaine<sup>1</sup> ([afichet@wu.ac.at](mailto:afichet@wu.ac.at))  
and Wolf H. Reuter<sup>1</sup>

**Keywords:** Partially Ordered Set Theory, Composite Indices, Index Functions, Fiscal Frameworks, Fiscal Rules, Budgetary Procedures

## 1. INTRODUCTION

Criteria and methods for measuring objects based on nominal or ordinal properties have always been of interest in the social sciences. Composite indices, computed by means of scoring or index functions, are one of the most common methods to rank or measure objects based on different properties. Composite indices are employed in various fields of social sciences, e.g. to measure the quality-of-life in a country, the degree of corruption of governments, industrial competitiveness, environmental sustainability, institutional quality, or economic performance.

However, the validity (and the robustness of the results obtained by the use) of composite indices has often been challenged due to unavoidable subjectivity in their calculation.<sup>3</sup> This shortcoming contributes to the disparities in the conditional ordering of objects and limits their comparability.

To address these issues we advocate the use of partially ordered set theory (POSET), which requires subjectivity only in the choice of the properties to be considered.<sup>4</sup> As an application, we consider the measurement of the stringency fiscal frameworks as and compare the results obtained using the POSET approach with those obtained using composite indices (used in previous studies).

The present paper makes the case for an increased use of POSET theory in the social sciences and provides a comparison of POSET indices and composite indices (from previous studies) measuring the "stringency" of fiscal frameworks using data from the OECD Budget Practices and Procedures survey (2007/08).

## 2. METHODS

### 1.1. POSET Theory

POSET theory is presented in this paper as an alternative approach to construct composite indices and ranks of objects. While employed in natural sciences, where it has proven as a reliable methodology for computing robust indices and reducing the requirements for subjective choice to a minimum, it has to the best of our knowledge not

---

<sup>1</sup> Vienna University of Economics and Business, Department of Economics, Welthandelsplatz 1, 1020 Vienna, Austria.

<sup>2</sup> Austrian Institute of Economic Research (WIFO), Arsenal, Objekt 20, 1030 Vienna, Austria.

<sup>3</sup> Subjective choices are necessary on various levels: i) a selection of variables due to the scarcity of comparable qualitative data; ii) a choice of weights for each property is necessary, depending on prior knowledge, expert views or beliefs; iii) the treatment of missing values that may limit the number of observations and aggravate the interpretation of the composite indices; and iv) the choice of an aggregation technique (the scoring scheme), the computing process and of the scale.

<sup>4</sup> This methodology is convenient in providing comparable indices or ranks of objects while reducing subjectivity to a minimum.



been applied in the social sciences so far. The application of POSET theory is advocated also for practical reasons by reducing the subjectivity in the assessment of indices and fully exploiting the available information. According to Brüggemann and Patil (2011) and De Loof et al. (2008) partially ordered set theory offers the possibility of ordering objects conditional on their characteristics. It is considered as an efficient alternative to composite indicators computed by index functions since the assignation of scores to each property is reduced to an ordinal rank. There are also established methods for checking the sensitivity of POSET indices with respect to the selection of properties (the remaining subjective choice) (Brüggemann and Patil, 2011).

We define a set  $O$  of  $i = 1, \dots, N$  objects (depicted by  $x_i \in O$ ) with  $J$  corresponding properties as  $q(x_i) = (q_1(x_i), q_2(x_i), \dots, q_J(x_i))$ . These properties are ordered in a set  $Q$  as  $q(x_i) \in Q_j \forall j$ . Keeping this in mind, we define a partially ordered set as a set of objects ordered alongside their properties by the following relation:

An object  $x_i \in O$  is written as  $x_i \leq_O x'_i$  if  $q_j(x_i) \leq_Q q_j(x'_i)$  where  $\leq_O$  and  $\leq_Q$  are the respective binary partial order relations between objects  $x$  and linear order relations between properties of these objects. Not comparable objects are written as  $x_i ||_O x'_i$  and are dealt with in the POSET theory by adding a linear extension approach, thus attributing a rank  $r$  to every object  $x_i \in O$ . For two objects  $x_i$  and  $x'_i$ , the linear extensions are defined such that  $x_i <_O x'_i$  implied  $r_i < r'_i$ . The set of all linear extensions of the POSET  $(O, \leq_O)$  is denoted as  $\varepsilon(O)$ . The set of linear extensions contains all possible rankings of objects  $x_i \in O$  obeying the POSET theory. Assuming that all linear extensions have the same probability and are uniformly distributed on an interval  $[0, |\varepsilon(O)|]$ , each rank of an object can be associated with a certain rank probability. Indeed, the expected value of the rank of an object  $x_i \in O$  is given by

$$p_i(r) = |\varepsilon_i^r(O)| / |\varepsilon(O)| \quad (1)$$

for  $r \in [1, n]$  where  $|\varepsilon_i^r(O)|$  is the number of linear extensions in  $\varepsilon(O)$  and  $|\varepsilon(O)|$  is the cardinality of the set of linear extensions also referring to the number of elements contained in the set. By construction, the sum of all expected probabilities for an object  $x_i$  is equal to 1:  $\sum_{r=1}^N p_i(r) = 1$ . Therefore, the expected rank of an object  $x_i \in O$ , also referred to as averaged rank  $\bar{r}_i$  is written as follows:

$$\bar{r}_i = \frac{\sum_{r=1}^N [r \cdot p_i(r)]}{N} \quad (2)$$

The POSET approach of ranking objects and, analogously, generating index values relates on the number of objects contained in  $O$  and on their properties contained in  $Q$ . As shown in Brüggemann et al. (2004), the computation of the approximate rank probabilities of more than  $n = 25$  objects exceeds today's computer capacities. We therefore apply the approach of Brüggemann et al. (2004) that addresses this issue by calculating the approximate average rank of objects using Local Partial Order Model techniques.

## 1.2. Data

The OECD dataset was collected during 2006 and 2007, as a revised version of the 2003 survey, using an online questionnaire for senior budget officials of 97 countries. It contains 89 questions on the countries' fiscal frameworks. In general, each category contains questions with a given set of prescribed answers, containing information on i)

whether fiscal rules, restrictions, transparency or other requirements on the fiscal authority in the budgetary process exist, and ii) if so, how stringent they are, and iii) which legal basis (constitution, law, informal rule) they have. In our analysis, we choose those questions matching as close as possible the ones included in the calculation of fiscal composite indices in previous studies.

### 3. RESULTS

This section compares the two methods (Composite indices vs. POSET) to construct indices of fiscal frameworks based on the literature: ACIR (1987); Bohn and Inman (1996); Wagner (2003); Alesina et al. (1999); Filc and Scartascini (2004); Von Hagen (1992); Gleich (2003); Debrun et al. (2008).

Investigating the degree of similarity between POSET theory (yielding POSET indices) and index functions (yielding Composite indices), 19 indices yielded by both methods are compared. Correlation coefficients are reported in Table 1. The principal shortcoming of composite indices lies in the assignment of points to particular characteristics of fiscal rules and budgetary procedures and the choice of weights is highly subjective that can lead to large sensitivity in response to varying point schemes and weights. In opposition, POSET theory reduces subjectivity to a minimum.

**Table 1. Correlation of Composite Indices and POSET Indices**

Index		Spearman Correlation	Kendall tau (indices)	Differences*		
				Mean & (Std. Dev.)	Min	Max
<b>ACIR</b>	Category I	0.60	0.60	0.35 (0.18)	0	0.75
	Category II	-0.52	-0.54	0.59 (0.19)	0	0.69
<b>Bohn and Inman</b>	No Carry Over	0.99	0.96	-0.00 (0.11)	-0.45	0.53
	Governor Veto	0.91	0.89	-0.02 (0.10)	0	0.48
<b>Wagner</b>	Deposit Rule	0.96	0.95	0.40 (0.14)	0.25	1.13
	Withdrawal Rule	0.64	0.58	0.76 (0.19)	0.38	0.88
<b>Alesina et al.</b>	Total	0.59	0.46	0.22 (0.23)	-0.46	0.85
	Borrowing Constraint	0.62	0.52	-0.11 (0.25)	-0.77	0.66
	Agenda Setting	0.92	0.79	0.04 (0.14)	-0.44	0.36
<b>Filc and Scartascini</b>	Fiscal Rules	0.65	0.47	0.26 (0.12)	-0.00	0.68
	Hierarchical Procedures	0.81	0.70	0.56 (0.20)	0.06	1.00
	Transparency	0.38	0.32	0.37 (0.19)	-0.02	0.95
<b>Von Hagen**</b>	Total	0.84	0.65	0.78 (0.07)	0.63	0.86
	Negotiations Structure	1.00	1.00	0.75 (0.00)	0.75	0.75
	Budget Approval	0.69	0.57	0.72 (0.05)	0.64	0.81
	Budget Implementation	0.85	0.68	0.78 (0.04)	0.67	0.85
	Budget Rules	0.89	0.75	0.77 (0.07)	0.66	1.13
<b>Gleich</b>	Total	0.77	0.59	-0.08 (0.20)	-0.44	0.44
	Preparation	0.80	0.66	-0.06 (0.21)	-0.26	0.26
	Legislation	0.82	0.68	-0.00 (0.22)	-0.79	0.79
	Implementation	0.79	0.67	-0.05 (0.20)	-0.63	0.63
<b>Debrun et al.</b>	Expenditure Rule	0.73	0.57	-0.22 (0.27)	-0.73	0.25
	Revenue Rule	0.36	0.29	-0.15 (0.37)	-0.62	0.52
	Budget Balance Rule	0.69	0.55	0.03 (0.24)	-0.55	0.38
	Debt Rule	0.44	0.34	0.14 (0.34)	-0.66	0.60

Notes: \* Mean differences and standard deviation of these differences between the replicated indices and the POSET indices from the literature as well as the minimum and maximum difference value. Composite indices are divided by maximum index scale (see Table 1) in order to be standardized between zero and one. All correlation coefficients contained in Table 4 are significantly different from zero (at the 5 % level). \*\* Replicated indices based on methodology by Von Hagen (1992), data taken from Hallerberg (2003).

The results reveal interesting features of the interrelationship between composite and POSET indices: the large majority of correlation coefficients reported in Table 1 are high, indicating that the POSET indices are approximating their counterparts based on an index function approach (or vice versa). Moreover, all correlation coefficients (for both Spearman's correlation coefficient and Kendall tau) are statistically significant. We find that POSET indices are very similar to composite indices if the underlying basis of questions or number of NA's is small, i.e. where the subjective choices are not so important due to the small amount of possible weightings or selections. If the datasets

and fractions of NA's are larger we do find larger difference between the two approaches as POSET allows considering more information in the index calculation.

#### 4. CONCLUSIONS

The OECD dataset is shown to be suited to replicate many fiscal rules measures from the literature and therefore enables a comparison of the composite indices with indices based on POSET theory. We compare POSET and composite indices of fiscal rule (using the same set of questions as previous studies but the full country sample) and find partly large similarities between indices obtained by POSET theory and composite indices. Of course, the large correlation for part of the indices does not necessarily imply that the used of these two kinds of indices in econometric analyses yields the same results. Overall, we argue that POSET indices can serve as an (more) objective benchmark for the measurement of performance and ranking analyses.<sup>5</sup>

POSET theory has been shown to be a well suited alternative to composite indices. It allows to fully exploit available information for the ordering of objects conditional on their properties, yields measures that are less sensitive to subjective decisions, improves comparability, and simplifies the calculation of rank indices by making designation of properties, assignation of weights and deletion of objects with missing data unnecessary.

#### REFERENCES

- [1] Brüggemann, R. and Patil, G. P., *Ranking and Prioritization for Multi-indicator Systems: Introduction to Partial Order Applications*, Vol. 5, Springer (2011)
- [2] De Loof, K., De Baets, B., De Meyer, H. and Brüggemann, R., A hitchhiker's guide to POSET ranking, *Combinatorial Chemistry & High Throughput Screening* 11(9) (2008) 734–744.
- [3] Brüggemann, R., Sørensen, P. B., Lerche, D. and Carlsen, L. Estimation of averaged ranks by a Local Partial Order Model, *Journal of Chemical Information and Computer Sciences* 44(2) (2004) 618–625.
- [4] ACIR, 'Fiscal discipline in the federal system: National reform and the experience of the States. Washington, D.C.', Advisory Council on Intergovernmental Relations. (1987)
- [5] Bohn, H. and Inman, R. P., Balanced-budget rules and public deficits: Evidence from the US States, in 'Carnegie-Rochester Conference Series on Public Policy', Vol. 45, Elsevier (1996) 13–76.
- [6] Wagner, G. A., 'Are state budget stabilization funds only the illusion of savings?: Evidence from stationary panel data', *The Quarterly Review of Economics and Finance* 43(2) (2003) 213–238.

---

<sup>5</sup> In the working paper version of this study, a further step shows significant and positive correlations between 1) global POSET indices, computed by taking into account all available and applicable information provided by the dataset (without restrictions) with 2) the corresponding specific indices that are confined to smaller sets of questions. This result is taken as a hint that the selection of relevant properties seems not to be too crucial for the indices computation (in the present setting).

- [7] Alesina, A., Hausmann, R., Hommes, R. and Stein, E., 'Budget institutions and fiscal performance in Latin America', *Journal of Development Economics* 59(2) (1999) 253–273.
- [8] File, G. and Scartascini, C., Budget institutions and fiscal outcomes: Ten years of inquiry on fiscal matters at the research department, in 'Presentation at the Research Department 10th Year Anniversary Conference. Office of Evaluation and Oversight. Inter-American Development Bank' (2004)
- [9] Von Hagen, J., 'Budgeting procedures and fiscal performance in the European communities', *Economic Papers* 96 (1992) pp.1–79.
- [10] Gleich, H., Budget institutions and fiscal performance in Central and Eastern European Countries, Technical report, European Central Bank (2003)
- [11] Debrun, X., Moulin, L., Turrini, A., Ayuso-i Casals, J. and Kumar, M. S., 'Tied to the mast? National fiscal rules in the European Union', *Economic Policy* 23(54) (2008) 297–362.
- [12] Hallerberg, M., 'Budgeting in Europe: Did the domestic budget process change after Maastricht?', Paper Prepared for the 2003 EUSA Conference, Nashville, TN (2003)

# A step towards communicating with indicators

Justyna Gustyn ([j.gustyn@stat.gov.pl](mailto:j.gustyn@stat.gov.pl))<sup>1</sup>

**Keywords:** statistics, indicator, communication, STRATEG system, development policy

## 1. INTRODUCTION

In today's increasingly complex and interconnected world, indicators constitute an essential resource for statistical data users, among others, policy-makers and the general public. With more focus on in-depth and fact-based comparisons and benchmarks while creating national and supra national policies, statistics and especially indicators are useful as a way of representing reality. By providing support in making evidence-based decisions and comparisons between policies and programmes, countries and regions, indicators contribute to better perception of changing socio-economic reality, objective performance measurement, increased transparency and accountability.[1] Nowadays, indicators are a powerful way of communicating information, and also a guidance for achieving desired goals.

## 2. MAIN OBJECT

### 2.1. The STRATEG system – a platform for communicating with indicators

In the face of growing importance of indicators in many areas, the Central Statistical Office of Poland is trying to address the challenge of developing and permanently implementing the concept of communicating with indicators into the activities of official statistics. With the aim to enrich communication services, the STRATEG system has been created and developed steadily as a system which perfectly fits with this concept. It is an innovative database application dedicated particularly to users engaged in monitoring cohesion policy as well as country and regional development. The system integrates indicators derived from different sources (statistical and non-statistical), offering the opportunity to present data in visually-attractive forms (charts, maps, reports), which considerably facilitate the process of data analysis.

### 2.2. Metadata – indicator profiles

With the aim of enhancing communication with users, detailed indicator profiles were prepared for resources gathered in the STRATEG system, presenting a set of useful information, inter alia:

- indicator name with its alternative name(s) used in the strategy or programme,
- strategic/programme documents with targets for monitoring,
- theme category to which an indicator belongs,
- indicator's description,
- overall methodological explanation,
- data source and data availability,
- available variables (age group, years, sex, urban/rural areas),
- territorial level of availability (the European Union (EU-27 and EU-28) and Member States, country level, lower spatial aggregation levels, functional areas),
- remarks concerning any changes (including methodological ones) having influence on the time series continuity.

---

<sup>1</sup> Central Statistical Office of Poland

While preparing indicator profiles, official materials and publications not only of the CSO but also of other foreign institutions such as the Eurostat, World Bank or OECD were used as the main information and methodological source.

Indicator profile	
Indicator name	GDP per capita at PPP [EU27=100]
Alternative name / indicator used in the strategy/programme	-
Strategic/programme documents / targets for monitoring	<p><b>National Strategy of Regional Development:</b></p> <ul style="list-style-type: none"> <li>- The main objective: growth, employment, cohesion</li> </ul> <p><b>National Strategic Reference Framework:</b></p> <ul style="list-style-type: none"> <li>- The strategic objective. Creation of the the conditions for growth of competitiveness of economy based on knowledge and entrepreneurship ensuring growth of employment and social cohesion</li> </ul> <p><b>National Development Strategy:</b></p> <ul style="list-style-type: none"> <li>- Key indicators</li> </ul> <p><b>Strategy for Development of the Świętokrzyskie Voivodship:</b></p> <ul style="list-style-type: none"> <li>- Contextual indicators</li> </ul> <p><b>Strategy for Development of the Śląskie Voivodship:</b></p> <ul style="list-style-type: none"> <li>- Priority area: (A) Modern economy</li> </ul> <p><b>Strategy for Development of the Lubuskie Voivodship:</b></p> <ul style="list-style-type: none"> <li>- The main objective. The use of the Lubuskie voivodship potentials for growing quality of life, stimulating the competitive economy, increasing the cohesion of the region and the effective management of its development</li> </ul>

**Figure 1. Screenshot of the indicator profile from the STRATEG system (strateg.stat.gov.pl)**

Apart from a wide set of indicators concerning various disciplines (e.g. national accounts, labour market, territorial cohesion, social capital) including methodological information, definitions and interpretational guidelines, STRATEG collects a range of analytical reports, statistical publications (national and regional) and short analytical comments on current socio-economic situation and trends concerning a given subject. The above mentioned functionalities, characterized by clarity, simplicity and open accessibility, constitute significant support while conducting analytical works.

Indicators gathered in the STRATEG database, besides facilitating measurement of progress against objectives, play an important role in communication. Such way of communicating information to the general public, allow them to monitor different aspects of public life and economy, thereby to look critically at undertaken activities and predict their possible effects.

### 3. PLANS FOR THE FUTURE – INDICATOR INTERPRETATION GUIDE

The examination of indicators as a knowledge practice requires explications of the methodological and analytical boundaries. In the constantly expanding “world of indicators”[2], it is important to enrich critical reflections on indicators by positioning various professional methodological commitments.

To fulfill the aforementioned aim, subsequent development stages of communicating with indicators via STRATEG involve preparation of indicator interpretation guide that will serve our users as navigation through the jungle of information available. Being aware that

statistical indicators of all kinds represent a methodological challenge, the manual serving as guidelines on key indicators of development monitoring will give a general understanding and scientific definition, clear descriptions facilitating analysis and interpretation of statistical measures, explanation of possible consequences or implications as well as reference to the source of information and suggested further reading.

Thanks to the signposts marking the way, it will allow users in-depth comprehension of indicators, among others, by explaining the purpose of measurement, way of use and possible interpretations. Despite some potential limitations in the use of indicators, i.e. dependence on the context, serving precise purposes or misleading interpretation of results, they provide input in global, national and local policies by translating often intangible, abstract objectives into measurable targets against which progress and achievements can be monitored.

#### **4. CONCLUSIONS**

By creation and launch of the STRATEG system, the Central Statistical Office of Poland has made a further step towards communicating information via indicators. STRATEG is not only a comprehensive tool providing information on the programming and monitoring of the development progress, in particular for representatives of government authorities engaged in the process of tracking the implementation of the development goals set at various management levels, but foremost a good example of effective communication with various groups of users.

A wide set of indicators from various areas gathered in the system is perceived as a very useful way of communicating information, among others, by providing warning signals, revealing trends and simplifying complex socio-economic phenomena, and thus helping users to diagnose the current situation, improve ability to take appropriate actions and provide assessment of performance or progress towards established objectives.

#### **REFERENCES**

[1] Eurostat, European Commission, *Towards a harmonized methodology for statistical indicators*. Part 1: Indicator typologies and terminologies. 2014 edition.

[2] *A World of Indicators. Knowledge Technologies of Regulation, Domination, Experimentation and Critique in an Interconnected World*, 13.10.2011 – 15.10.2011 Halle an der Saale, in: H-Soz-Kult, 04.08.2011, <http://www.hsozkult.de/event/id/termine-16974>>.

# Geostatistics Portal – a platform for statistical data geovisualization

Mirosław Migacz ([m.migacz@stat.gov.pl](mailto:m.migacz@stat.gov.pl))<sup>1</sup>

**Keywords:** GIS, geostatistics, census, grid, INSPIRE

## 1. INTRODUCTION

Up to 2013 statistical data published by Central Statistical Office of Poland (CSO) was limited to official indicators, publications, press releases and tables in data banks. The closest thing to spatially referenced statistics were maps inserted into publications as static images. In 2013 CSO released the Geostatistics Portal – a web platform for spatial visualization of census' and other statistical surveys' results.

## 2. METHODS

### 2.1. Building the GIS and its usage in the 2010-2011 census round

In order to carry out two nationwide censuses in 2010 and 2011 Polish official statistics decided to create a geodatabase with dwelling locations for all of the country. An extensive collection of statistical maps has been scanned and georeferenced as the primary reference material. On the basis of census legislation numerous spatial datasets have been acquired free of charge from public administration authorities. All these datasets have been processed to serve as reference material for address point acquisition. In half a year 210 GIS operators in regional units of Polish official statistics have created a statistical geodatabase with 5,7 million address points – the most complete and accurate source of information on dwelling locations in Polish public administration. The database was used in both censuses to prepare a spatialized survey frame, navigate enumerators to respondents and for visualization of census results.

Agricultural Census 2010 and Population and Housing Census 2011 were the first censuses in Poland that were carried out without use of paper. All statistical data in the 2010-2011 census round was collected with reference to address point coordinates. This allows publication of census results aggregated to any spatial unit, whether it's a statistical, administrative unit or a grid cell. The only restriction is the need to maintain statistical data confidentiality.

### 2.2. Geostatistics Portal – a platform for census data visualization

In order to publish census results CSO prepared the Geostatistics Portal. The portal provides tools for creating all sorts of thematic maps.

Statistical data available for the portal are absolute values – these can be directly visualized on various kinds of diagram maps. Users also have at their disposal a set of normalization attributes that are used for on-the-fly calculation of relative values (statistical indicators) to be presented on choropleth maps.

---

<sup>1</sup> Central Statistical Office of Poland



For all types of presentation users select thematic phenomena from a predefined list, define the aggregation level (territorial unit) of the output as well as the following (if applicable): symbol, color range, number of classes and classification methods.

Spatial reference of statistical data kept in the portal database goes as low as LAU2 level (municipality). However, users within Polish official statistics will have access to more detailed data. They will be able to create spatial queries for an aggregated value of a phenomenon within a user defined polygon as well as perform advanced spatial analysis on microdata. Results of such analysis may then be published directly in the Portal or become part of a publication, provided they meet requirements of statistical data confidentiality.

Apart from thematic maps and spatial analyses in the Geostatistics Portal users can find spatial data maintained within official statistics (such as statistical unit boundaries). The portal serves as a publishing platform for spatial data services Central Statistical Office is obliged to provide according to the Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). A discovery service has been established as well as view and download services for two spatial data themes: statistical units and population distribution (demography). Access to all services (discovery, view, download) is public and free of charge.

### **2.3. Geostatistics Portal – visualization of statistical data other than census results and plans for future development**

This year a new dataset has been introduced to the Geostatistics Portal – the Local Data Bank. It contains statistical data of a broad scope that are available at a municipality (LAU2) level. The Bank is regularly updated and holds historical statistics as well. Up to now the statistical data was only available for view and download in tables. The Geostatistics Portal provides an extensive set of tools to visualize the data on maps.

Central Statistical Office of Poland actively participates in EU projects ([1], [2], [3]) that cover merging statistical and geographic data and presenting statistical data on grids. Methodology developed in these projects will allow CSO to publish georeferenced statistical data in divisions other than administrative. A population map for units below municipality level (statistical regions and census enumeration areas) is to be published, as well as population maps on kilometer grids of various cell sizes.

Even though the Geostatistics Portal already offers a broad choice of geovisualization tools, there are plans for developing new ones. This includes new cartographic presentation methods (e.g. cartograms), a more convenient way to visualize and compare time series and tools for on-the-fly visualization of user-supplied data.

## **3. RESULTS**

### **3.1. Spatial address databases**

Spatial address databases containing boundaries of statistical regions and census enumeration areas as well as statistical address points (locations of dwellings) have been prepared prior to the 2010-2011 census round. The address points served as a basis for census survey frames, while statistical unit boundaries served as a supporting layer for enumerator management.

### **3.2. Geostatistics Portal**

The Geostatistics Portal is the main platform for statistical data visualization on maps in official statistics. Built for purposes of publishing georeferenced census results it now also contains data from the Local Data Bank and serves as an INSPIRE spatial data services broker.

### **3.3. Products of EU projects**

As a result of Central Statistical Office's participation in EU projects such as the GEOSTAT project (stages: 1A and 1C) or "Merging statistical data and geographic information in Member States" following products emerged:

- total population map for 2006 on a 1km<sup>2</sup> grid [1],
- total population map for 2011 on a 1km<sup>2</sup> grid [2],
- population maps for 2011 on a 1km<sup>2</sup> grid by sex and economic age groups [2],
- methodology for acquisition of address points for enterprises using data possessed by official statistics as well as other spatial data sources (e.g. datasets from the mapping agency) [3],
- methodology for aggregation of population data assigned to X, Y coordinates to grids [3],
- methodology for automated aggregation of population data assigned to X, Y coordinates to various forms of irregular divisions of space (e.g. geodetic precincts, statistical regions, census enumeration areas) [3],
- methodology for calculating statistical indicators for spatial planning using layers from the Database of Topographic Objects, e.g. building density, road density indicators [3].

## **4. CONCLUSIONS**

Central Statistical Office of Poland has made giant progress in the field of geographic information over the last 6 years. Creation of spatial address databases allowed efficient execution of the census and various options for dissemination of georeferenced census results. Regularly updated spatial address databases serve as basis for frames of other surveys conducted by official statistics.

Launching of the Geostatistics Portal was a giant step forward in providing users with a convenient and intuitive way to access statistical data on maps. With a huge database and a wide range of visualization tools at their disposal, users can design their own dynamic map presentations rather than rely on static maps that were included in publications so far. Central Statistical Office constantly works on improving the Geostatistics Portal – publishing new data and developing new tools for geovisualization and spatial analysis – all this to bring statistical data closer to the people and authorities and make governance easier on all administrative levels.

## **REFERENCES**

- [1] GEOSTAT 1A – Representing Census data in a European population grid – final report.
- [2] Interim report on the activities carried out within the grant agreement No 08143.2013.003-2013.466. Action entitled: Producing national population grid datasets of the census 2011 – GEOSTAT 1C.

- [3] Final technical implementation report on the activities carried out under Eurostat grant agreement No 50502.2012.001-2012.519. Action entitled: Merging statistical data and geospatial information in Member States.

# Experiences using LUCAS data in Finnish Land Cover monitoring - Current activities and future plans

Markus Törmä (markus.torma@ymparisto.fi)<sup>1</sup>, Elise Järvenpää, Pekka Härmä, Lena Hallin-Pihlatie, Suvi Hatunen, Minna Kallio

**Keywords:** LUCAS, Land Cover, Land Use, CORINE, Copernicus

## 1. INTRODUCTION

The aim of this paper is to describe the ideas, data and methods how LUCAS in-situ data are being used in the validation processes of the Finnish Corine Land Cover (CLC) 2012 classification and High Resolution Layers which are both produced as part of the Copernicus Land Services programme coordinated by the European Environment Agency (EEA). The aim is also to share experiences gained in the validation process and to discuss the possibilities of using LUCAS national land monitoring in the future.

### 1.1. Land cover monitoring in Finland

The Finnish Environment Institute SYKE is a research and expert organization, which is dedicated to studying phenomena relating to environmental changes, and developing related change management solutions. SYKE is responsible of the maintenance and development of an environmental spatial data infrastructure (ESDI), which serves both the whole environmental administration and external users, such as national and European bodies. In addition to GIS and environmental data, SYKE also produces operational remote sensing products.

In Finland there are several governmental agencies producing accurate spatial databases and registers covering the whole of Finland. These datasets are in many cases based on information acquired from the ground or from the interpretation of aerial images. The existing GIS data and environmental registers form a base for present and future land monitoring activities.

### 1.2. CORINE and Copernicus programmes

European Commission introduced the CORINE programme in order to gather information relating to the European environment. CORINE Land Cover classifications (CLC) have been produced using satellite images and visual interpretation with mapping scale of 1:100000 and 25 hectare minimum mapping unit. The classification nomenclature is hierarchical and contains five classes at the first level, 15 classes at the second level and 44 classes at the third level [1][2].

In Finland, CLC has been made differently in order to produce more detailed national land cover information at the same time, and have been based on automated interpretation of satellite images and data integration with existing digital map data. Map data provides information describing land use and soils and it has been produced by Finnish National Land Survey (NLS), Finnish Forest Research Institute METLA, Ministry of Agricultural and Forestry, Population Register Centre and SYKE. Satellite images have been used in estimation of continuous variables describing vegetation type and coverage, as well as in updating map data. Generalization from national high resolution CLC classification to European version has been made using fully automated

---

<sup>1</sup> Finnish Environment Institute SYKE, Mechelininkatu 34a, PL 140, 00251 Helsinki, Finland

procedure. So far, CLC classification has been produced for years 2000 and 2006 in Finland, and the version for year 2012 has just been completed [3].

Up to recent years CORINE Land Cover mappings have been project based work organized by European Environment Institute. Each participating country has been responsible for data production in its own territory. 2011 onwards CLC updates became part of the European land monitoring which is organized within Global Monitoring of Environment and Security (GMES) programme by DG ENTR of the European Commission [4]. Nowadays CORINE Land Cover classification belongs to operational Copernicus programme under Land Services coordinated by EEA and European Space Agency (ESA).

In Copernicus Land Services the CORINE Land Cover updates are continued, but also five additional pan-European High Resolution Layers (HRL) [4] are produced describing the main land cover characteristics: artificial surfaces (e.g. roads and paved areas), forest areas, agricultural areas (grasslands), wetlands, and small water bodies.

HRL-datasets are interpreted from satellite images on produced by private companies but verified and enhanced by Member States. Presently SYKE also verifies and enhances the High Resolution layers as part of the Copernicus Land Services programme, together with Metla and the Finnish Geodetic Institute.

### **1.3. LUCAS survey**

Since 2006, EUROSTAT has been carried out an area frame survey on the state and the dynamics of changes in land use and cover in the European Union called the LUCAS survey. The surveys are done every three years and the latest LUCAS survey (2012) covers all 27 EU countries and observations on more than 270 000 points. From LUCAS survey 3 types of information are obtained:

- Micro data covering land cover, land use and environmental parameters associated to the single surveyed points;
- Photographs; and
- Statistical tables in NUTS-levels [5].

SYKE has also applied for LUCAS Grants 2014 (Provision of harmonized land cover/land use information: LUCAS and national systems)) together with Finnish Forest Research Institute METLA. The overall strategy of this application is further develop Finnish bottom-up-approach techniques in land monitoring in order to take into account the data needs of EUROSTAT. This includes combination of spatial data with in-situ field surveys completed in National Forest Inventory. This approach will be tested in providing statistical tables for the year 2012 according to the data specifications of EUROSTAT and compared with the results of LUCAS 2012 survey.

## **2. METHODS AND RESULTS**

### **2.1. Using LUCAS data in validation of Corine Land Cover datasets**

Following processing steps are needed to perform the validation CLC classification using LUCAS-points.

- Because of differing classification nomenclature, we have to define how LUCAS-classes are transformed to CLC classes.
- Sample selection from all LUCAS-points: We prefer points that have been visited and photographed. Water areas are the exception, then we prefer points which are observed farther away.
- Find corresponding CLC class from classification for LUCAS-points.
- Compute error matrix and accuracy measures like overall accuracy and classwise producer's and user's accuracies.

## **2.2. Using LUCAS data in validation of High Resolution Layers**

The processing steps are very similar that in previous case, the difference is that we have to define the HRL-classes of LUCAS-points. We are also currently doing the validation of HRLs using own sample points and we would compare the validation results of HRLs computed using LUCAS and own points.

## **3. CONCLUSIONS AND FUTURE PLANS**

Land monitoring in Finland has been based on data integration of information collected and maintained in several organizations responsible for sectorial operational monitoring programmes. Due to changes in the operational environment of national and European land monitoring, new information needs have to be met. The new needs set new demands to existing database management systems and to the contents of environmental registers and other datasets. There is a clear need for spatially, temporally and thematically more accurate land monitoring data. For example regarding agricultural areas, monitoring of changes (location and size of new, unused or overgrown agricultural areas) is one of the critical issues at the moment – phenomena that are not suitable to observe using regular CLC data. SYKE is now trying to meet these requirements by developing a kind of time series dataset of various themes utilizing existing GIS and EO datasets. The overall goal (ambition) is to complement coarse “snap shot land monitoring data” with more exact and continuous yearly time series of several different land monitoring themes.

In the future agreements between EU and member states it should be discussed and agreed on how land monitoring resources are best used based on long term cooperation between EU and national institutions. If data production of European land cover is based on coordinated and long term programme, data needs of EU can be also taken into account in national monitoring programmes. All this would enable faster production of data and availability of national expertise and data sources, which guarantees also high data quality and national usage of produced data.

## **REFERENCES**

- [1] Commission of the European Communities, Corine Land Cover – Technical Guide, Office for Official Publications of the European Communities (1994), <http://www.eea.europa.eu/publications/COR0-part1>.
- [2] M. Bossard, J. Feranec, J., Otahel, CORINE Land Cover Technical Guide – Addendum 2000, European Environment Agency Technical Report 40 (2000), <http://www.eea.europa.eu/publications/tech40add>.

- [3] P. Härmä, R. Teiniranta, M. Törmä, R. Repo, E. Järvenpää, M. Kallio, The production of Finnish Corine land cover 2000 classification, In International Archives of Photogrammetry and Remote Sensing volume XXXV Part B4 (2004), 1330-1335.
- [4] EEA, GMES Initial Operations 2011-2013 Land monitoring Services: Annex I – Tender Specifications, European Environment Agency (2011), <http://www.eea.europa.eu/about-us/tenders/eea-ses-11-004-framework/tender-specifications>.
- [5] Eurostat, Land cover / use statistics- Overview, Eurostat (2015), <http://ec.europa.eu/eurostat/web/lucas/overview>.

# Forecasting skyrocketing unemployment with big data

María Rosalía Vicente ([mrosalia@uniovi.es](mailto:mrosalia@uniovi.es))<sup>1</sup>, Ana Jesús López ([anaj@uniovi.es](mailto:anaj@uniovi.es))<sup>1</sup> and Rigoberto Pérez ([rigo@uniovi.es](mailto:rigo@uniovi.es))<sup>1</sup>

**Keywords:** big data, unemployment, forecasting, Google trends, ARIMAX.

## 1. INTRODUCTION

In the last few years the term “big data” has gained popularity in order to refer to the vast amount of digital data which information and communication technologies are making possible to gather, store, diffuse and share in unprecedented ways.

Economists have started to explore the potential of the “big data” coming from online search behaviour since the internet has become a major source of information [1, 2]. In this sense, several papers have investigated the usefulness of internet search data in order to improve the nowcasting and forecasting of economic indicators, with special attention to unemployment [3-7]. Nonetheless, most of the empirical evidence on this field has focused in countries with low/moderate unemployment rates (France, Germany, the United Kingdom, and the United States, among others).

This paper follows this line of research and explores whether online search data helps to improve the nowcasting and forecasting of unemployment in the context of high rates of employment destruction over time. In particular, the present analysis focuses on forecasting the figures of unemployment in Spain. This country reveals as a very interesting case due to the sharp increases in unemployment caused by the economic crisis. In fact, Spain has one of the highest unemployment rates in Europe and doubles the European Union’s average.

The analysis presented in this paper is based on a time series approach, ARIMAX models, including explanatory variables from both the demand and the supply sides of the labour market: a business sentiment index (the Employment Confidence Indicator) and online searches on “job offer” on Google. The obtained results confirm the usefulness of internet search data as both an economic indicator and a forecasting tool.

## 2. DATA AND METHODS

### 2.1. Data

The variable of interest is given by registered unemployment in Spain, that is, the monthly number of job demands registered by the Spanish public employment services [8]. The period of analysis is from January 2004 to December 2012, taking the year 2013 as forecasting horizon.

With the aim of properly analysing (and then forecasting) the evolution of unemployment, explanatory variables from both the demand and the supply sides of the labour market are taken into account. On the supply side, data on online searches for job vacancies is incorporated into the model. Following the seminal paper by [1], data from Google Trends’ service is used [9]. In particular, data was collected on the queries for “oferta de trabajo” and “oferta de empleo”, which are two different and common ways to translate “job offer” into Spanish. On the demand side, that is, regarding employers, the

---

<sup>1</sup> University of Oviedo



Employment Confidence Indicator (ECI) is included. This monthly indicator shows the balance between the positive and negative opinions of industrial firms about the current employment situation and their perspectives three-months ahead [10].

## 2.2. Methods

The analysis of the unemployment time series follows the well-known Box-Jenkins approach. More specifically, the identification of the considered series starts with the stationarity analysis through the augmented Dickey-Fuller unit root test and the KPSS stationarity test. The existence of seasonal unit roots can be checked through the DHF test.

Once the unemployment series has been transformed in order to achieve the stationarity requirements, and its correlogram has been analysed, the next step is to identify the baseline model. After testing several models, the ARIMA(0,1,2)(0,1,1) shows the best performance and is selected as the baseline for unemployment forecasts:

$$\text{Baseline B1: } (1-L)(1-L^{12})Y_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta_1 L^{12})u_t$$

This first specification can be improved by taking into account the existence of a structural break, and, in particular, by including a level shift (LS) starting in March 2008 and a level shift with trend (t LS). Hence, the baseline B2 expression is derived:

$$\text{Baseline B2: } (1-L)(1-L^{12})Y_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta_1 L^{12})u_t + \gamma_1 LS_t + \gamma_2 t LS_t$$

In order to test whether the data coming from job online searches can provide a more complete description of the evolution of unemployment, ARIMAX models are specified. In particular, the considered explanatory variables refer to the previously described Employment Confidence Indicator (ECI) and the Google index coming from searches related to job offers with alternative terms (Google-T stands for “oferta de trabajo” while Google-E refers to “oferta de empleo”):

$$\text{Model M1: } (1-L)(1-L^{12})Y_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta_1 L^{12})u_t + \gamma_1 LS_t + \gamma_2 t LS_t + \beta_1 X_t^{ECI}$$

$$\text{Model M2: } (1-L)(1-L^{12})Y_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta_1 L^{12})u_t + \gamma_2 t LS_t + \beta_1 X_t^{ECI} + \beta_2 X_t^{\text{Google-T}}$$

$$\text{Model M3: } (1-L)(1-L^{12})Y_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta_1 L^{12})u_t + \gamma_2 t LS_t + \beta_1 X_t^{ECI} + \beta_3 X_t^{\text{Google-E}}$$

Regarding forecasting evaluation, the most common measures have been considered, including the Root Mean Squared Error, the Mean Percentage Error, the Mean Absolute Percentage Error and Theil's U index.

## 3. RESULTS

Table 1 shows the estimation of the two proposed baseline models (B1-B2) together with the ARIMAX models including explanatory variables from both the demand and the supply sides of the labour market (M1-M3).

Focusing on the ARIMAX, a significant negative coefficient is estimated for the Employment Confidence Indicator, confirming that the more positive the perspectives, the lower the unemployment. This negative and statistically significant association between employers' perspectives and unemployment evolution is observed for the three proposed ARIMAX models (M1-M3).

When the variables related to online job searches are included, leading to proposals M2 and M3, it is observed that they significantly improve the previous models; in fact, model M2 results to be the most suitable option according to the significance of the estimated coefficients, the standard deviation of residuals and the information criteria.

**Table 1. Estimation results for ARIMA and ARIMAX models on Spanish unemployment from January 2004 to December 2012**

	Baseline B1	Baseline B2	Model M1	Model M2	Model M3
$\theta_1$	0.7853 ***	0.7603 ***	0.7422 ***	0.6858 ***	0.6863 ***
$\theta_2$	0.4055 ***	0.4006 ***	0.3888 ***	0.3763 ***	0.3766 ***
$\Theta_1$	-0.4618 ***	-0.5526 ***	-0.5339 ***	-0.6607 ***	-0.6555 ***
$\gamma_1$ (Level shift)		58439.3 **			
$\gamma_2$ (Level shift with trend)		-751.266 **	-258.788 **	-339.137 ***	-304.633 ***
$\beta_1$ (Employment Confidence Indicator)			-1206.42 ***	-704.939 *	-785.996 *
$\beta_2$ (Google index for “oferta de empleo”)				304.563 **	
$\beta_3$ (Google index for “oferta de trabajo”)					308.017 *
S.D. of innovations	33237.26	33043.72	32428.39	31212.74	31598.58
Akaike Criterion	2259.380	2258.660	2255.088	2249.829	2252.163
Schwarz Criterion	2269.595	2273.983	2270.412	2267.706	2270.040
Normality test Chi-square	Chi-2=2.57 p=0.27	Chi-2=1.79 p=0.40	Chi-2=1.34 p=0.51	Chi-2=2.41 p=0.30	Chi-2=2.49 p=0.29

Table 2 summarises some of the main indicators of forecasting performance. Significant improvements are observed in the ARIMAX models when Google-related variables are included. In fact, the Mean Squared Error show that Google-based models improve the forecasting accuracy in 15% with regard to the baseline model B2.

**Table 2. Evaluation statistics of unemployment forecasting in the horizon January-December 2013**

	Baseline B1	Baseline B2	Model M1	Model M2	Model M3
Root Mean Squared Error	219440	64065	67653	61639	59056
Mean Percentage Error	-3.3073	1.2408	0.1319	0.8527	0.6042
Mean Absolute Percentage Error	3.5837	1.2408	1.1794	1.1678	1.075
Theil's U	3.2791	0.9023	0.9707	0.8678	0.8289
Bias proportion	0.5191	0.8899	0.0122	0.4635	0.2593
Regression proportion	0.1412	0.0302	0.0115	0.0340	0.0511
Disturbance proportion	0.3474	0.0798	0.9763	0.5025	0.6896

#### 4. CONCLUSIONS

The internet has grown to be a major source of information not only because of all the content that is online but also because the activities that take place online can be tracked. Hence, a new line of research has emerged, with its focus on the use of internet-based data to forecast economic variables.

This paper has aimed to contribute to this field by paying attention to the use of job search-related data to the analysis of unemployment. While all the previous evidence has focused on countries with quite stable labour markets, this paper investigates whether online search-related data can be useful in the context of important economic shocks. In particular, the case of Spain is explored, a country that has been badly hit by the economic crisis.

In order to provide an accurate picture on the evolution of unemployment, ARIMAX models have been used including explanatory variables from the demand and the supply sides of the labour market: more specifically, the Employment Confidence Indicator and Google Trends indicators on “job offer” searches. Both variables are found to be statistically significant, and their estimated coefficients show the expected signs. Moreover, our estimations indicate that the inclusion of Google Trends indicators significantly improves the forecasting of unemployment’s figures.

Overall, the presented evidence highlights the importance and usefulness of internet search-related data for the nowcasting and forecasting of economic variables. Such results are especially interesting since traditional data collection is costly and many countries face important budgetary restrictions because of the economic crisis that they have not been able to overcome yet. The importance of this type of data is expected to keep growing in the nearly future. Hence, there is an increasing need to further explore the application of these new data sources in order to improve the nowcasting and forecasting of economic statistics.

## REFERENCES

- [1] H. Choi and H. Varian, Predicting the present with Google trends (2009), [http://google.com/googleblogs/pdfs/google\\_predicting\\_the\\_present.pdf](http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf).
- [2] B. Edelman, Using internet data for economic research, *Journal of Economic Perspectives* 26 (2012), 189-206.
- [3] C. Anvik and K. Gjølstad, Just google it. Forecasting Norwegian unemployment figures with web queries, Center for Research in Economics and Management Working Paper 11 (2010).
- [4] N. Askitas and K. F. Zimmermann, Google econometrics and unemployment forecasting, *Applied Economics Quarterly* 55 (2009), 107-120.
- [5] F. D’Amuri and J. Marcucci, Google it! Forecasting the US unemployment rate with a Google job search index, MPRA Paper 18732 (2010).
- [6] M. Ettredge, J. Gerdes and G. Karuga, Using web-based search data to predict macroeconomic statistics, *Communications of the ACM* 48 (2005), 87-92.
- [7] Y. Fondeur and F. Karamé, Can Google data help predict French youth unemployment?, *Economic Modelling* 30 (2013), 117-125.
- [8] Ministry of Employment and Social Security, Movimiento laboral registrado (2014), <http://www.empleo.gob.es/series/>
- [9] Google, Google trends (2014), <http://www.google.com/insights/search>
- [10] Ministry of Industry, Energy and Tourism, Encuesta de coyuntura industrial (2014), <http://www.minetur.gob.es/es-ES/IndicadoresyEstadisticas/Industria>

# Projection of road sensors to the Dutch road network

Martijn Tennekes (m.tennekes@cbs.nl)<sup>1</sup> and Marco Puts (m.puts@cbs.nl)<sup>1</sup>

**Keywords:** Visual inspection, spatial data, big data, traffic statistics

## 1. INTRODUCTION

Road sensors measure the number of passing vehicles every minute. In the Dutch network of highways, there are approximately 20 thousand of those road sensors, resulting in a huge data source. A key step in the production of traffic statistics from this data source is to project the geographic locations of the road sensors on the Dutch road network. To achieve this, road segments have to be defined based on the locations of the road sensors and subsequently the lengths of these segments.

The method that is described in this paper consists of two main parts. First, the main routes per highway per direction are deduced from the detailed road network. Second, these main routes split into road segments based on the locations of the road sensors, and the entrance and exit ramps. We illustrate that visual inspection throughout the whole process is crucial.

All geographic data inspection and editing is done in R, especially with the recently developed package tmap [1]. This package contains a flexible plotting method that is similar to ggplot2 [2], but tailored to spatial data. Also, some processing functions that were needed for this project were added to the package.

## 2. METHOD

The first part of the process is the necessary pre-processing of geographic locations of the road sensors and the road network which is described in section 2.1. The extraction of the main routes is discussed in section 2.2. Finally, in section 2.3, we describe the calculation of the road segment lengths.

### 2.1. Pre-processing geographic location data

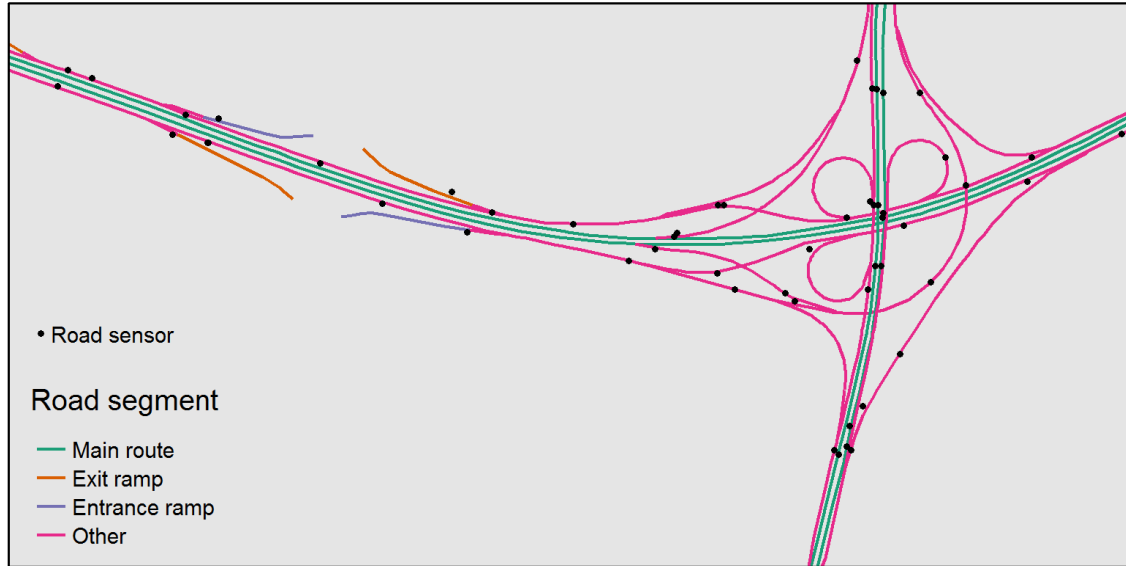
For statistical interference of geographic data, it is important to use a proper map projection. Unprojected map coordinates, known as latitude-longitude coordinates, often lead to inaccurate measurements for distances, area sizes, and directions. A good map projection preserves one or more of these properties. For the task at hand we use the Dutch National Grid (Rijksdriehoekstelsel), which is a Cartesian coordinate system that is optimized for the Netherlands. It preserves distances, which means that the difference between any two coordinates in the Netherlands corresponds approximately to the real distance in meters.

Besides the geographic coordinates, the metadata of the road sensors consist of the road names, the direction, and the type of carriageway. Only the road sensors from the main carriageways of the main Dutch highways were selected for the further process.

---

<sup>1</sup> Statistics Netherlands, Heerlen, The Netherlands

The information of the Dutch road network is contained in a ESRI shape file. It consists of almost ten thousand polylines that represent different road segments. In Figure 1, the road network including road sensors around an interchange is illustrated. Obviously, the main route segments (coloured green) are selected in order to determine the main roads (see section 2.2). The exit and entrance ramp segments are needed to determine which road sensors correspond to which main road segments (see section 2.3).



**Figure 1. Road sensors on the interchange between the A12 (horizontal) and the A27 (vertical)**

## 2.2. Main routes

Although the main routes in the road network shape file (depicted in green in Figure 1) seem fluent polylines, they cannot be used immediately to calculate main road lengths by two reasons. First of all, each main road consists of multiple polylines which should be glued together. Also, the polylines may represent parallel carriageways in the same direction, which is undesirable since our aim is to calculate the length of the main route.

We introduce the following algorithm to create the main route polylines per road per direction:

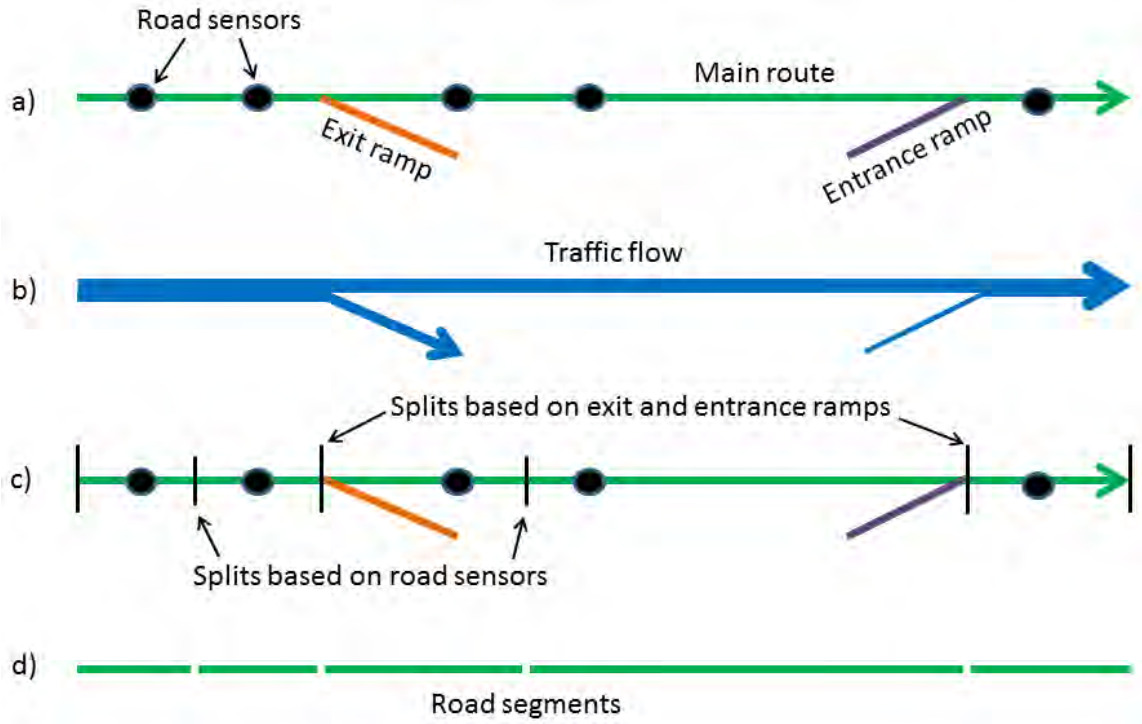
- i. All polylines get additional points at every  $d$  meters. By default, let  $d=100$ .
- ii. The points from all polylines are collected, and undoubled based on their location.
- iii. A minimum spanning tree is created from all points.
- iv. Edges that are longer than  $d$  are removed. This is only the case if the road consists of disjoint parts. The remaining components represent the disjoint parts of the road.
- v. The longest path in each component is the polyline that represents the main route. It is reversed if necessary.

Next, the geographic locations of the road sensors are confronted with the corresponding road polylines. As for the majority of the road sensors, their locations are not too far from the polyline. In those cases, the locations on the polylines at the shortest distance are determined. However, in some cases, the locations of the road sensors are many

kilometres away from the highway polyline. This is mainly caused by metadata inconsistency due to errors or difference in reference period. This inconsistency leaves two options. Either the far off road sensors are removed from the data, or the main route polylines are updated with the locations of the road sensors. In the latter case, the locations are added to the points that are collected in step ii. Since road sensors may be farther away from each other than  $d$  meters, step v is skipped.

### 2.3. Road segments

Our aim is to define the lengths of the road segments that correspond to the road sensors. Each main route is considered as a straight, one-dimensional, line. An example is given in Figure 2a. This highway consists of one exit ramp, one entrance ramp, and five road sensors.



**Figure 2. Deriving road segments: a) simplified example of a highway, b) addition of splits, c) derived road segments.**

Observe that the traffic flow along this road only depends on the exit and entrance ramps. This is illustrated in Figure 2b, where the traffic flow has three levels. Accordingly, the road is split into three segments. Each segment that contains more than one road sensor is subdivided in such a way that the middles are taken as split points. In our example, the first and the second road segment are split since each contains two road sensors (see Figure 2c). It also may occur that a part does not contain any road sensors. In that case, the statistical outcomes for that part are imputed. Hence, the five resulting road segments are depicted in Figure 2d. The lengths of these road segments correspond approximately to the real length in metres.

In order to project the road sensors to the original main routes, which is needed prior to the analysis above, a series of main route points are created, one at every 10 meters. Each road sensor is projected at the closest 10-meter-point. The points at which the exit and entrance ramps respectively diverge and converge (points of bifurcation) are determined in the same way.

### 3. RESULTS

First results show that the road network is nicely segmented based on the location of the road sensors. However, on some roads, no loops were found between two ramps, leading to a not completely covered network. Therefore, it is necessary to impute the statistical outcomes regarding the uncovered road segments.

Furthermore, also loops were found on roads that were not completely included in the road network shape file. These roads segments of these roads have been constructed with the locations of the road sensors as described as in section 2.3. However, the quality of these road segments will have to be assessed.

Finally, the total lengths of the road segments per highway have been compared to official figures [3]. The algorithm to determine the main routes, as described as in section 2.2, appears to be accurate at first sight for most highways. However, for some highways, the found differences in road length were considerably large. This problem is probably due to inconsistencies in highway definitions or reference periods.

### 4. CONCLUDING REMARKS

The method that is described in this paper can directly be used to obtain the lengths of the road segments that the road sensors represent. Therefore, it is possible to translate the processed vehicle count data that the road sensors generate, to traffic statistics, such as vehicle-kilometres, traffic intensities and traffic densities. Furthermore, it is possible to produce regional numbers, for instance based on the NUTS (Nomenclature of territorial units for statistics) [4] classification.

Obviously, the processing of the road network only needs to be executed when there are changes in the road network, or periodically, say every quarter of a year. As for the second part of the method, the definition of the road segments, it is important to know whether the road sensors function properly. If a road sensor does not work, the road segments need to be redefined.

It is important to realise that highways are not independent of each other. There are many interchanges between them, especially in the Netherlands. By the same reasoning as in section 2.3, the traffic flow on the main routes within the interchange is lower than on the main routes outside the interchange, since part of the traffic flows to the interchanging highway at the start of the interchange, and new traffic enters the main route from the interchanging highway at the end of the interchange. These middle parts can also be regarded as road segments. Analogue to the methodology described in section 2.3, the statistical outcomes of these segments can be imputed if road sensors are missing.

### REFERENCES

- [1] M. Tennekes, tmap: Themeatic Maps. R package version 0.6. <http://CRAN.R-project.org/package=tmap> (2014)
- [2] H. Wickham, ggplot2: elegant graphics for data analysis. Springer New York (2009)
- [3] Rijkswaterstaat, Actuele Wegenlijst RWS per 1 juli 2013 (2013) [https://nis.rijkswaterstaat.nl/portalcontent/logon/p2\\_32.html](https://nis.rijkswaterstaat.nl/portalcontent/logon/p2_32.html)
- [4] European Commission, NUTS - Nomenclature of territorial units for statistics (2012) [http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts\\_nomenclature/introduction](http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction)

# An exercise in producing flows statistics from big data sources

Pilar Rey del Castillo ([Pilar.Rey-del-Castillo@ec.europa.eu](mailto:Pilar.Rey-del-Castillo@ec.europa.eu)), Vidal Miguel Lázaro Toribio ([Miguel.Lazaro-Toribio@ec.europa.eu](mailto:Miguel.Lazaro-Toribio@ec.europa.eu))

**Keywords:** Big Data, short-term indicators, data editing, imputation, MapReduce

Producing reliable statistics from the huge numbers of data files created by the activity of mobile phones, remote sensors, software logs and other electronic sensing devices has become an important task for statisticians. One of the first issues that arise when using big data sources is that their size and/or complexity can make it difficult to process the data using traditional statistical procedures.

The paper presents an exercise in producing short-term indicators of the evolution of a flow variable from big data files. This type of data is typically the result of the activity of sensors counting the occurrences of specific events over time. The initial step of collecting the information needed using the MapReduce paradigm is presented first. Then standard statistical procedures used for the processing of similar flow variables are considered for data editing and imputation. Lastly index numbers are computed in order to describe the trend in the corresponding variables over time. One finding of the exercise, albeit an expected one, is that a software tool such as RHadoop allows the results to be produced very quickly, because the information can be collected and posterior analysis and processing performed in a single stage.

## 1. INTRODUCTION

It is widely accepted that the vast amount of data originating from ICT tools represents a new opportunity for official statistics. As a result, national statistical offices and international organisations have launched a range of initiatives in the area, of varying scope and focusing on different aspects.

The Sandbox is a web-accessible environment where researchers from different institutions can explore various tools and methods needed for producing statistics and thus study the feasibility of applying them to Big Data sources. The Sandbox has been established as part of a wider international cooperation project promoted by the United Nations Economic Commission for Europe [1].

The datasets uploaded on the Sandbox for the purpose of studying their potential uses, advantages and disadvantages, include pre-processed data from road sensors. In the Netherlands, the National Data Warehouse for Traffic (NDW) collects this information. Cooperation between Statistics Netherlands and the NDW has allowed data for 2012 for the roads in South Limburg to be uploaded on the Sandbox. The data represents the number of vehicles passing each minute on around 800 traffic loops in the area.

The traffic sensors are a rich source of information, providing data not only on the vehicle count but also on, e.g., speeds, distances between loops and geographical location. This allows sophisticated analysis of the traffic to be performed, at a very detailed level and with a high level of precision. This paper only studies a single variable –the number of vehicles by



minute– as an example of the analysis that could be performed for any phenomenon using this type of flow data. Similar methods could be used, for example, to compute tourism indicators using software to count the number of people caught on video-cameras located in strategic tourist locations [2].

The general aim of the exercise is to develop indicators of the evolution of the phenomena measured by the counting variables over time. The following sections present a summary of the steps taken to construct the indicators.

## 2. METHODS

One dataset has been uploaded on the Sandbox for each day in 2012. The datasets include data on the number of vehicles detected by sensors for each minute on each of the loops and lanes of the speedways in the region of South Limburg. The sensors usually have a small processing capability that allows a single message containing the sum of the readings to be forwarded to the base station, rather than each reading individually [3].

This volume of information cannot be processed using conventional statistical software and instead requires methods specifically developed for Big Data. The Sandbox infrastructure provides different options and the software RHadoop [4] has been selected. It is a combination of Hadoop, one of the leading systems for storing and performing operations on Big Data, and R, the powerful statistical programming language. Hadoop uses the MapReduce programming model, designed for processing and generating large data sets in a large number of computers –collectively referred to as a cluster– using a parallel and distributed algorithm [5].

The aim of the exercise being to construct indicators of the evolution by hour, the first step is to calculate the total number of cars by road, loop and hour. Each loop has a different number of measurements or transmissions of information to the base station, and the number of transmissions with missing information is also computed for data editing purposes. All the calculations are performed using a MapReduce algorithm.

The next step is to perform data editing to ensure completeness and validity: this will strongly depend on the specific big data source. In this case, a possible problem is that some sensor nodes may fail or be blocked due to physical damage, lack of power or environmental interference. To detect these failures, data with more than a certain proportion of missing information in the readings by hour are not validated. These data together with any other data that are missing are later imputed by means of a procedure that uses the information available for the other loops on the same road. Lastly, aggregative index numbers measuring the evolution in the number of vehicles by road over time are computed.

## 3. RESULTS

The first output from the MapReduce step is the list of the number of vehicles and the number of readings and missing readings by road, loop and hour. The number of vehicles  $x_{it}$  corresponding to the loop  $i$  for the time  $t$  (year, month and hour) with number of missing readings  $m_{it}$  and number of readings by hour  $h_i$  is validated if  $|(h_i - m_{it})/h_i| \leq 0.03$ , i. e.,

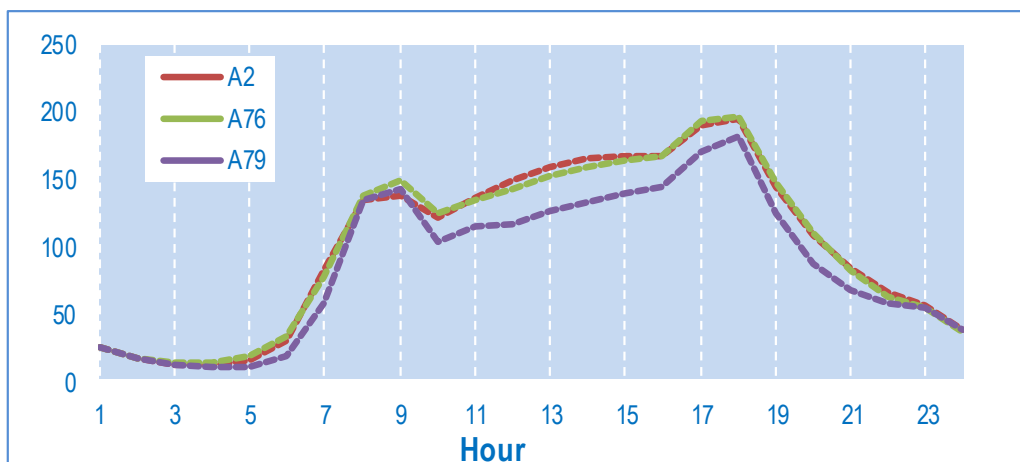
the data are only accepted if they have less than or equal to 3% missing readings. The results of the data validation process appear on Table 1:

**Table 1. Percentage of validated data**

ROAD	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<b>A2</b>	9.7	22.4	21.3	40.0	51.1	54.9	20.2	55.4	63.8	29.1	15.3	16.9
<b>A76</b>	11.8	24.5	24.4	40.4	55.2	59.0	27.0	56.4	64.3	27.0	10.0	17.4
<b>A79</b>	49.9	85.2	53.2	83.3	88.8	84.0	73.5	79.5	93.2	92.1	71.4	66.6

The missing and not validated information for the early months in the year could be imputed, for example, using backcasts from ARIMA time series models [6]. But, as the data will probably show significant seasonal patterns over the year (and information is only available for one year), it seems preferable to manage without the data for the first four months. Forecasting could also have been used to impute the missing data for the remainder of the year, but a different method that uses only the information provided by the readings was chosen instead. To apply this method, the way to compute the index numbers is considered. As they are flow data, a simple aggregative index [7] is used, that is, the index of time  $t$  with respect to  $t_0$  for each road is  $I_t = \sum_i x_{it} / \sum_i x_{it_0}$ , where the sum is extended to the  $i$  loops in the corresponding road. In practice, the indices can be computed as chain-linked ones  $I_t = I_{t-1} \cdot (\sum_k x_{kt} / \sum_k x_{kt-1})$ , where, in this case, the sum is extended to the  $k$  loops having data validated for both periods  $t$  and  $t - 1$ . The missing data  $x_{it}$  is then imputed as  $\hat{x}_{it} = x_{it-1} \cdot (\sum_k x_{kt} / \sum_k x_{kt-1})$ , considering the valid information corresponding to the same road. Using this simple method of imputation, the indices are always computed using all the information available, and are not deteriorated by a repeated lack of information on some units.

The index numbers are subsequently processed so as to give 100 averages over the year. This produces index numbers by day and hour for each road. Averages by hour, week or month and an average index for all the roads can also be calculated from these indices. As an example, Figure 1 shows the pattern of traffic by hour and road.



**Figure 1. Traffic by hour and road**

#### 4. CONCLUSIONS

An exercise of producing short-term indicators of the evolution over time from Big Data sources has been performed using RHadoop software. Although more sophisticated indicators may be more appropriate for the particular traffic phenomenon used as an example here, the simple aggregative indices could be suitable for measuring the evolution of flow variables that count the occurrences of interesting phenomena such as, for example, people walking by strategic tourist points.

An important finding is that the software tool RHadoop provides a framework for working in an efficient way, as it allows the information to be collected and posterior analysis and processing to be performed in a single step.

#### REFERENCES

- [1] United Nations Economic Commission for Europe, "The role of Big Data in the Modernisation of Statistical Production",  
<http://www1.unece.org/stat/platform/display/msis/Final+project+proposal3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production>
- [2] Barandiaran, J., Murguia, B., and Boto, F., "Real-Time People Counting Using Multiple Lines", 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), Klagenfurt, Austria, pp. 159-162, May 7-9, 2008.
- [3] Chen, W-P., and Hou, J.C., "Data Gathering and Fusion in Sensor Networks," in Handbook of Sensor Networks: Ivan Stojmenovic, ed., Wiley & Sons, 2005.
- [4] "RHadoop and MapR",  
[https://www.mapr.com/sites/default/files/rhadoop\\_and\\_mapr.pdf](https://www.mapr.com/sites/default/files/rhadoop_and_mapr.pdf)
- [5] Lämmel, R. "Google's Map Reduce programming model — Revisited", Science of Computer Programming, Vol. 70, No. 1, 2008.
- [6] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., Time Series Analysis: Forecasting and Control, 4<sup>th</sup> Edition, Wiley & Sons, July 2008.
- [7] Stone, R., and Prais, S. J., "Systems of Aggregative Index Numbers and Their Compatibility", The Economic Journal, Vol. 72, No. 247, pp. 565-583, September 1952.

# Weighting classes versus Dual System Estimation for population estimates using a census or administrative sources

Owen Abbott ([owen.abbott@ons.gov.uk](mailto:owen.abbott@ons.gov.uk))<sup>1</sup> and Helen Ross ([helen.ross@ons.gov.uk](mailto:helen.ross@ons.gov.uk))<sup>1</sup>

**Keywords:** census, population, PES, capture-recapture, non-response.

## 1. INTRODUCTION

The 2001 and 2011 UK Censuses used Dual System Estimation (DSE) as part of the methodology for estimating coverage, which in turn provided the total population estimates. The role of DSE was to estimate the non-response in the post-enumeration survey (called the Census Coverage Survey), which was undertaken shortly after the census. A ratio estimator was then applied to the adjusted survey estimates, with the census counts as the auxiliary, to estimate the total population.

It is well known that the application of DSE to human populations is problematic, as the assumptions underpinning the method are often violated. For instance, violation of the assumptions of independence, homogenous capture probabilities and linkage error all lead to biases. In the 2001 and 2011 Censuses a number of adjustments were required to mitigate against some of these biases. This has always been an issue for countries that used DSE in their coverage studies.

This paper considers an alternative weighting class approach. The key attractions of such an approach are that high quality individual level linkage is no longer required as it is replaced with linkage to the address sampling frame, and the estimator is less susceptible to over-coverage in the source used to estimate the weights (in this case the census). This type of approach may therefore be suitable for use with administrative records, which may have much higher over-coverage. Of course, the approach does have its flaws much like the DSE. The key issue is that it makes no adjustment for persons missed in addresses captured by the survey - in the UK this is about 2 to 3 per cent in a coverage survey. This paper presents some simulation studies comparing the two estimation methods, and summarises the advantages and disadvantages of both.

## 2. METHODS

### 2.1. Dual System Estimation

Given a census and corresponding Census Coverage Survey (which is carried out in a sample of postcodes), we have the following:

$X_{ij}^a$ , the count of persons in age-sex group  $a$  in household  $j$  in postcode  $i$ . The survey observes this quantity to be  $z_{ij}^a$ , and the census observes this quantity to be  $x_{ij}^a$ .

$X_i^a$ , the count of persons in age-sex group  $a$  in postcode  $i$ , observed in the census as  $x_i^a$

---

<sup>1</sup> Office for National Statistics, UK.

Define  $H_i$  to be the households included in the sample,  $S$ , in postcode  $i$ . Of these,  $R_i$  are responding households.

We can use the census to correct for non-response in the survey using a Dual System Estimator (DSE) as was used in the 2011 Census, the non-response weight being estimated as

$$\hat{w}^a = \frac{\sum_{i \in S} X_i^a}{\sum_{i \in S} M_i^a} \quad (1)$$

where  $M_i^a$  is the count of matched persons between the survey and the census from matching at individual level. The population totals can then be estimated by

$$\hat{T}_{DSE}^a = \sum_{i \in S} \hat{w}^a z_i^a \quad (2)$$

The DSE is well known to be susceptible to biases when:

- the census is inflated with over-coverage such as duplicates, as  $x$  is too large;
- there is matching error;
- there is a correlation between the probabilities of response in the census and CCS.

## 2.2. Weighting Class Estimation

An alternative is to consider a weighting class adjustment (see Lohr, 1999). In the census context, this requires us to partition the auxiliary data (the census) into classes, and also to link the census at household level to the CCS so that we know which households responded to the CCS and which did not. This is done at household level as we have information from the survey about which households responded and which did not. The linkage at household level will be crucial, as will the information about responding and non-responding households. The weighting-classes themselves are formed using the data from the census, and can be at household level or individual level.

Given that we have the age-sex group of the individuals on the census, and they can be associated with the households, an alternative is to form classes by age and sex, using the census to work out the totals by age and sex and also the number by age and sex in responding households. Therefore the weighting-classes are defined by  $a$ , and the response probability is

$$\hat{\varphi}_a = \frac{\sum_i \sum_{j \in R} w_{ij} x_{ij}^a}{\sum_i \sum_{j \in S} w_{ij} x_{ij}^a} = \frac{\sum_{i \in R} X_i^a}{\sum_{i \in S} X_i^a} \quad (3)$$

where  $w_{ij}$  is the sampling weight for household  $j$  in postcode  $i$ . The response probability is the total number of people in an age-sex group on the census in households that responded out of the total number of people in an age-sex group on the census.

As the weighting-class is defined at person level, the weights for household  $j$  in postcode  $i$  are unchanged, and effectively we have now constructed weights at individual level. For simplicity, however, the population totals can then be estimated by

$$\hat{T}_{WC}^a = \frac{\sum_{i \in S} w_i z_i^a}{\hat{\varphi}_a} \quad (4)$$

### 2.3. Simulation Study

Simulations to explore the performance of the weighting class estimator versus the DSE were undertaken. The basis for the simulations is described in ONS (2013). A scenario where the within household coverage in the CCS was perfect was included. A perfect response PCS simulation provided a benchmark. A single area was used for the study, which was a rural area with approximately 10 per cent survey non-response overall. 200 simulation replicates were used.

## 3. RESULTS

Table 1 shows the relative bias at the total population level for the simulations. Both scenarios are shown together. It shows that compared to the baseline simulation with no non-response adjustment, both the DSE and weighting class estimator have a lower bias. As expected, the DSE performs best as its assumptions have been simulated to be met. However, the weighting class method is comparable to the DSE when the effect of within household non-response is removed.

**Table 1 – Relative bias for total population estimates**

<b>Method</b>	<b>Scenario</b>	
	<b>Base</b>	<b>Base, no within hh NR</b>
Perfect	0.01%	0.01%
No NR adjustment	-10.46%	-7.83%
Age-sex weighting class	-2.59%	0.28%
DSE with over-count correction	0.16%	0.17%

## 4. CONCLUSIONS

It is clear that the weighting class approach proposed in this paper is a plausible method for making non-response adjustments to the CCS. In particular, it does not require matching at individual level unlike a dual-system estimate. However, the price paid is that it cannot make adjustments for persons who are missed from the survey even when their household responds. Similarly to DSE, it also makes assumptions of uncorrelated response and high quality matching which leads to bias when (inevitably) these assumptions are not met.

Whilst much is understood about DSE in a census context, and there are already some methods available to help account for some of the biases when it is applied in practice, the weighting class approach is less well understood. Further studies are required before any decisions are made about which estimator should be adopted for the estimation of the population.

The consideration of alternatives to a DSE is also relevant for considering the use of administrative sources for population estimation as a replacement for a census. In this context highly accurate matching may not be possible due to privacy constraints. Methods that rely less on accurate matching therefore become much more attractive. Work is underway to explore the differences in performance in this context as part of the research exploring alternatives to a traditional census.

## REFERENCES

- [1] S. L. Lohr, Sampling: Design and Analysis, Second edition. Boston: Brooks/Cole (1999), 266-267.
- [2] Office for National Statistics, Beyond 2011: Producing Population Statistics Using Administrative Data: In Theory. Methods & Policies Report (M8). Available at: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html> (2013).

# Population size estimation with different imputation techniques for incomplete covariates

**Susanna C. Gerritse**

Utrecht University

**Bart F. M. Bakker**

Statistics Netherlands, VU University Amsterdam

**Peter G. M. van der Heijden**

Utrecht University, University of Southampton

**Keywords:** Predictive Mean Matching, EM-algorithm, Capture-recapture, Census  
**Introduction**

One commonly used way to estimate the size of a population is to link the individuals from two or more registers, and then estimate the missed part of the population via capture-recapture methods [1, 2, 3, 4, 5]. Three registers of Statistics Netherlands have been linked to estimate the number of usual residents in the Netherlands. Usual residence is defined as the individuals that are staying in the Netherlands for longer than a year [6]. The three registers used are the Dutch Population Register (PR), an employee register (ER) and a crime suspects register (CSR). We are interested in usual residence as defined by a residence duration of longer than a year in the Netherlands, for which we need a measure of residence duration. We can deduce residence duration for the PR and the ER. In the PR we have a date of registration that can be used as a PR residence duration. In the ER we can use joblengths to estimate the residence duration. However, for the CSR we cannot deduce residence duration and another method has to be used.

When considering incomplete categorical data, EM algorithms to complete the categorical datasets are commonly used [4, 7]. In this project, both the EM algorithm as well as multiple imputation have been used to complete missing data [8, 9, 10]. Since the missing variable residence duration will be used as a covariate to estimate the usual residence under capture-recapture models, we need to know which imputation method is best. The two different imputation techniques used are the Expectation Maximization (EM) algorithm [9] and Predictive Mean Matching (PMM) by means of multiple imputation using chained equations [11].

## **Methods**

The PMM imputation was conducted via the program MICE in R [11]. We used the people that are registered in the ER but not in the PR (ERnPR) as donor for the residence duration for the CSRnPRxER. We have chosen this subpopulation because we assume that the ERnPR population will best resemble the individuals that are only in the PR. We are aware that this is a heavy assumption and that there is a chance this assumption is false. However, no other source is available to impute residence duration for the individuals only in the CSR. Additionally, we cannot draw a parallel between those registered in the PR and not registered in the PR, because this assumption is less likely than the CSR individuals to resemble the non-CSR ER individuals'.

In earlier research the EM algorithm has been used to complete an incomplete table with expected values according to a prespecified loglinear model [4, 7, 9]. Partially observed covariates occur when a covariate is only present in one register and not the other, and are usually ignored because they lead to missing data in one or more registers. However, not only can ignoring such a covariate lead to biased estimates [4, 7], we also need this covariate to estimate the usual residents of the Netherlands.



Using the package CAT in R we can complete categorical data that deal with missing cells, enabling the use of partially observed covariates in our analyses. The donor population here is the rest of the dataset that is fully observed. This also means that individuals in the PR will be used to impute the partially observed residence duration in the CSR. As we specified above, the CSRnPRxER population will probably more closely resemble the ERnPR population. Thus using also the PR population to impute the partially observed covariate may give biased results, and may give a higher estimate of the usual residents.

Table. 1. Observed values for the three registers.

		CSR	CSR	Total
PR	ER	1	2	
1	1	2,115	259,804	261,919
1	2	4,862	350,551	355,413
2	1	355	112,529	112,884
2	2	5,087	0	5,087
	Total	12,419	722,884	735,303

Table 1 shows the observed values for the linked registers. The zero cell in the table is a structural zero and has to be estimated and divided into usual residents and non-usual residents. This zero cell will be estimated via capture-recapture. To that end a Poisson loglinear model with a log link will be fit using generalised linear model (GLM) in the software R. The function STEP in R then selects the best fitting model. To prevent that we overfit the model, we search for a model that fits the data well and is as parsimonious as possible with the Bayesian Information Criterion (BIC). When the sample size is large the BIC criterion has a larger penalty for the number of parameters in the model than the Akaike Information Criterion (AIC) and therefore leads to more parsimonious models.

Four covariates were used in the loglinear model: Nationality group, age, sex and usual residence. Nationality group has 8 categories: (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Antillean, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American, and (8) not mentioned elsewhere. The countries are clustered according to likely migration motives, migration legislation, regulations of the PR and size. For age, we use four categories: (1) 15-24 (2) 25-34 (3) 35-49 and (4) 50-64 years of age. Sex has the categories (1) male and (2) female [10].

The number of records in the CSR that cannot be linked to the two other registers is very large. 37% percent of the individuals registered in the CSR but not in the PR and ER were missing a large amount of linkage information, such that they could not be linked. There are also erroneous captures, because in this cell, persons who do not live in the Netherlands are captured because they have committed a crime. The experts we consulted were not able to give estimates for the number of “criminal tourists”. In order to investigate the effect of both reasons we simulate that 10%,20% or 30% of the unlinked persons in the CSR is a criminal tourist.

## **Results**

Results can be found in Table. 2 and Table. 3. Table 2 shows estimates of the total population size after 0, 10, 20 and 30 percent has been taken off the CSRnPRxER individuals. Totals are provided for both the PMM as well as the EM imputed dataset. As can be seen from this table the number of estimated usual residents is lower for the PMM imputed data than for the EM imputed data. Also, given a constant 10% reduction seems to have less of an impact on the PMM imputed data than on the EM imputed data.

Table 2. The estimates for the missed portion of the population. The first three columns show the estimates after a PMM imputation, the last three columns show the estimates after an EM imputation.

	PMM				EM		
	Total n	> 1 year	< 1year		Total n	> 1 year	< 1year
No reduction	868.503	262.045	606.458		902.024	416.032	485.993
10% reduction	777.617	234.285	543.332		805.212	371.788	433.724
20% reduction	679.362	203.027	476.335		687.680	319.958	367.722
30% reduction	594.774	176.181	418.593		477.636	219.070	258.566

Table 3 shows the estimates for the population size estimate for the 8 nationality groups, split out over the number of people estimated to reside longer and shorter than a year in the Netherlands. The difference between the PMM and the EM imputed estimates can be clearly seen.

Table 3. Estimates for the missed portion of the population per nationality group, after a reduction of 30% on the CSRNPRxER cell.

	PMM			EM		
European Union	130.839	55.944	74.895	145.048	82.341	62.707
Polish	233.385	57.187	176.198	186.300	80.122	106.178
Other Europe	161.555	34.559	126.996	91.064	30.861	60.204
Other Western	15.476	4.938	10.538	9.620	3.989	5.631
Turky and Morocco	2.587	1.553	1.034	2.294	1.480	815
Afghan, Iraq, Iran.	7.487	4.601	2.886	4.529	3.091	1.438
Other Balkan, Asian	35.286	13.220	22.066	22.365	9.911	12.454
Middle East	8.159	4.179	3.980	16.415	7.274	9.141
Total	594.774	176.181	418.593	477.636	219.070	258.566

The EM imputation gives higher estimates than the PMM imputations. This is to be expected since for the EM imputation we use the distribution of the known observed values on residence duration of the PR and the ER on the unknown categories of residence duration for the CSRNPRxER. However, since it is conceivable that the CSRNPRxER population resides shorter in the Netherlands than those that are registered in the PR, and even the ER, using those registered in both the PR as well as the ER as donors to fill the unobserved residence duration values of the CSRNPRxER population might result in more usual residents than is the case.

Given our considerations on the 37% of the CSRNPRxER population that could not be linked, we assume that the 30% reduction in this cell we think it is most probable that most of these unlinkables will not belong to the population. Additionally, because of the considerations on the donor population, we assume that the estimated resulting from the PMM imputed data are more reliable than those resulting from the EM imputed data. Thus our missed portion of the population, that reside in the Netherlands for longer than a year are 176.181 individuals.

## **Conclusions**

We have seen that the EM imputation gives a higher estimate of the missed portion of the population than when we impute via PMM. This is to be expected due to the nature of the EM imputation, where we impute the unknown values with the known values of also the PR registered individuals. However, it is conceivable that the PR registered individuals

reside longer in the Netherlands, or plan to do so, hence they register themselves in the PR. In the PMM imputation, we do not use the PR registered individuals to impute the values of residence duration for the CSR population that have not been linked to either the PR or the ER. Here we used the ER registered individuals that cannot be linked to the PR to impute the residence duration of the CSR.

Additionally, the PMM imputation has been shown to be more robust also to erroneous captures, whereas the EM imputed dataset showed a much bigger change when erroneous captures were introduced. This can probably for the large part be attributed to the different donor populations. When data for covariates are missing researchers should take into consideration which imputation technique works best on their specific missing data. Even though the EM imputation is commonly used to complete incomplete categorical data, in our case the PMM imputation seems to give a more plausible population size estimate of the missed part of the population. Additionally, PMM has an advantage in being more flexible when choosing the donor population

## **References**

- [1] S., Fienberg, (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59, 409-439.
- [2] Y. Bishop, Fienberg, S., and Holland, P. (1975). *Discrete multivariate analysis, theory and practice*. New York: McGraw-Hill.
- [3] IWGDMF (International Working Group for Disease Monitoring and Forecasting), 1995, Capture- recapture and multiple record systems estimation. Part 1. History and theoretical development. *American Journal of Epidemiology*, 142, 1059-1068.
- [4] P. G. M., Van der Heijden, J. Whittaker, M. Cruyff, B. F. M. Bakker and H. N. van der Vliet, (2012). People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates, *Annals of Applied Statistics*, vol. 6, nr. 3, pp. 831-852.
- [5] B., Baffour, J.J. Brown, and P.W.F Smith, (2013). An investigation of triple system estimators in censuses. *Statistical Journal of the International Association for Official Statistics*, 29, 53-68.
- [6] European Parliament, 2008, Regulation (EC) No 763/2008 of the European Parliament and of the council of 9 July 2008 on population and housing censuses, In: *Official Journal of the European Union*, 13.8.2008, pp. L 218/14-L 218/20.
- [7] E. Zwane, and P. G. M. van der Heijden, (2007). Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registers, *Statistics in Medicine*, 26:1069–1089.
- [8] Office for National Statistics (ONS, December 2012). 2011 Census: Methods and Quality report: Item edit and imputation process,
- [9] S. C., Gerritse, P.G.M. van der Heijden & B.F.M. Bakker, (Accepted), On the robustness of the population size estimator under violation of the assumption of independence, accepted for publication in *Journal of Official Statistics*.
- [10] B F.M. Bakker , S. C. Gerritse, P. G.M. van der Heijden, D. J. van der Laan, H. N. van der Vliet, and M. Cruyff, (2014). Estimation of non-registered Usual Residents in the Netherlands, ultimo September 2010, Conference of European Statistics Stakeholders, Rome, Italy, 24 - 24 november 2014.
- [11] Van Buuren, S., 2012, *Flexible imputation of missing data* (Boca Raton: Chapman & Hall/CRC Press)

# **Measuring Uncertainty in ONS population estimates: capturing variability in statistics from combinations of census, administrative and survey sources**

**Keywords:** Statistical uncertainty, bootstrapping, cohort component method, population estimates

## **1. INTRODUCTION**

This paper describes innovative methods for measuring variability in the Office for National Statistics' (ONS') mid-year population estimates. ONS uses the cohort component method to update the mid-year population estimates each year. This involves a complex combination of data and statistical processes, combining sources including census, survey and administrative data and using a variety of estimation procedures. Estimates from these data sources have sampling and non-sampling errors, including capture and recording errors, coverage, timeliness and inter-censal drift issues. Comprehensive and quantifiable information on non-sampling error for many of these sources is not available. The complexity of trying to estimate the uncertainty in the mid-year estimates is also compounded by the fact that the data that we might use as comparators are used within the estimation process itself. To overcome this, ONS has developed a simulation-based approach. This focuses on the three components of the mid-year estimates that contribute most to uncertainty: the population base, which relies on the decennial census, and international and internal migration (For more background information, see ONS 2008-12 [1]).

Uncertainty measures for 2002-10 were published by ONS as research statistics in 2012 (see ONS 2012 [2]). We are currently working on the production of uncertainty measures for 2011-13. These will incorporate changes to reflect new methods that ONS has adopted for the production of mid-year estimates. In addition, the 2011 census has allowed us to rebase the pre-2011 mid-year estimates and evaluate the approach taken to produce uncertainty measures.

## **2. METHODS**

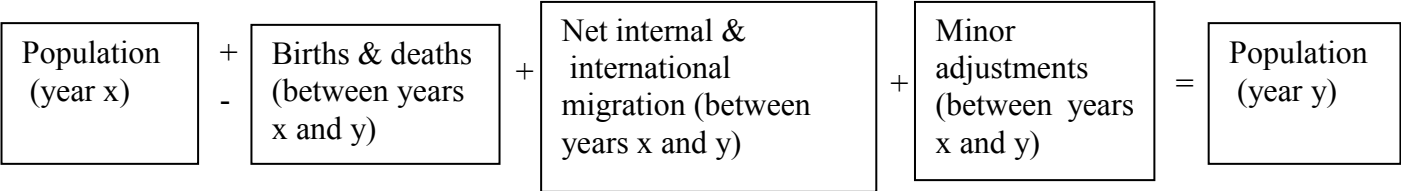
The methods to estimate uncertainty were developed by ONS in collaboration with academics at the Southampton University Statistical Science Research Institute. The uncertainty measures mirror the mid-year estimates in their basic structure.

### **2.1. Creating the mid-year population estimates using the cohort component approach.**

ONS produces annual estimates of the resident population of England and Wales as at 30 June each year. The estimates are provided at a range of geographies down to local authority level. The cohort component method is used to update these each year. In brief, components of demographic change (natural change (births less deaths), net international migration and net internal migration) are added to the previous year's aged-on population base. Minor adjustments are also applied for example for armed forces personnel. In census year, the census population is used as the population base. Rather than ageing on the population by one year, the population is only aged on by the period of time between census and 30 June. Similarly the components only need to account for

change during this period rather than the whole year (for more background information, see ONS 2014 [3]. The further we move away from the decennial census, the greater the uncertainty in the estimate becomes. The cohort component method is summarised in Figure 1.

Figure 1 Summary of the cohort component method.

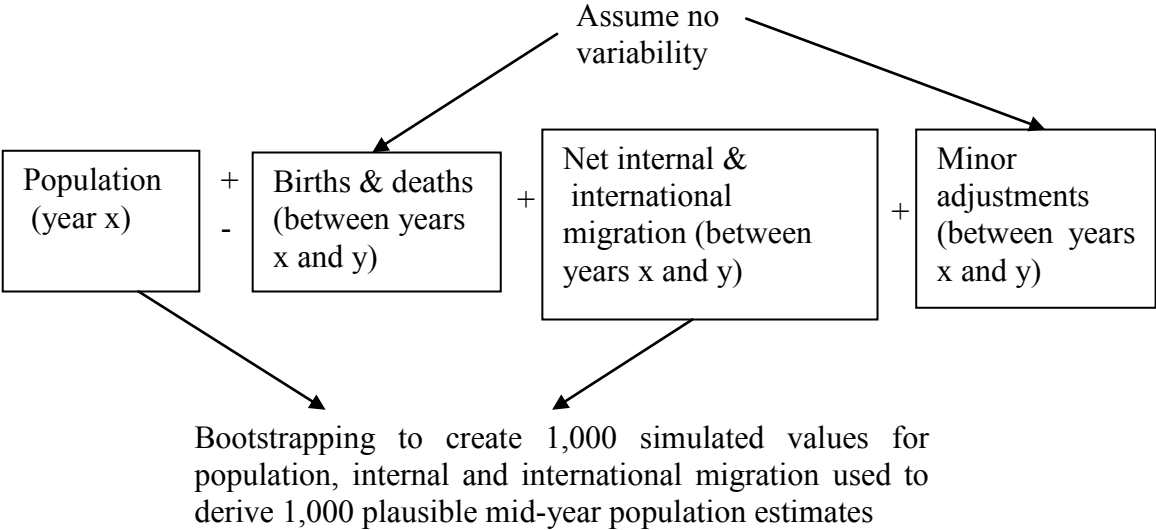


### 2.2. Measuring uncertainty in the mid-year population estimates.

Our approach focuses on the three components of the mid-year estimates that contribute most to uncertainty:

- The population base
- Internal migration
- International migration

Figure 2 Summary of the process for measuring uncertainty



The method assumes that there is no error in the births, deaths and minor change components.

For the census, international and internal migration, we drew on the observed data and replicated mid-year estimation processes to generate 1,000 possible, alternative values for each local authority. These simulations were recombined, mirroring the cohort component approach, to create 1,000 possible estimates for each local authority. Using these we were able to derive, empirically, error distributions around the mid-year

estimate for each local authority. The approach is summarised in Figure 2. The simulated estimates are rolled forward each year. This ensures that the simulated distribution (across local authorities) for the composite includes the uncertainty from previous years and new uncertainty for the current year.

### 3. RESULTS

For the 2002-10 series, uncertainty measures for the 376 local authorities in England and Wales included:

- lower and upper 95% bounds for the uncertainty range
- the uncertainty measure as a percentage of the population,
- the percentage contribution that the 2001 census, internal migration and international migration made to the overall measure of uncertainty for each local authority
- an interactive map showing uncertainty levels for local authorities by year

The uncertainty measure for a typical local authority through the last decade is illustrated in Figure 3. This shows how uncertainty grew through the decade after census year. Uncertainty in our estimates is increasingly the result of the growing influence of internal and international migration. Figure 4 demonstrates this for the local authority that is represented in Figure 3.

Figure 3 Uncertainty range around the mid-year population estimate from 2002 for a local authority within England and Wales

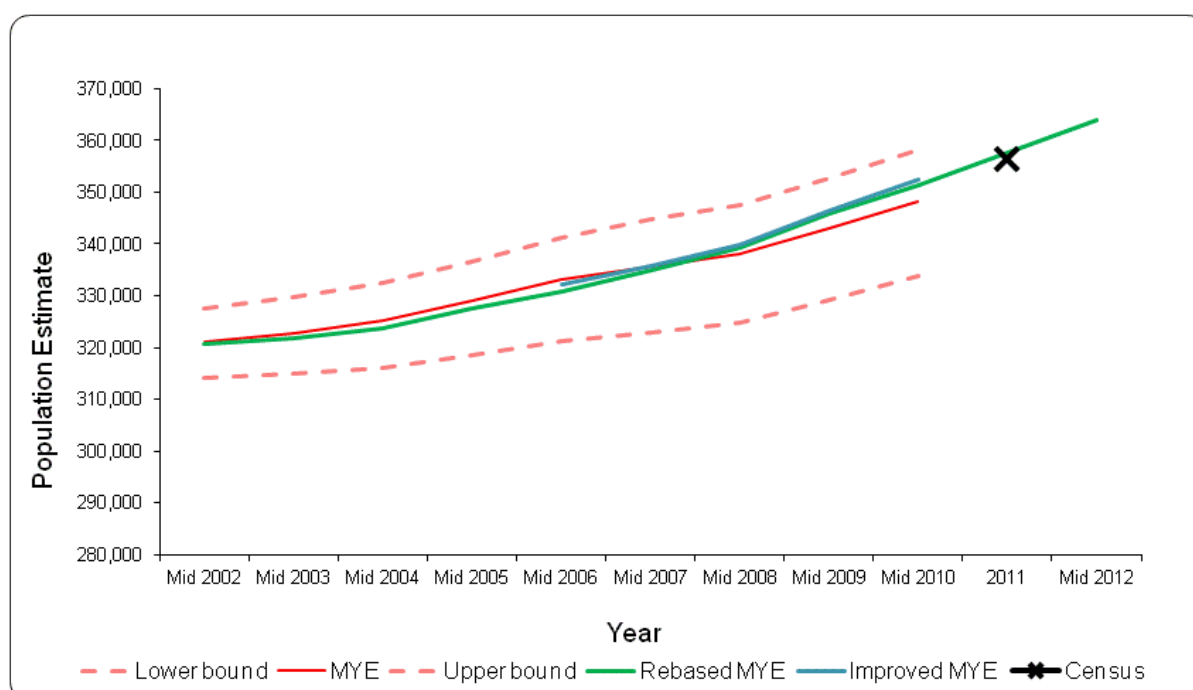
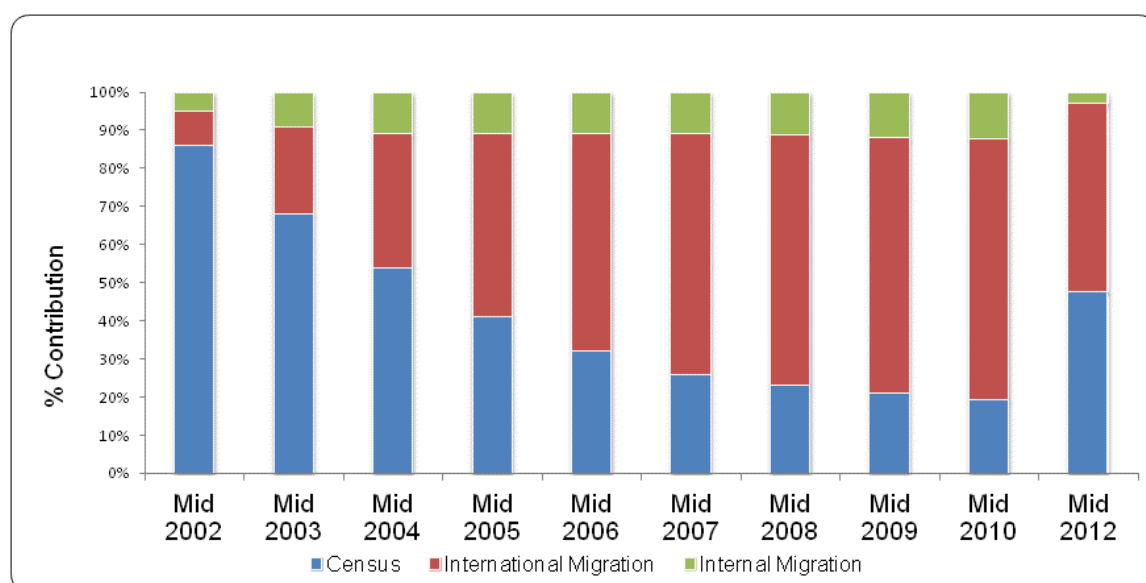


Figure 4 The proportional contribution that census, international and internal migration make to overall uncertainty for a local authority over the decade



#### 4. CONCLUSIONS

We have adapted the uncertainty measures to mirror methodological changes in the production of the mid-year estimates since 2011. The 2011 census results have provided a benchmark against which we have evaluated our approach. Similar methods could be used to measure uncertainty around other estimates that combine administrative and survey data.

#### REFERENCES

- [1] Office for National Statistics ONS (2008 – 2012) Migration Statistics Improvement Programme (MSIP) available at: <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/index.html>
- [2] Office for National Statistics (2012) *Uncertainty in local authority mid-year population estimates* available at: <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/uncertainty-in-la-mypes/index.html>
- [3] Office for National Statistics (June 2014) *Methodology Guide for Mid-2013 UK Population Estimates (England and Wales)* available at <http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/population-estimates-for-las/methods-guide-for-mid-2013-population-estimates.pdf>

# Creating a new framework for census workplace data

David Martin (D.J.Martin@soton.ac.uk)<sup>1</sup>, Samantha Cockings<sup>1</sup>, Andrew Harfoot<sup>1</sup>, Bruce Mitchell<sup>2</sup>, Ian Coady<sup>2</sup>

**Keywords:** census, small area data, automated zone design, workplace zones, geodemographic classification

## 1. INTRODUCTION

The 2011 census in England and Wales included a series of questions to individuals about their employment and principal place of work. Outputs from previous censuses have been reported primarily on a residence basis, with the 2001 small area output geography being created using automated zone design techniques, based explicitly on standardization of residential population characteristics. Although some travel-to-work data were produced, it has not therefore previously been possible to analyse worker and workplace characteristics at actual locations of work. Just four small univariate tables were produced in 2001 relating to persons at place of work. For 2011 an entirely new, additional, set of spatial units have been created specifically for the reporting of workplace statistics by re-application of the automated zone design to the record-level census data. These new spatial units are known as workplace zones (WZs) and have permitted a new family of census outputs including digital boundary data, small area tabulations and, currently, work to produce a multivariate classification of WZs. This paper will describe the construction of the new framework for census workplace data. The full paper will include detailed examples, with maps, which cannot be incorporated into this abstract for reasons of space.

## 2. METHODS

The 2001 census in England and Wales introduced the use of automated zone design techniques for the creation of the smallest census output areas (OAs). Digital boundary polygons were generated around every small postcode in the country and census results were pre-tabulated for each of these building block polygons. Using an automated zone design algorithm [1], the building blocks were iteratively recombined until OAs were produced which met multiple design criteria, including a minimum population threshold value, target population and household sizes, internal homogeneity of housing type and tenure and a spatial compactness measure. The population threshold was included in order to ensure that an appropriate level of confidentiality was maintained in any statistical tables produced for these OAs. The approach was considered to be a significant success, subsequently being adopted for the generation of larger spatial units for the publication of intercensal neighbourhood statistics and was adopted in slightly amended form for the production of OAs for the 2011 census [2], using the AZTool software (<http://www.geodata.soton.ac.uk/software/AZTool/>).

The OA geography is thus an optimised set of spatial units based on the actual distribution of residential locations. However, during the consultation phases of 2011 census preparation interest was expressed by major users, particularly in central government and the commercial sector, for an alternative set of spatial units which would be optimised for the analysis of the workplace data captured by the census questionnaire.

---

<sup>1</sup> Geography and Environment, University of Southampton, Southampton, SO17 1BJ, United Kingdom

<sup>2</sup> Office for National Statistics, Titchfield, PO15 5RR, United Kingdom



These data include not only individual characteristics such as occupation, educational qualifications, hours worked, mode and distance travelled to work but also main business of the employer. From an analytical perspective, many of these characteristics are potentially just as important aggregated to location of employment as aggregated to place of residence, enabling a detailed understanding of local labour markets and patterns of employment. The census asks for address of usual residence and also, for those in work, address of workplace of their main job. The geographical referencing codes are therefore present to produce small area aggregate statistics for workplaces but residential OAs are highly unsuitable as output zones for these data. This is due to the generally inverse spatial relationship between places of residence and employment. Census OAs were explicitly designed around target residential population sizes of 125 households, resulting in a mean residential population of 300 persons. The distribution of workers within OAs is highly skewed, with many containing no places of employment at all, while the largest, in the City of London, contained almost 80,000 employees in the 2001 census. In residential areas, the small numbers of workers would make the data far too disclosive to permit publication of useful statistics, while in major employment centres there is strong user demand for much higher resolution data on the detailed distribution of employment and worker characteristics, for example for transport and business planning.

The creation of a new set of output zones for workplace statistics is therefore an appropriate problem for the re-application of automated zone design techniques with different design criteria. An additional challenge is that the confidentiality requirements of workplace statistics are rather different to those of residential statistics. Although the number of workers can be considered as broadly equivalent to the number of residents, it is not possible to directly equate employers with households due to their very different size distribution. It is a legal requirement of the census outputs that no single employer be identifiable in the workplace statistics, just as no household should be identifiable in the residence-based statistics. Although employers are not named in the census outputs, it would potentially be possible to identify the workforce characteristics of a large employer which dominates a local labour market if appropriate measures are not taken to combine its workforce with other workers from the local area before producing aggregate statistics. Before taking the decision to commit to production of 2011-based workplace data for a new national coverage of WZs, a substantial prototyping exercise was undertaken [3] using record-level data from 2001 census returns for six different local authority districts in order to fully assess the feasibility of the project and particularly the impact of differing design criteria and protection of workplace confidentiality.

The automated zone design procedure has therefore been re-framed for the unique challenge of generating appropriate zones for the publication of 2011 census workplace statistics. This work was undertaken after the production of the residential OAs in order to ensure that there would be spatial consistency between the boundaries of the two sets of units. Digital boundary polygons were generated around the entire set of postcodes recorded in the census as workplaces (this definition includes the residential locations of individuals who work from home) and clipped to the boundaries of the 2011 OAs. Pre-tabulation of census workplace statistics to workplace polygons allowed these to be input to the AZTool software as building blocks and the zone design algorithm to be re-run using new design criteria. In this case, the design criteria were set to ensure that no WZ would contain fewer than three postcodes and 200 workers. Industry of employment was used to provide some control over internal homogeneity of WZs. Each OA was initially assessed to consider whether it could form a WZ in its own right, whether it required merging with another unit or offered the potential for subdivision. In rural and residential neighbourhoods mergers would be expected, while subdivision should be possible in most urban centres with high employment densities.

Alongside 2001 and 2011 censuses, the Office for National Statistics has also published a multivariate classification of output areas (an Output Area Classification, known as OAC [4]). This essentially involved the application of a nested k-means classification to a large multivariate OA-level dataset in order to assign each OA into three levels of a national classification hierarchy known as supergroups, groups and subgroups. The 2001 OAC has had considerable success as a free, general purpose geodemographic classification and a similar methodology has been re-applied for creation of an equivalent product based on 2011 data [5]. These classifications are effective in grouping residential areas with similar characteristics but are of course entirely based on individual and household characteristics at place of residence. The new workplace statistics afford a novel opportunity to undertake a national classification of WZs, using a broadly similar methodology, in order to produce a consistent system for identifying neighbourhoods having similar workplace characteristics, such as industrial and retail parks, town centres, centres of manufacturing industry, residential areas with high levels of homeworking, etc. There has again been considerable user interest in such a new classification, which is currently being generated from the newly-published WZ-level workplace statistics.

### **3. RESULTS**

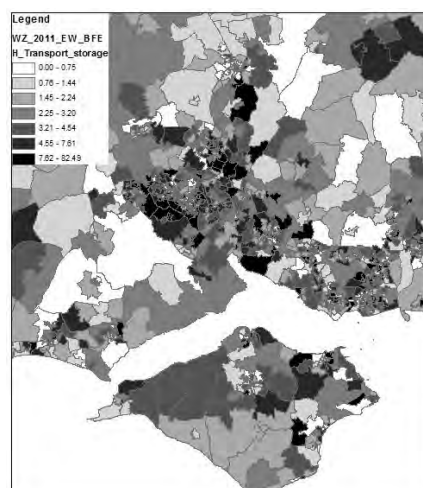
The automated zone design approach has been successfully applied to the generation of 2011 census WZs, resulting in a new digital boundary dataset for 53,578 WZs [6]. 33,206 of these WZs resulted from mergers of OAs while 16,539 resulted from subdivision of OAs – the latter being in areas of high employment density. The remainder were either unchanged (2,289) or represented more complex splits and mergers (1,544). The basic population denominator for the new WZs is persons in employment, with a mean of 493 persons in each WZ and a maximum of 11,985. Although still including a few large populations, reflecting very large employers, the framework provides enormously more spatial and statistical detail than that available from any previous census and with much higher resolution in most commercial districts.

The WZs are the basis for 23 new statistical tables including univariate (e.g. country of birth) and bivariate tables (e.g. occupation by highest level of qualification, distance travelled to work by age). These are possible due to the design control over the numbers of persons and places of employment and represent an enormous enhancement to the benefits delivered from the census workplace questions compared to the 2001 approach in which the use of residential OAs greatly restricted the information about workplace populations. In common with all other 2011 census standard outputs, the boundaries and tables have all been published as Open Government Data. The new family of WZ-based census outputs were published in 2014 and work is now under way on the generation of a standard geodemographic classification of WZs, to parallel the OA classification recently published. The WZs will also form the destinations in tables of 2011 travel-to-work data. Figure 1 provides just one example of the new outputs that have become possible, mapping the proportion of the employed population working in transport and storage industries at the WZ level for the Solent region on the south coast of England.

### **4. CONCLUSIONS**

Automated zone design techniques proved highly effective in the 2001 census of England and Wales for constructing a high resolution small area geography for the publication of aggregate census results. The same approach has been retained in 2011 for a residence-based system of small areas. The use of an automated zone design approach however offers additional opportunities to increase the value of the enumerated census data by generating a second set of geographical units for the publication of workplace-based

statistics. This approach has been successfully employed to the 2011 census outputs and boundaries and tables have been published during 2014. This is a novel approach which harnesses the potential of automated zone design in order to produce a second small area geography from the same record-level census results while controlling for a different set of design constraints and confidentiality requirements. There is now interest in the potential to further develop the new WZ geography, perhaps by aggregation to larger spatial units, to support publication of other workplace-related statistics.



**Figure 1. Percentage employment in transport and storage industries in the Solent Region, by 2011 census workplace zones**

## REFERENCES

- [1] D. Martin, A. Nolan, M. Tranmer, 'The application of zone design methodology to the 2001 UK Census' *Environment and Planning A* (2001) 33, 1949-1962
- [2] S. Cockings, A. Harfoot, D. Martin, D. Hornby 'Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales' *Environment and Planning A* (2011) 43, 2399-2418
- [3] D. Martin, S. Cockings, A. Harfoot, 'Development of a geographical framework for Census workplace data' *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2013) 176, 585-602
- [4] D. Vickers D, P Rees, 'Creating the UK National Statistics 2001 output area classification' *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, (2007) 170, 379-403
- [5] Office for National Statistics, 'Methodology note for the 2011 area classification for output areas' Titchfield, Office for National Statistics <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology.pdf>
- [6] B. Mitchell, 'Workplace zones: a new geography for workplace statistics' (2014) Titchfield, Office for National Statistics [https://geoportal.statistics.gov.uk/Docs/An\\_overview\\_of\\_workplace\\_zones\\_for\\_workplace\\_statistics\\_V1.01.zip](https://geoportal.statistics.gov.uk/Docs/An_overview_of_workplace_zones_for_workplace_statistics_V1.01.zip)

# Uncertain Population Forecasting: A Case for Practical Uses

Jakub Bijak (j.bijak@soton.ac.uk)<sup>1</sup>, Isabel Alberts<sup>2</sup>, Juha Alho<sup>3</sup>, John Bryant<sup>4</sup>, Thomas Buettner<sup>5</sup>, Jane Falkingham<sup>1</sup>, Jonathan J. Forster<sup>1</sup>, Patrick Gerland<sup>5</sup>, Thomas King<sup>6</sup>, Luca Onorante<sup>7</sup>, Nico Keilman<sup>8</sup>, Anthony O'Hagan<sup>9</sup>, Darragh Owens<sup>10</sup>, Adrian Raftery<sup>11</sup>, Hana Ševčíková<sup>11</sup>, and Peter W.F. Smith<sup>1</sup>

**Keywords:** forecast uncertainty, official statistics, population forecasting, use of forecasts

## 1. INTRODUCTION: PROBABILISTIC POPULATION FORECASTS REVISITED

In this paper we revisit the case for practical uses of the uncertainty assessments in official population forecasts, concerned with the future population size and structures.

Already in the 1970s a group of statistical demographers, uneasy with the deterministic ‘projections’, suggested the use of probability distributions to present the forecast uncertainty [1]. At the time, however, available technical resources were insufficient [2].

Since 1980s, statistical demography has been developing rapidly, especially in the area of stochastic population forecasting. Several authors have argued that probability distributions would allow the forecasts users to prepare appropriate contingency plans [3], and to make and analyse derived and conditional forecasts, when needed [4].

Despite methodological developments and recommendations, only in a few countries have probabilistic population forecasting methods been put in official statistical practice – the Netherlands and New Zealand. There have been hardly any policy applications of decision analysis or similar techniques, with an exception of [5].

In July 2014, the United Nations Population Division issued the first official probabilistic population projections for all countries, based on [6]. Much of the media coverage showed an understanding of the probabilities reported<sup>12</sup>. In this context, we want to re-open the discussion on practical advantages and obstacles of probabilistic forecasting.

## 2. CHALLENGES AND OPEN QUESTIONS

The current practice in official population forecasting is inadequate, as deterministic forecasts are bound to fail. Probabilistic forecasts, with probability distributions of possible outcomes, contain warnings about the uncertainty, which itself is an ethical virtue. On the other hand, single-number forecasts are easier to grasp in cognitive terms.

---

<sup>1</sup> University of Southampton, Southampton, UK.

<sup>2</sup> German Weather Service, Offenbach, Germany.

<sup>3</sup> University of Helsinki, Helsinki, Finland.

<sup>4</sup> Statistics New Zealand, Christchurch, New Zealand.

<sup>5</sup> United Nations Population Division, New York, NY.

<sup>6</sup> University of Newcastle, Newcastle upon Tyne, UK.

<sup>7</sup> Central Bank of Ireland, Dublin, Ireland.

<sup>8</sup> University of Oslo, Oslo-Blindern, Norway.

<sup>9</sup> University of Sheffield, Sheffield, UK.

<sup>10</sup> Aviation Training Consultant, Ireland.

<sup>11</sup> University of Washington, Seattle, WA.

<sup>12</sup> D. Carrington, “World Population to hit 11bn in 2100.” *The Guardian*; or Q. Schiermeier, “World Population Unlikely to Step Growing This Century.” *Nature News*, 18 September 2014.

There is a need for an analytical framework for supporting decisions under uncertainty. Deterministic scenarios are problematic, and answer an incorrect question: what *would* happen under given assumptions – when the real question is: what *will* happen [1, 7].

Various reasons put forward in the past for a meagre uptake of probabilistic methods (see e.g. [8]) can now be largely addressed, thanks to advances in methodology and training. The four key contemporary challenges can be found elsewhere, as mentioned below.

The first challenge is the user attitude towards forecasting uncertainty and towards risk in general. Perception of uncertainty depends on the risk attitude of the users, and can be either seen as a lack of knowledge, or as added information that can help make decisions.

The second challenge results from the specificity of user needs and circumstances. The horizons for various forecasts and decisions differ, and so do their consequences. A few pre-defined variants are thus unlikely to correspond to specific user needs.

The third challenge is to deal with information, which may be incomplete, conflicting, or superfluous. Here, the role of expert judgement and appropriate elicitation becomes crucial [9], also with respect to the utility (loss) functions of the users. The perceptions of probability or utility (loss) are not uniform, and cognitive biases are likely to occur [10].

Finally, the fourth challenge is related to validation, calibration and testing of probabilistic forecasts, chiefly through comparing them with known outcomes [3]. Here, the aim could be to minimise the errors for a model with well-calibrated uncertainty [11].

### **3. WHERE NEXT? PRACTICAL RECOMMENDATIONS**

To address these challenges, the discourse about uncertainty needs to change: from a lack of knowledge, to *additional* confidence, knowledge or information. Being explicit and transparent about uncertainty is also related to honesty, humility, and trust. This approach has proved successful in aviation, contributing to a substantial increase in safety levels.

To convince the users and producers of forecasts about the added value of uncertainty, the experience in other areas can be looked at. As population forecasts are a key input for many policy areas, they will also contribute to decisions regarding other policy variables.

Addressing the second challenge requires bespoke approaches, with forecasts tailored to the specific needs of different users [10], from high-level, strategic decision-making, to practical, operational planning, requiring quantitative input [12]. The appreciation of benefits of probabilistic forecasts can help justify the resources for their development.

Tailoring predictions, and eliciting the relevant information, requires interaction with users. The prerequisites here involve an open, two-way dialogue, with frequent exchange of information between forecasters and users, which would benefit from insights from cognitive science on statistical literacy, education and training.

As to validation, more methodological research is still needed, especially on calibrating the forecasts. Further work needs to acknowledge that time series of observations are likely not independent. In such cases, methods of risk management can be promising.

In order to achieve a paradigm shift in practical applications of probabilistic population forecasts, the focus should not be on methods, but on possible impacts and consequences of decisions. As a prerequisite, various sources of uncertainty need to be acknowledged, ideally within a joint and coherent framework, such as the one of the Bayesian statistics.

## NOTES AND ACKNOWLEDGEMENTS

The full version of this paper will be forthcoming in *Journal of Official Statistics* in 2015. The paper originates from the workshop on “The use of probabilistic forecasts with focus on population applications”, London, 19 June 2014. The financial support of the ESRC and EPSRC is gratefully acknowledged. We also thank other workshop participants for stimulating discussions. The views expressed in the paper are exclusively those of the authors, and should not be attributed to any institution with which they are affiliated.

## REFERENCES

- [1] N. Keyfitz, “On Future Population.” *Journal of the American Statistical Association* 67 (1972), 347–363.
- [2] J. M. Hoem, “Levels of Error in Population Forecasts.” Article No. 61 (1973). Oslo: Statistisk Sentralbyrå.
- [3] J. M. Alho, and B. D. Spencer, “The practical specification of the expected error of population forecasts.” *Journal of Official Statistics*, 13 (1997), 203–226.
- [4] R. D. Lee, “Probabilistic Approaches to Population Forecasting.” *Population and Development Review*, 24 (1998), 156–190.
- [5] J. M. Alho, S. E. Hougaard Jensen, and J. Lassila (eds.), “Uncertain Demographics and Fiscal Sustainability” (2008), Cambridge: Cambridge University Press.
- [6] A. E. Raftery, N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig, “Bayesian probabilistic population projections for all countries.” *Proceedings of the National Academy of Sciences*, 109 (2012): 13915–13921.
- [7] D. J. Hand, “Deconstructing Statistical Questions.” *Journal of the Royal Statistical Society A*, 157 (1994), 317–356.
- [8] W. Lutz, and J. R. Goldstein, “Introduction: How to Deal with Uncertainty in Population Forecasting?” *International Statistical Review*, 72 (2004), 1–4.
- [9] A. O’Hagan, C. E. Buck, A. Daneshkhah, et al., “Uncertain Judgements: Eliciting Expert Probabilities” (2006), Chichester: Wiley.
- [10] A. E. Raftery, “Use and Communication of Probabilistic Forecasts” (2014), mimeo, University of Washington. <http://arxiv.org/abs/1408.4812> (as of 22 August 2014).
- [11] T. Gneiting, F. Balabdaoui, and A. E. Raftery, “Probabilistic forecasts, calibration and sharpness.” *Journal of the Royal Statistical Society B*, 69 (2007), 243–268.
- [12] J. Bijak, “Forecasting International Migration in Europe: A Bayesian View” (2010), Dordrecht: Springer.

# Mass Appraisal at the Census Level - Israeli Case

Larisa Fleishman ([larisaf@cbs.gov.il](mailto:larisaf@cbs.gov.il))<sup>1</sup>, Yury Gubman ([yuryg@cbs.gov.il](mailto:yuryg@cbs.gov.il))<sup>1</sup>

**Keywords:** Prediction model, dwelling value, register statistics, quantile regression

## 1. Introduction

Information about the value of the entire national housing stock may serve as a basis for the development of various statistical outputs, such as small area estimation using dwelling data at a census level, housing and land value estimators in high spatial resolution. This data may be also used for improving estimates of economic welfare and inequality, and for longitudinal analyses, for instance in respect to studying trends in household savings over the years.

In Israel, the main source of information about dwelling value is the record of real-estate transaction prices kept by the Israel Tax Authority (ITA). Apart from prices of dwellings sold during a given period, these files include an assortment of property characteristics such as area, number of rooms, story, and others. In addition, the Household Expenditure Survey (HES), conducted by the Central Bureau of Statistics annually, provides information about subjective valuation of dwelling value by dwelling owners.

In our paper, we develop a methodology for the estimation of dwelling values for all types of dwellings at a nationwide level. We consider and treat the issue of representativeness of the existed data sources on dwelling values. Clustering procedure based on quantile regression analysis along with geographical stratification allows improving the prediction accuracy. The developed methodology was applied for privately owned residential properties in localities that have populations in excess of 2,000, in Census Tracts (CTs) in which more than 50 percent of residents are Jews (hereinafter: “the Jewish urban sector”), for 2011. It should be noted that for Arab sector, very small number of transactions is available. Moreover, the physical, environmental, socio-economic and socio-cultural characteristics of the Israeli Arab housing market constitute a major factor in differentiating it from the Jewish housing market. Thus, the dwelling value estimation in the Arab sector should be based on different methodology fitted for the current stage of housing market in this sector. The development and implementation of such methodology is beyond the scope of this paper.

## 2 Mass Appraisal

In most cases, a hedonic model for housing prices is the underpinning of estimation methods for dwelling value prediction. In most cases, information about dwelling value is harvested from files of transactions actually consummated in the specified period [1]. Yet, this data may not be a representative sample of the entire dwelling stock because it pertains only to sale transactions that match homeowners’ valuations. This condition does not prevail among other dwellings in the region that were offered for sale but not sold. This issue, known in the literature as sample selection bias, may also skew a dwelling valuation produced by a statistical model based on transaction data [2]. As for subjective dwelling valuation, researchers claim that no such bias exists since such data reflect the price level more accurately by yielding a representative random population of properties sampled in surveys [3]. Subjectively estimated dwelling valuations that are reported in surveys, however, are susceptible to an upward bias [4].

---

<sup>1</sup> Israeli Central Bureau of Statistics

The literature proposes that a distinction should be made between models for the explanation of the phenomenon and models for its prediction [5]. In the latter case, the main criterion for the insertion of a variable is its forecasting ability as opposed to explanatory ability or the statistical significance of the regression coefficient in the former. In several studies, a division into estimation cells is proposed [6]. To examine estimation quality and choose the best model, the studies noted above use several indices. Mean Absolute Percentage Error (MAPE), Median Absolute Percentage Error (MedAPE), etc. The MAPE index is the most common; the smaller its value is, the more accurate the prediction.

### 3. Methodology

In Stage 1, a hedonic model was used to investigate the representativeness of the transaction file, with the economic conditions and the annual number of sales transactions treated as givens. In this matter, differences in the effect of the characteristics of a property on its value were examined for dwellings sold versus dwellings not sold. In Stage 2, based on the results of the foregoing analysis, models were fitted for the estimation of the entire population of dwellings. Then, with the help of accuracy indices the best prediction model was chosen. The model used in the current study is given by:

$$(1) \quad Y_{ijkl} = \log(P_{ijkl}) = \lambda_0 + \lambda_1 Asset_i + \lambda_2 Building_j + \lambda_3 CT_k + \lambda_4 Locality_l + u_{ijkl}$$

where  $P_{ijkl}$  denotes the value of property  $i$  in building  $j$  in region  $k$  and locality  $l$ .  $Asset_i$  denotes the characteristics of dwelling  $i$  (area, number of rooms, property tax rate),  $Building_j$  - the characteristics of building  $j$  (year of construction, building type, number of dwellings in building),  $CT_k$  - locational variables of a census tract  $k$  and residents' demographic and economic characteristics (number of buildings, residents' median age, residents' average annual income, proximity of CT to center of locality, proximity of CT to Tel Aviv city, and whether CT fronts the sea;  $Locality_l$  - locality characteristics (population and district of locality); and  $u_{ijkl}$  - random noise with variance  $\sigma$ . After log transformation, the explained variable is approximately normally-distributed, justifying the use of the ordinary least squares (OLS) method to estimate Equation (1).

It was found that the coefficients of the hedonic model for the ITA transaction data resemble those of HES. Thus, both of these data sources may potentially serve as a basis for mass appraisal. However, the literature on the subject supports the premise that transaction price is the best proxy for property market value [7]. Thus, the model based on ITA data may be used for mass appraisal of the entire housing stock in a given area and at a certain time. As well, we can conclude that underrepresentation of certain types of dwellings in the ITA data is unlikely to be crucial. Yet this issue is worth considering at a prediction stage.

Differences in dwelling price level between geographical districts (defined according the official administrative division) are found to be statistically significant in a pairwise comparison between every two districts (using t-test). These findings support geographical stratification.

Quantile regression analysis was performed to test the distribution of the estimated coefficients over centiles of the dependent variable. In this method, the regression coefficients were calculated for specific percentiles of distribution of the dependent variable [8]. It was found that the coefficients of the regression model are stable enough across most of the distribution: from the 10<sup>th</sup> centile to approximately the 95<sup>th</sup>. At the extremes, the coefficients are unstable and differ from the values obtained in the middle of the distribution range.



The results of these analyses suggest the following stratification: (1) inexpensive dwellings (lowest decile, all districts) (2) expensive dwellings (five uppermost centiles, all districts), and (3) seven cells differentiated by district for dwellings remaining after the removal of those in (1) and (2). This division is consistent with the literature as it takes account of not only the spatial aspect, but also the differences in the contribution of the factors that affect the prices of most expensive and inexpensive dwellings as against the rest of the housing stock [9].

#### 4. Results and conclusions

A prediction model was estimated independently for each cell. Table 1 shows the distribution of the absolute percentage errors calculated by the cross-validation method. The accuracy indices presented in Table 1 are superior to those reported in most of the studies.

**Table 1: Distribution of absolute percentage errors**

Index	Mean (MAPE)	Median (MedAPE)	Percentiles			
			10	25	75	90
<b>Value (pct.)</b>	20.12	12.28	2.21	5.65	22.44	36.86

Table 2 contrasts the distribution of predicted values for 2011 with the distributions of transaction prices and subjective valuation data for the same year.

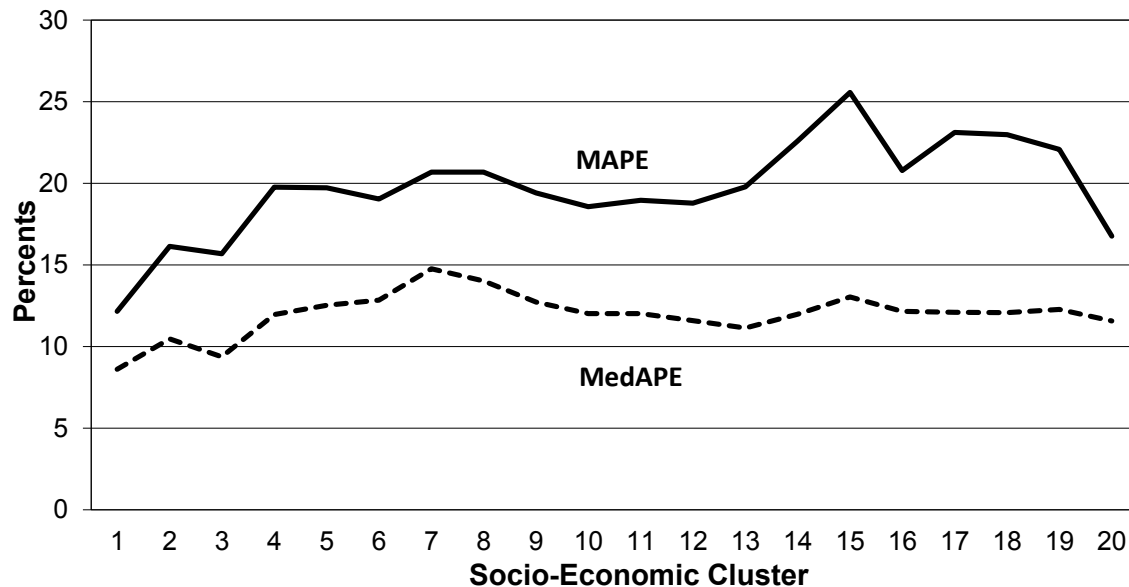
**Table 2: Distribution of predicted values, actual prices, and subjective valuations**

Source	Index	Mean	Median	10 <sup>th</sup> percentile	25 <sup>th</sup> percentile	75 <sup>th</sup> percentile	90 <sup>th</sup> percentile
Predicted values		862,26,,1	86,1968,2	4716954	5876919	867796185	268516789
ITA data		86,9268,2	9,,6,,,	87,6,,,	77,6,,,	8687,6,,,	8694,6,,8
HES data		864996789	868,,6,,,	,276,,,	9,,6,,,	861,,6,,,	267,,6,,,

The estimated standard deviation of the predicted values was NIS 16,079 on average, about 1.3 percent of the mean value presented in Table 2. It shows that the proposed predictors are sufficiently stable.

Property price level in a given residential area is often analyzed in context of its socio-economic profile [10]. This approach stems from the well-known correlation between property prices and various effects reflecting socio-economic characteristics of population in a given area. Figure 1 shows the distribution of MAPE and MedAPE indices over CT's socio-economic clusters. Figure 1 demonstrates that there is certain variance in the accuracy indices over socio-economic clusters. In particular, the attained accuracy of the predicted values (MAPE) is greater for the socio-economically strongest and weakest regions than in the remaining regions, justifying the estimation by cells parsed by the value of properties in the region as described above. MedAPE curve shows that the robust estimator of accuracy in all clusters is smaller than MAPE index and no meaningful variance in its distribution is observed among the clusters.

In order to examine the performance of the proposed methodology at different time points, it has been applied to the 2012 and 2013 data. It appears that the proposed method applied to additional time points gains the accuracy indices stable during the addressed period (2011-2013).



**Figure 1: Distribution of the accuracy indices by CT's socio-economic cluster**

Dwelling values estimated at the nationwide level allow one to produce new statistical products at high geographic resolutions on a range of topics, e.g., the behavior of the housing market, the economic profile of residential areas, and welfare and inequality, to name only a few. Value data at the individual-record level also facilitate estimations for small geographical units, and population groups defined by socio-economic and demographic characteristics.

## Sources

- [1] A.C. Goodman and T.G. Thibodeau, Housing market segmentation and hedonic prediction accuracy, *Journal of Housing Economics*, 12(3), (2003), pp.181-201.
- [2] N. Nguyen and A. Cripps, Predicting housing value: A comparison of multiple regression analysis and artificial neural networks, *Journal of Real Estate Research*, 22(3), (2001), pp.313-336.
- [3] J.E. Zabel, Controlling for quality in house indices, *Journal of Real Estate Finance and Economic*, 19, (1999), pp.223-241.
- [4] J.L. Goodman and J.B. Ittner, The Accuracy of home owners' estimates of house value, *Journal of Housing Economics*, 2, (1992), pp.339-357.
- [5] G. Shmueli, To explain or to predict? *Statistical Science*, 25(3), (2010), pp.289-310.
- [6] N. Lozano-Gracia and L. Anselin, Is the price right?: Assessing estimates of cadastral values for Bogota, Colombia, *Regional Science Policy & Practice*, 4(4), (2012), pp.495-508.
- [7] H.S. Banzhaf and O. Farooque, Interjurisdictional housing prices and spatial amenities: Which measures of housing prices reflect local public goods? *Regional Science and Urban Economics*, 43(4), (2013), pp.635-648.
- [8] R. Koenker and Jr.G. Bassett, Regression quantiles, *Econometrica: Journal of the Econometric Society*, (1978), pp.33-50.
- [9] R. Reed, The contribution of social area analysis: modelling house price variations at the neighbourhood level in Australia, *International Journal of Housing Markets and Analysis*, 6(4), (2013), pp. 455-472.
- [10] F. Des Rosiers, M. Theriault, Y. Kestens and P. Villeneuve, Landscaping and House Values: An Empirical Investigation. *Journal of Real Estate Research*, 23, (2002), pp. 139-161.

# **A European toolbox for a modular design and pooled analysis of social survey programmes**

Martin Karlberg (Martin.Karlberg@ec.europa.eu)<sup>1</sup>, Fernando Reis<sup>1</sup>, Cristina Calizzani<sup>1</sup>, Fabrice Gras<sup>1</sup>

**Keywords:** Streamlining, modularisation, integration, pooling, recomposition.

## **1. INTRODUCTION**

### **1.1. The European Social Surveys**

Collectively, the European Social Surveys aim to capture a wide range of societal phenomena, such as unemployment, health, education, income and living conditions, rendering European social statistics which meet the quality criteria of official statistics as set out in the European Statistics Code of Practice. The current core social surveys are: the Labour Force Survey (LFS), the European Statistics on Income and Living Conditions (SILC), the Adult Education Survey (AES), the European Health Interview Survey (EHIS), the Information and Communication Technology household survey [ICT(HH)], the Household Budget Survey (HBS) and the Time Use Survey (TUS).

Historically, needs for new European social statistics have been addressed in three different ways: (i) needs in a statistical domain with an established survey are addressed by expanding it (e.g. more detailed living condition information needs being covered by additional items in the EU Survey on Income and Living Conditions); (ii) punctual needs covering a few pieces of information in a new domain or in a domain without an established survey are addressed by incorporating a few additional items in a larger European Survey (e.g. lifelong learning variables from the education statistics domain were added to the Labour Force Survey in 2003 since no dedicated European adult education survey existed at that time); (iii) major needs in a new domain or in a domain without an established survey are addressed by setting up an altogether new European Social Survey (e.g. the harmonised AES was set up at European level in 2007).

### **1.2. Problem Statement**

There are certain drawbacks with the current system. For instance, strategies (i) and (ii) of adding items to existing surveys tend to lead to rather large “overloaded” questionnaires – and sometimes, oversampling of certain items may occur if they are piggy-backed onto a larger survey without subsampling. Still, this is often the adopted solution due to the fact that option (iii) of setting up an altogether new European Social Survey is a major operation, with a considerable cost and a long lead time (e.g. the development of the harmonised European Adult Education Survey took around 4 years between 2004 and 2007 and the first time the survey was carried out at the same time in all countries was in 2010, i.e. six years after the inception of the AES). Other inefficiencies result from “the same question being asked” in different social surveys – without this information being pooled to render estimates of higher precision, as well as “almost the same” question being asked in different social surveys (e.g. education participation of an individual is asked with reference to the last 12 months in the AES, with reference to the last 4 weeks in the LFS and simply asking for the studentship status at the time of the interview in SILC), possibly leading to seemingly inconsistent presentations of what is in essence the same phenomenon.

---

<sup>1</sup> Eurostat, European Commission, L-2920 Luxembourg

The need for improvement and modernisation is acknowledged by the European Statistical System (ESS); the 2011 Wiesbaden Memorandum calls for the ESS to strengthen the capacity for reaction and adaptation, and presents efficiency gains in the production of social statistics as the optimal way to address the emerging needs for new or improved social statistics.

### **1.3. Contribution of this paper**

While the Wiesbaden Memorandum requires a *common architecture* for European social statistics, and calls for a *streamlining* of the core social surveys conducting microdata collection on persons and households, it does, understandably, leave the precise details on how to achieve these targets open. Various initiatives have been launched, and in this paper, we will present the solutions provided in a study [1] commissioned by Eurostat.

## **2. METHODS**

### **2.1. The envisaged new system – requiring harmonisation and modularisation**

The methods developed in the study are only applicable if the variables to which they are applied have been harmonised (the “student” concept discussed in section 1.2 being one example) across the European Social Surveys. ESS work on this is currently underway.

Under the new system, the (almost 3000) variables of the European Social Surveys would be allocated into (say 150-200) mutually exclusive groups of (say 10-15) variables, called *modules*. How these modules actually will be designed is being addressed elsewhere [2], but it should be noted that for the “recomposition” methods proposed below (see Section 2.4) to be applicable, it is necessary that this modularisation exercise has taken place for the European Social Surveys.

The modules would be distributed into *instruments* in such a way that each instrument consists of a fixed set of standard re-usable modules and each module may be present in many instruments. These instruments would thus replace the current set of European survey questionnaires and normally be higher in number than today. The current system of social surveys could be viewed as a special case of the envisaged future system.

Data compilation for each instrument may take place independently of the other instruments or with different degrees of coordination with them. The estimation of target parameters will be based on pooling (see Section 2.2) of the required input data from all instruments with which they have been compiled. This is expected to produce benefits in the form of reduced response burden, reduced sample size, reduced cost, increased precision and increased analytical potential of the data. Moreover, the system makes it possible to (i) reuse modules in different instruments, thereby creating “crossings” allowing for the simultaneous observation of two modules from different domains in a single instrument, and to (ii) integrate new modules in the system with a short lead time. This would allow it to respond quickly and efficiently to emerging needs for statistical information.

### **2.2. Pooling**

To make use of the fact that the same question is asked in different social surveys, the study has identified methods currently available for pooling data across surveys in such a way that the precision is increased, and estimates that are consistent between surveys are obtained. The methods of Merkouris [3] have been identified as suitable for microdata, while the methods of Renssen and Nieuwenbroek [4], which take point and variance estimates as inputs, are possible to use in case microdata are not available.

### 2.3. Allocation

Assuming as given the following general setup:

- (i) a certain set of instruments has been defined,
- (ii) there may be overlaps between different instruments (in the sense that the same module may be present in several instruments),
- (iii) precision criteria associated with each module and each “crossing” (i.e. each set of modules which have to be simultaneously observed) have been specified,

the study demonstrates that (given a cost function reflecting NSI operating cost and/or response burden), the simplex algorithm can be applied to arrive at an optimal allocation of sample sizes of instruments (i.e. the one yielding the lowest cost while respecting the minimum effective sample size criteria).

### 2.4. Composition of instruments

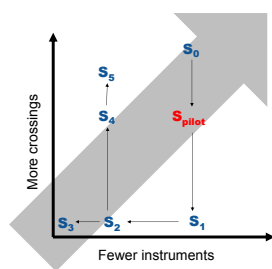
Once the European Social surveys have been modularised, i.e. their micro level variables have been organised into modules, the number of possible ways in which the original survey questionnaires could be recomposed into instruments is enormous. Thus, while the algorithm of Section 2.3 is easily applied if the set of instruments is defined, it doesn't in itself provide any guidance as to what the best way to compose instruments is.

In the study, this is tackled by means of *simulated annealing*, which is a random search algorithm. Based on a pre-specified number of instruments, “admissibility criteria” and precision requirements, the algorithm starts from a given scenario (any “admissible” instruments composition) and randomly tries out alternative “admissible compositions”, calculating their costs as outlined in Section 2.3.

## 3. RESULTS

For the purpose of investigating the characteristics of the instrument composition procedure, it was tried out in a simulation study [5] for a subset of LFS, SILC and AES variables, using precision criteria from a mid-sized EU member state.  $S_0$ , the baseline scenario used, closely mimics the current organisation of the three social surveys. Five alternative scenarios ( $S_1$ ,  $S_2$ ,  $S_3$ ,  $S_4$ ,  $S_5$ ), each one with a different set of design constraints, were tried out. Some of them ( $S_2$ ,  $S_3$ ,  $S_4$ ,  $S_5$ ) allowed for a larger number of instruments in relation to today's surveys. All of the alternative scenarios imposed fewer requirements (in relation to  $S_0$ ) concerning variables that have to be together in one questionnaire. One scenario ( $S_5$ ) offers a greater analytical potential by including crossings for which needs were expressed in a user survey. The expected outcome of the study would be that:

- (i) when additional crossings are added to the requirements, more data collection constraints improving the analytical potential are introduced, and cost will thus increase;
- (ii) when more instrument become possible, this translates to a lifting of data collection restrictions: modules with low sample size requirements are not “stuck” in instruments with a large sample size.



#### **S<sub>pilot</sub> characteristics:**

**Close to S<sub>0</sub>, i.e. the current split in LFS, SILC and AES**

#### **Difference: fewer crossings**

Instead of requiring all LFS variables to be simultaneously observed, only the “most popular crossing” (i.e. those downloaded most frequently from Eurostat) are required

**Result:** Some modules can move between instruments (no longer “stuck” in one “survey”) → **lower costs**

Some modules end up with “new neighbours” (added bonus) yielding new crossings → **New analysis possibilities**

**Figure 1. Outcome of the application of the instrument composition algorithm for various scenarios for a hypothetical subset of LFS, SILC and AES modules.** The grey shaded arrow represents the cost of the optimal solution of each scenario.

The actual outcome in terms of cost (as represented by the grey diagonal arrow) for the scenarios described above is seen in Figure 1. For instance, S<sub>1</sub>, which requires fewer crossings than S<sub>0</sub>, does render a less costly solution – and S<sub>2</sub>, which allows for more instruments than S<sub>1</sub>, does also decrease cost. Since S<sub>5</sub> doesn’t require that all currently co-administered variables are put together in the same instrument, it is less costly than the baseline scenario, although new possibilities for analysis (new crossings) are added.

In addition, for the purpose of a possible future pilot study, a “pilot” scenario S<sub>pilot</sub> rather close to today’s situation (S<sub>0</sub>) was also investigated. As expected, Figure 1 confirms it to be slightly less costly than the baseline scenario S<sub>0</sub>.

## **4. CONCLUSIONS AND NEXT STEPS**

The study that we have conducted demonstrates that a modular system of social surveys would be supported by a methodological toolbox for (i) pooling data between instruments, thus making maximum use of the information collected, (ii) shifting sample units between instruments to reduce overall cost while respecting statistical information needs (including precision criteria) and (iii) composing instruments from the set of modules available in a flexible way. By progressively moving from a system of “survey stovepipes” (like the current one) to an integrated social surveys architecture based on the concept of “instruments”, we would be able to achieve efficiency gains and flexibly meet new information needs by efficiently accommodating increased information requirements in the form of new variables and new crossings. The new system could be phased in, starting out with pooling of data between surveys, and only proceed to more complex issues such as instrument composition, at a later stage.

Numerous methodological challenges still exists (e.g. the full integration of longitudinal aspects in the algorithms, and the incorporation of complex indicators such as poverty rates, alternatives to the “simulated annealing” search algorithm), and in practical terms, an interface allowing “qualified practitioners” to design the instruments by specifying business rules would still be needed. Then, as experiences elsewhere [5] have shown, there are obviously a wide range of operational concerns to take into account for the implementation of an integrated survey architecture.

However, in spite of the remaining methodological challenges, much of the efforts needed if an integrated system is to see the light of day are now on the “content side”; the tools are there, and the various stakeholders need to agree on common definitions of the variables used in two or more European Social Surveys and define the various modules.

## REFERENCES

- [1] Stavropoulos, P. (2014), *Development of methods and scenarios for an integrated system of European Social Surveys – Final Report*. Produced by Agilis S.A. under Eurostat contract 61001.2011.005-2012.426.
- [2] Reis, F. (2013). *Links Between Centralisation of Data Collection and Survey Integration in the Context of the Industrialisation of Statistical Production*; WP2 presented at the UNECE Seminar on Statistical Data Collection.
- [3] Merkouris, T. (2010) “An Estimation Method for Matrix Survey Sampling” *ASA, Proceedings of the Section on Survey Research Methods*, 4880–4886.
- [4] Renssen, R. H. and Nieuwenbroek, N. J. (1997), “Aligning Estimates for Common Variables in Two or More Sample Surveys,” *JASA* 92, 368–375.
- [5] Ioannidis, E. (2014), *Instrument Design and Sample Size for the Pilot*. (Deliverable of the contract reported in [1].)
- [6] Smith, P. (2009) ”Survey harmonisation in official household surveys in the United Kingdom”, *ISI Invited Paper Proceedings*. Durban, South Africa.

# Sampling coordination of business surveys: a new method implemented at INSEE

Emmanuel GROS<sup>1</sup> ([emmanuel.gros@insee.fr](mailto:emmanuel.gros@insee.fr))

**Keywords:** Sampling coordination, permanent random numbers, coordination function, response burden, stratified sampling.

## 1. INTRODUCTION

The public statistical system carries out each year a significant number of businesses and establishments surveys. The objective of the negative coordination of samples is to foster, when selecting a sample, the selection of businesses that have not already been selected in recent surveys, while preserving the unbiasedness of the samples. This coordination contributes to reduce the statistical burden of small businesses – large businesses, from a certain threshold, are systematically surveyed in most surveys.

This paper presents the new sampling coordination method currently used at Insee. This method, using Permanent Random Numbers (PRN) assigned to each unit, is based on the notion of coordination function, defined for each unit and each new drawing, which transforms permanent random numbers.

## 2. METHODS

We present here the main principles of the method, detailed in [1], limited to the case of stratified simple random sampling. This method was proposed by C. Hesse in 2001 in [2], and studied by P. Ardilly in 2009 in [3].

### 2.1. A PRN method resting on the concept of coordination functions

The concept of coordination function plays an essential role in the method.

A coordination function  $g$  is a measurable function from  $[0,1]$  onto itself, which preserves uniform probability: if  $P$  is the uniform probability on  $[0,1]$ , then the image probability  $P^g$  is  $P$ . It means that for any interval  $I = [a, b[$  included in  $[0,1]$  :

$$P[g^{-1}(I)] \stackrel{\text{def}}{=} P^g(I) = P(I) = b - a$$

The length of the inverse image of any interval under  $g$  equals the length of this interval: a coordination function preserves the length of intervals – or union of intervals – by inverse image.

Each unit  $k$  of the population is given a permanent random number  $\omega_k$ , drawn according to the uniform distribution on the interval  $[0,1]$ . The drawings of the  $\omega_k$  are mutually independent.

We consider a sequence of surveys  $t = 1, 2, \dots$  ( $t$  refers to the date and the number of the survey), and we denote by  $S_t$  the sample corresponding to survey  $t$ . Suppose that one has defined for each unit  $k$  a “wisely chosen” coordination function (see 2.2)  $g_{k,t}$  which changes at each survey  $t$ .

---

<sup>1</sup> INSEE – France's National Institute for Statistics and Economic Studies – 18 boulevard Adolphe Pinard, 75675 Paris Cedex 14, FRANCE



The drawing of the sample  $S_t$  by stratified simple random sampling is done by selecting, within each stratum  $(h,t)$  of size  $N_{(h,t)}$ , the  $n_{(h,t)}$  units associated with the  $n_{(h,t)}$  smallest numbers  $g_{k,t}(\omega_k)$ ,  $k=1\dots N_{(h,t)}$ .

### **Proof**

The  $N_{(h,t)}$  random numbers  $(\omega_k)$  associated to the  $N_{(h,t)}$  units of the stratum have been independently selected according to the uniform probability on  $[0,1]$ , denoted  $P$ . Since we have  $P^{g_{k,t}} = P$  for each  $k$ , the  $N$  numbers  $g_{k,t}(\omega_k)$  are also independently selected according to  $P$ . Then, using a well-known result, the  $n_{(h,t)}$  smallest values  $g_{k,t}(\omega_k)$  give a simple random sample of size  $n_{(h,t)}$  in the stratum.

## **2.2. Construction of a coordination function from the cumulative response burden**

**① Response burden and coordination function:** Let  $\Omega$  denote the vector of random numbers  $\omega_k$  given to the population units  $k$ , and  $\gamma_{k,t}$  be the response burden of a questioned business  $k$  at survey  $t$ . The cumulative burden for unit  $k$  is a random variable, function of  $\Omega$ , equal to:

$$\Gamma_{k,t}(\Omega) = \sum_{u \leq t} \gamma_{k,u} \cdot \mathbb{I}_{k \in S_u}(\Omega) \quad (1)$$

We wish to define, for each unit  $k$ , a coordination function  $g_{k,t}$  based on  $\Gamma_{k,t-1}$ , the cumulative burden of unit  $k$  until survey  $t-1$ . To meet the objective of negative coordination – to draw as a priority, for a given sample selection, units that have had the lowest response burden during the recent period – and taking into account the selection scheme of the units – the higher the probability for the unit to be selected the smaller the number  $g_{k,t}(\omega_k)$  –, a desirable property for any coordination function is the following:

$$\Gamma_{k,t-1}(\Omega^{(1)}) < \Gamma_{k,t-1}(\Omega^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \leq g_{k,t}(\omega_k^{(2)})$$

where  $\omega_k^{(i)}$  ( $i=1,2$ ) denotes the  $k^{\text{th}}$  component of vector  $\Omega^{(i)}$ . This condition is not easy to handle, because the function  $\Gamma_{k,t-1}(\Omega)$  is a function of vector  $\Omega$ : it depends not only on the random number  $\omega_k$  given to unit  $k$ , but on all the other random numbers  $\omega_1 \dots \omega_N$ . We will see on **③** how we can replace this function by a function  $\Gamma'_{k,t-1}(\omega_k)$  which depends only on  $\omega_k$ . The desirable property for any coordination function  $g_{k,t}$  will become :

$$\Gamma'_{k,t-1}(\omega_k^{(1)}) < \Gamma'_{k,t-1}(\omega_k^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \leq g_{k,t}(\omega_k^{(2)}) \quad (2)$$

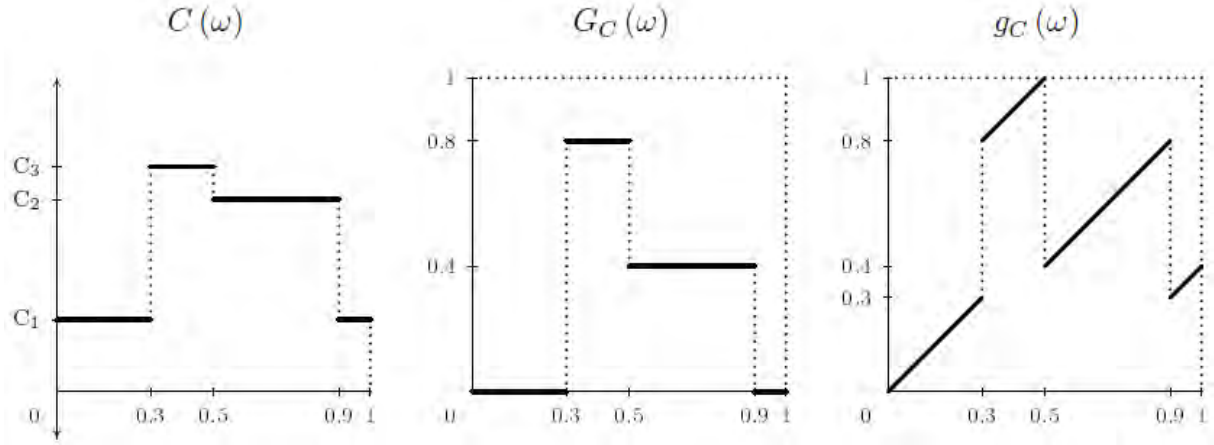
**② Construction of a coordination function:** For the sake of simplicity, we omit the subscripts  $k$  and  $t$ . So  $\omega$  is now a simple real number between 0 and 1. We note  $C$  the cumulative burden function – supposed to be a bounded measurable function:  $\omega \in [0,1] \rightarrow C(\omega) \in \mathbb{R}$  – and we wish to associate to it a coordination function  $g$  such that:

$$C(\omega^{(1)}) < C(\omega^{(2)}) \Rightarrow g(\omega^{(1)}) \leq g(\omega^{(2)}) \quad (2')$$

Let us define the function  $G_C = F_C(C)$ , with  $F_C$  the cumulative distribution function of  $C$ :

$$\forall \omega \in [0,1], G_C(\omega) = P(u | C(u) < C(\omega))$$

We can show that the range of  $G_C$  is included in  $[0,1]$ , and that  $G_C$  satisfies (2'), but is not a coordination function if  $C$  has "levels", that is subsets of  $[0,1]$  where  $C$  is constant ( $G_C$  has then the same levels). However, we can construct a bijective coordination function on  $[0,1]$   $g_C$  equal to  $G_C$  outside the levels and composed of line segments having a slope equal to 1 on the levels of  $G_C$ , as illustrated in the next figure, where  $C$  is a step function, with 4 levels:



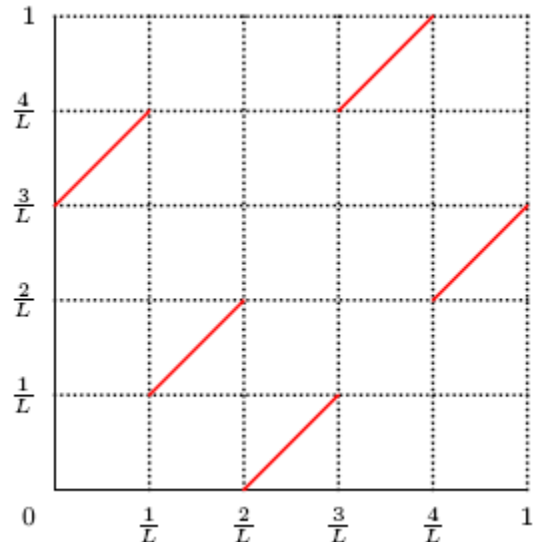
**③ Application to stratified simple random sampling:** As mentioned in **①**, we need to replace the cumulative burden function  $\Gamma_{k,t}$ , function of vector  $\Omega$ , by an approximate cumulative burden function  $\Gamma'_{k,t}$ , which should be a function of  $\omega_k$  close to  $\Gamma_{k,t}$ . This is done via a two-steps procedure:

① First, we replace  $\Pi_{k \in S_u}(\Omega)$  by its conditional expectation given  $\Omega_k = \omega$ , which is the best approximation, in the L2-norm sense, of this indicator function depending only on  $\omega_k$ . We show that this conditional expectation is equal to  $b_{k,t}(g_{k,t}(\omega))$ , where  $1 - b_{k,t}$  is the cumulative distribution function of a  $\text{beta}(N_h, N_h - n_h - 1)$  distribution in a stratum  $h$ . This leads to an expected cumulative burden function  $\Gamma_{k,t}^e = \sum_{u \leq t} \gamma_{k,u} \cdot b_{k,u}(g_{k,u}(\omega))$ ;

② Then, as the expected cumulative burden function is not a step function or a function that can be easily "computed", we divide the interval  $[0,1]$  into  $L$  – a "large enough" integer (at least greater than 50) – equal subintervals  $I_\ell \left[ \frac{\ell-1}{L}; \frac{\ell}{L} \right]$ ,  $\ell = 1 \dots L$ , and we approximate function  $b_{k,t}$  by a step function  $\beta_{k,t}$  constant on each subinterval  $I_\ell$ .

Finally, the cumulative burden function  $\Gamma_{k,t}$  is replaced by the approximated expected cumulative burden function  $\Gamma_{k,t}^{ae} = \sum_{u \leq t} \gamma_{k,u} \cdot \beta_{k,u}(g_{k,u}(\omega))$ , which is a step function, constant on each  $I_\ell$ . So we are in the same context as in the example presented in **②**: from the function  $\Gamma_{k,t}^{ae}$ , we construct a "G" function, also constant on each  $I_\ell$  and then a coordination function  $g$  which looks like in the opposite example with  $L=5$ . It is entirely defined by a permutation  $\sigma$  on  $\{1, 2, 3, \dots, L\}$ , according to the following formula:

$$\forall \omega \in \left[ \frac{\ell-1}{L}; \frac{\ell}{L} \right] \quad g_\sigma(\omega) = \frac{\sigma(\ell)-1}{L} + \left( \omega - \frac{\ell-1}{L} \right)$$



The permutation  $\sigma$  is defined by the fact that the coordination function  $g_\sigma$  has to satisfy equation (2).

### 2.3. Sample coordination between surveys based on different kind of units

The methods allows the coordination of samples relating to surveys based on different kind of units, for example legal units and local units. This “multi-level” coordination is obtained by defining a permanent link between the legal unit and one of its local units – the head office for example – and by assigning to this “principal local unit” the same permanent random number as the legal unit – the PRN of other local units being drawn according to the uniform distribution on the interval  $[0,1[$ . So, the response burden of principal local units can be taken into account in the cumulative response burden of legal units for the drawing of legal units samples, and reciprocally, the response burden of legal units can be taken into account in the cumulative response burden of principal local units for the drawing of local units samples.

## 3. RESULTS

A simulation study has been conduct to assess the properties of this coordination method. 20 legal units samples and 8 local units samples have been drawn with the multi-level procedure describe in §2, each sample being coordinated with the whole of past samples, with  $\gamma_{k,t}=1$  for all units  $k$  and all samples  $t$ . We compare the results, in terms of distribution of legal units response burden, with those, on the one hand of a sequence of 28 independent drawings, and on the other hand of the “level by level” coordinated drawing of the 20 legal units samples and independently the coordinated drawing of the 8 local units samples. The following table shows the high efficiency of the multi-level coordination procedure.

Cumulative response burden of legal units, except take-all stratum	Frequency according to the sampling scheme			Differences between drawings:		
	Independant drawings	"Level by level" coordinated drawings	Multi-level coordinated drawings	Independant versus "level by level" coordinated	"level by level" versus multi-level coordinated	Independant versus multi-level coordinated
0	4 670 676	4 651 954	4 634 250	-18 722	-17 704	-36 426
1	410 016	439 355	474 286	29 339	34 931	64 270
2	40 095	34 824	18 230	-5 271	-16 594	-21 865
3	8 072	4 679	4 125	-3 393	-554	-3 947
4	2 142	813	737	-1 329	-76	-1 405
5	578	93	92	-485	-1	-486
6	121	5	2	-116	-3	-119
7	20	0	1	-20	1	-19
8	3	0	0	-3	0	-3

## 4. CONCLUSIONS

The sampling coordination method presented in this paper proves, via many simulations studies conducted on simulated as well as real data, to be very efficient – providing significant gains in terms of response burden allocation over the population units – as well as outstandingly robust vis-à-vis sampling design parameters. It is used operationally at Insee since the end of 2013.

## REFERENCES

- [1] F. Guggemos and O. Sautory, Sampling Coordination of Business Surveys Conducted by Insee, Proceedings of the Fourth International Conference of Establishment Surveys, June 11-14, 2012, Montréal, Canada.
- [2] C. Hesse, Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+, Insee working paper E0101 (2001).

- [3] P. Ardilly, Présentation de la méthode JALES+ conçue par Christian Hesse, internal Insee working paper (2009).

# Avoiding duplicate collection of flow data: estimating intra-EU inbound tourism using partner data

Christophe Demunter ([christophe.demunter@ec.europa.eu](mailto:christophe.demunter@ec.europa.eu))<sup>1</sup>, Krista Dimitrakopoulou<sup>1</sup>

**Keywords:** *tourism statistics, inbound tourism, Regulation 692/2011, micro-data, partner data.*

## 1. INTRODUCTION

The recent revision of the legal basis for European statistics on tourism [1] includes the transmission by the Member States to Eurostat of micro-data on tourism demand. One reason for requesting micro-data was to allow better possibilities of exploiting the data and thus better addressing user needs. Another reason was that "tourism in the Union has a predominantly intra-European dimension, which means that micro-data emanating from harmonised European statistics on the demand for outbound tourism already provide a source of statistics on inbound tourism demand for the Member State of destination, without imposing additional burden, thus avoiding duplicated observation of tourism flows" (recital (7) of the Regulation).

This abstract/paper explores the possibilities to estimate inbound tourism flows into countries of the European Union (EU) and European Free Trade Association (EFTA) on the basis of partner data obtained via the tourism demand side surveys (covering domestic and outbound tourism) organised by all other countries (extracted from the micro-data stored in Eurostat's secure environment).

The underlying principle is simple: an outbound flow for one country is an inbound flow for another country. Given that more than 75% of all tourism trips made by residents of the EU had another EU Member State as the main destination [2], using partner data to estimate intra-EU inbound flows can also cover a very significant part of the total inbound tourism flows. In the current system of tourism statistics, inbound tourism is only partially covered via accommodation statistics (namely arrivals and nights spent at tourist accommodation establishments), meaning the approach discussed in this abstract/paper can significantly contribute to bridging a known information gap, in a very cost-effective (see Principle 10 of the *Code of Practice* [3]) and burden neutral way (see Principle 9 of the *Code of Practice*).

This abstract/paper looks into the data availability of this sample based source, presents a few concrete results of previously unavailable statistics on inbound tourism and makes a preliminary assessment of the coherence with other data sources.

## 2. METHODS

Section 2 of Annex I of Regulation 692/2011 lays down the requirements for the statistics on tourism trips and visitors making the trips. Each year, the Member States transmit to Eurostat a micro-data file containing a sample of observed tourism trips of at least one overnight stay (= statistical unit). For each trip, a number of study variables in the sphere of tourism (month of departure, duration, destination, means of transport, means of accommodation, expenditure, etc.) and explanatory socio-demographic

---

<sup>1</sup> European Commission, Eurostat, Unit G-3 – Short-term business statistics and tourism

breakdown variables (age, gender, etc.) are transmitted. In most countries, the data is collected via household surveys [4], asking households or individuals about trips made during a reference period of – typically – three months preceding the interview.

The Regulation aims at harmonised output, but via recommended guidelines compiled in a methodological manual [5] there is also a certain degree of input harmonisation. This opens perspectives for creating new datasets by combining the different countries. For example: inbound tourism in Austria can be estimated by combining all the sampled outbound trips in the other EU Member States where the main destination of the trip was Austria.

## 2.1. Sample size

All countries rely on samples to collect tourism demand data. Before studying in more detail the availability and reliability of observations for bilateral flows within the EU, this subsection covers a few general aspects of the sample size for this area of statistics.

For the reference year 2012 – the most recent reference year with near complete data at the time of writing this abstract/paper – a total sample of 490 000 tourism trips made by residents of the EU (no data for Sweden) and Switzerland is available. However, for the purpose of this research, only the 154 000 outbound trips are relevant (the remaining 336 000 trips being domestic trips within the country of residence of the tourist). The 114 000 outbound trips with a destination within the EU (or 74% of all observed outbound trips) will be the basis for the current research.

The share of outbound trips – in other words the popularity of travelling abroad – differs strongly across the Union (see Figure 1), this obviously will have an effect on the number of observed outbound trips in the sample.

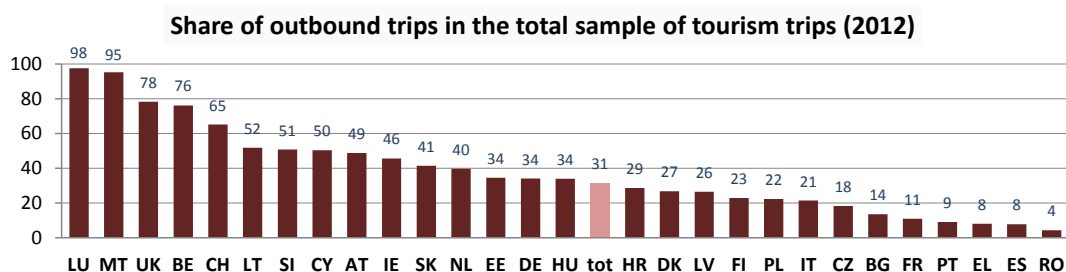


Figure 1. Share of outbound trips in the total sample of tourism trips, by reporting country.

## 2.2. Availability and reliability

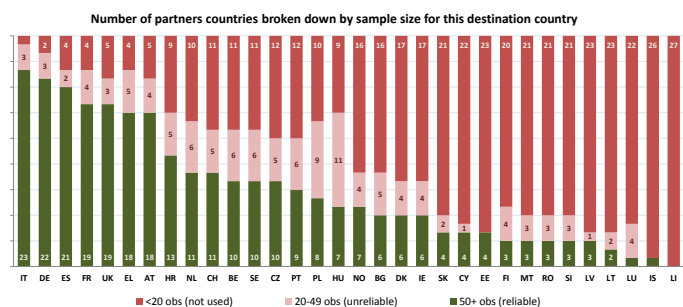
The feasibility for a given Member State to use partner data of the other 27 Member States (outbound flows) to estimate its inbound flows, will depend on the likelihood that a reliable number of observations is made. This in turn depends on two main criteria: attractiveness as a tourism destination and the sample size in the other countries that generate tourism flows to this destination. For Spain, France, Germany and Italy more than ten thousand observations of inbound trips are available, for Malta, Finland, Lithuania, Slovenia and Luxembourg, however, fewer than one thousand trips were observed.

For this area of statistics, the dissemination rules (see [5], section 3.6.8) lay down that estimates shall be flagged 'unreliable' if based on 20 to 49 sample observations and will not be published if based on fewer than 20 sample observations. All estimates based on

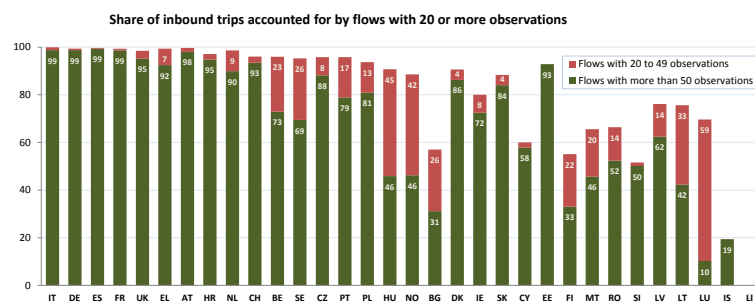
50 or more sample observations are published unflagged (and are available to users in Eurobase).

For the purpose of this abstract/paper, 784 flows were considered – i.e. 28 reporting countries (EU-28 – Sweden + Switzerland) combined with 29 destinations (28 EU Member States + EFTA as one grouped 29<sup>th</sup> entity), ignoring the 28 out-of-scope domestic flows. 94% of these flows are included in the sample while 6% was not observed. While only 36% of the flows is available without restrictions (i.e. 50 or more observations), these major flows represent 91% of all observed trips. 15% of flows are available but flagged as unreliable (5.5% of all observed trips), while 43% of flows cannot be published (3.5% of all trips).

Figure 2 shows for each destination country the number of partner countries for which the observed flows are reliable, unreliable or not for dissemination respectively. For the bigger countries, or for popular tourism destinations, a majority of the flows is available, however for smaller countries reliable inbound data is available for only few partner countries. When taking into account the weight of these flows (see Figure 3), the picture improves significantly: for most of the destination countries more than 80% of the inbound trips is accounted for by flows with 20 or more observations.



**Figure 2. Number of partner countries broken down by sample size for a given destination country (not including trips by residents of Sweden)**

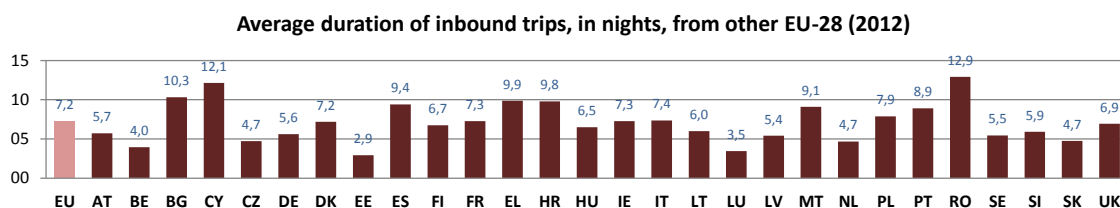


**Figure 3. Share of intra-EU inbound trips accounted for by flows with 20-49 and more than 50 observations (not including trips by residents of Sweden)**

### 3. RESULTS

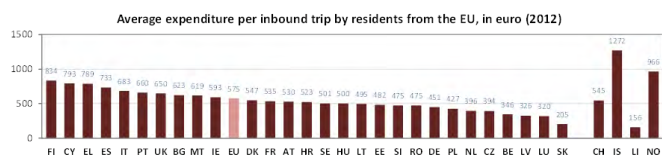
#### 3.1. Preliminary inbound data

Although publishing data on inbound tourism is not the objective of this abstract/paper, three graphs showing examples of the potential of this new, derived data are included. Figure 4 gives the average duration of inbound trips by residents of the EU. The proximity of the destination obviously plays a role. For example, the average length of trips to Belgium is 4.0 nights (see Figure 4), for British tourists to this country this is 3.5 nights while for Hungarian tourists this is 6.1 nights (not shown in the graph).

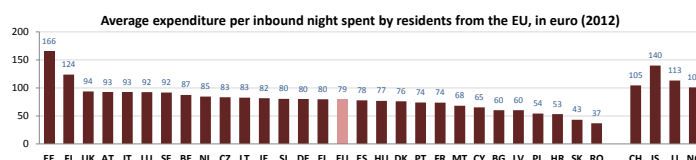


**Figure 4. Average duration of inbound trips (in nights) from other EU-28, by destination country.**

Figures 5 and 6 take a look at the average expenditure per inbound trip and per inbound night respectively. Inbound visitors coming from other EU Member States spent, in 2012, on average 575 euro per trip, ranging from 205 euro in Slovakia (a country with many short trips by tourists from neighbouring countries) to 793 euro in Cyprus (longer trips, on average 12.1 nights) and 834 euro in Finland. Per night, visitors from the EU spent on average 79 euro – ranging from 37 euro in Romania (many trips spent at non-rented accommodation) to 166 euro in Estonia (the country with the shortest average length of stay – 2.9 nights – for inbound trips, mainly made by Finnish tourists).



**Figure 5.** Average expenditure per inbound trip by EU residents, in euro.

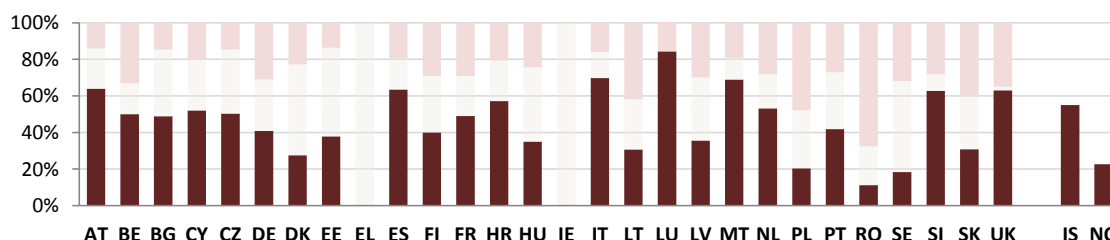


**Figure 6.** Average expenditure per inbound night spent by EU residents, in euro.

### 3.2. Coherence

Besides availability and reliability of the data, coherence with other data sources measuring a comparable or similar phenomenon are essential to assess the feasibility of the outlined partner data. It should be kept in mind that the reference data may (also) suffer from shortcomings in terms of quality.

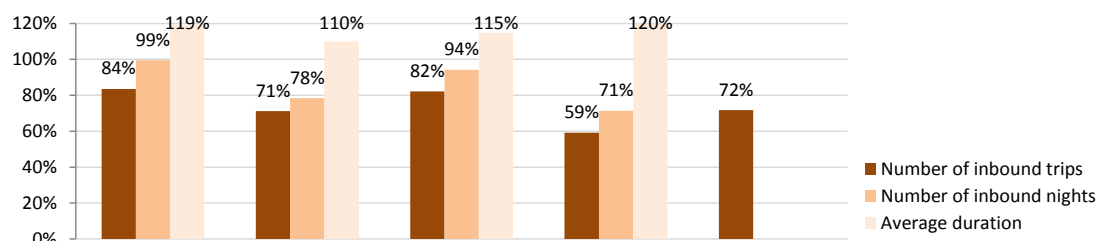
Within the data available at Eurostat, coherence was evaluated with the accommodation statistics (night spent by non-residents at accommodation establishments < estimate for total inbound nights?). Figure 7 shows that coherence is fine for all countries but that big differences exist. The gaps can be explained – among others – by the importance of non-rented accommodation (see the light-coloured upper bar showing the estimated share of non-rented accommodation in the total inbound nights), e.g. Romania, and by data collection thresholds and under-coverage in accommodation statistics, e.g. Denmark.



**Figure 7.** Ratio of nights spent at tourist accommodation establishments (i.e. rented accommodation) by non-residents from EU-28 and estimated number of nights on inbound trips by residents from EU Member States (lower, dark bar); the upper bar gives the estimated share of inbound trips spent at non-rented accommodation.

A second coherence check consists of a comparison with existing inbound tourism data. Five Member States kindly provided data coming from their border surveys (the analysis shown in Figure 8 is anonymised). The outcomes show that the data from the tourism demand surveys used in this abstract/paper is systematically lower than the border surveys. This can most likely be explained by the recall bias (respondents forgetting to report one or more trips for the requested reference period) which has been estimated in other studies ([6], [7]) to amount to an underestimation by 20 to 25%. Deviations tend to be bigger for trips than for nights (the memory effect will be less problematic for longer trips), as a consequence the average duration of trips is overestimated on the basis of the partner data from the tourism demand side surveys.





**Figure 8. Ratio of inbound demand side data and inbound border survey data (2012)**

#### 4. CONCLUSIONS

The first analysis of the tourism demand micro-data as a potential source for estimating in a cost-effective and burden neutral way the inbound tourism in EU (and EFTA) countries is very promising.

Although methodological issues such as sample size and recall bias can distort the availability of the data or the completeness and accuracy of the results (and need to be carefully quantified and where possible adjusted), this source can be a unique source of inbound tourism statistics previously not available and complementary to the existing arsenal of tourism statistics at national level.

Using partner data to obtain information on inbound tourism is a good example of the relevance of establishing common frameworks for the systematic development, production and dissemination of harmonised European statistics, not only in the area of tourism statistics but in all areas where intra-EU flows play a dominant role (e.g. trade in goods, trade in services, Balance of Payments, etc.). The abstract/paper showed the possibilities (and limitations) to enrich the socio-economic knowledge about tourism by re-using existing data collected by partner countries in the European Statistical System.

#### REFERENCES

- [1] Regulation (EU) 692/2011 of the European Parliament and of the Council concerning European statistics on tourism and repealing Directive 95/57/EC.
- [2] Eurostat, News Release 101/2014, 27 June 2014, *In 2012, 85% of trips abroad by EU residents were in Europe* and Eurostat, Statistics Explained, *Tourism trips of Europeans (multiple articles).*
- [3] European Statistical System, *European Statistics Code of Practice for the national and community statistical authorities* (adopted by the ESSC on 28 September 2011).
- [4] Eurostat, *Online reference metadata for the data collection "Annual data on trips of EU residents (tour\_dem)".*
- [5] Eurostat, *Methodological Manual for Tourism Statistics.*
- [6] Spanish Institute for Tourism Research (IET, Tourspain), *Memory Effect in the Spanish Domestic and Outbound Tourism Survey (FAMILITUR)*, paper presented at the 9th International Forum on Tourism Statistics, Paris, 19-21 November 2008.
- [7] S. Roux, J. Armoogum, J-L. Madre, F. Potier, G. Cernicchiaro (INRETS, France), *Sampling strategies and correction of measurement errors for tourism travel surveys*, paper presented at the 10<sup>th</sup> International Forum on Tourism Statistics, Lisbon, 22-23 November 2010.

# The harmonisation of mirror data using simultaneously estimated accuracies

Arie ten Cate ([arietencate@gmail.com](mailto:arietencate@gmail.com))<sup>1</sup>

**Keywords:** mirror data, harmonisation, reporting errors, international trade, foreign debt

## 1. INTRODUCTION

Mirror data are bilateral data where each quantity is reported twice. For instance, with international trade data we may have two values for each trade flow: the value reported by the exporter and the value reported by the importer. The EU does not produce consistent, harmonised statistics, based on the mirror data of the trade among its member countries. (The ESA is a “harmonised methodology”, not producing harmonised data in the above sense.)

The harmonisation of mirror data requires information about the accuracy of the various reporters. Such information can be obtained from the mirror data themselves: reporters which show on average a large mirror discrepancy might be inaccurate. Or, they might trade relatively much with inaccurate reporters.

This dilemma can be solved by the simultaneous estimation of all accuracies from all mirror data. This requires a stochastic model. Such a model needs an identifying restriction; a simple way to see this is the need to distinguish for example between all reporters reporting exactly correct and all reporters reporting exactly 10% too much.

This applies not only to international trade but also to international migration, direct foreign investment, foreign debt, etcetera. Without loss of generality I use the wording of international trade.

### 1.1. The basic principles and the literature

The following two principles are proposed:

- One identifying restriction is enough.
- All reporters are treated symmetrically. In other words: no prior information about a particular reporter is assumed.

None of the following papers satisfies both principles:

The first principle is not satisfied by [1]; see [2], appendix E, for a discussion. In [3], equation (5), an identifying restriction for export reporting is assumed and also one for import reporting (without a constant term), instead of one restriction for both.

The second principle is not satisfied in [4] and in [5], where the immigration reported by Sweden is assumed to have no bias.

Both principles are not satisfied in [6], part of the Integrated Modelling of European Migration (IMEM) project. Here is no need of an identifying restriction because the true

---

<sup>1</sup> Retired.

value of each migration flow is replaced by a linear model of migration, including an error term. Hence the harmonised data cannot be used for studying the causes of migration, because they are built into the data.

Finally, the procedure of the GTAP organisation does not use a simultaneous model at all and gives only an ordering of the countries by accuracy; see [7], also used at p.35 of [8] and discussed in appendix F of [2].

## 2. MODELS

Any research project aimed at modelling mirror data faces the choice whether or not to use the sign of the discrepancies. Using the sign is consistent with the assumption that countries differ in their bias. Ignoring the sign is consistent with the assumption that the countries differ in their error variance.

Both possibilities occur in the papers discussed above, as follows: signed are [1], [4], [5], [6]; unsigned are [3] and [7]. None of these discuss this choice.

Below, both possibilities are presented in a formal model. For any particular set of countries, both models can be estimated from the mirror data of these countries. The data might be available for several time periods. With  $T$  time periods and  $n$  countries we have  $Tn(n - 1)$  discrepancies. As we shall see below, we have  $2n$  reporting error parameters to be estimated.

When the parameters are percentages then the log is taken of the reported data. In the end, the antilog is taken of the harmonised values. See also equation (5.5) of [2].

### 2.1. The bias model

With the bias model, I assume that each country has an export bias and an import bias. The harmonised estimate of the trade flow from country  $i$  to  $j$  is the average of the two reported values corrected for the export bias of  $i$  and the import bias of  $j$ . This follows [1].

These biases can be estimated with least squares. I choose the following restriction: the expected value of the average of the reported total export and the reported total import is equal to, say,  $\gamma$  times the true total trade. In the absence of contradicting information, I choose  $\gamma = 1$ .

### 2.2. The variance model

With the variance model, I assume that each country has an export error variance and an import error variance. The harmonised estimate of a trade flow is the weighted average of the two reported values, with the weights varying inversely with the variances. This follows [9].

The estimation of these variances is much harder than the estimation of the biases above. In a way, the two models are a mirror image of each other: assuming normally distributed errors and apart from the identifying restriction, the bias model and the variance model can be written respectively as

$$\Delta_{ijt} \sim N(\beta' x_{ij}, \sigma^2) \quad \text{and} \quad \Delta_{ijt} \sim N(\mu, \beta' x_{ij})$$

where  $\Delta_{ijt}$  is the discrepancy of the trade flow from country  $i$  to  $j$  in time period  $t$ . The  $x_{ij}$  is a vector with only two nonzero elements, depending on  $i$  and  $j$ . For example in the bias model, the expectation  $\beta'x_{ij}$  is the difference between the export bias of country  $i$  and the import bias of country  $j$ . Unlike the bias model, the variance model is not a linear regression model. Given  $\mu$  (which might be a cif/fob margin), it can be seen as a Generalised Linear Model (GLM) with gamma-distributed data, introduced by [10]. The GLM allows us to estimate this with standard software. The likelihood function may have multiple local maxima.

Here, the identifying restriction is as follows: the variance of the total reported exports is, say,  $\gamma$  times the variance of the total reported imports. In the absence of contradicting information, I choose  $\gamma = 1$ .

### 3. RESULTS

Both of the above models have been applied to two small data sets concerning international trade.

#### 3.1. European trade in services

In [2] and [11] the trade in services between France, Germany, Italy and the UK is studied. In [2], the variance model is estimated both with least squares and with GLM.

As one might expect with the trade in services, being intangible, the reported values show large discrepancies. For example the trade flow from France to Germany as reported by Germany is more than three times as large as reported by France.

The bias model shows large biases for France and Italy. The variance model has three local maxima, with Italy consistently having large variances and the UK having consistently small variances.

#### 3.2. World trade in goods

In [12] the table 2 of [1] is used, with the trade in goods between several countries and blocks of countries, averaged over 1962-1987. Here the discrepancies are not as extreme as in the services trade data discussed above.

Based on the average discrepancy per country, the USA is a very accurate reporter. The simultaneous estimation of all accuracies retains this result. However, it increases the import accuracy of Japan, reducing at the same time the export accuracy of New Zealand and the import accuracy of the block consisting of non-EC Western Europe. This holds for both models of section 2.

At page 305 of [1], the possible under-reporting by the EC of imports from outside the EC is discussed. Indeed the estimated bias model shows an under-reporting with two percent; this is small compared to its standard error (less than one half) and small in the sense of being negligible in practice.

### 4. CONCLUSIONS

By making explicit the distinction between the two models of section 2 and by introducing for each of the two models a simple identifying restriction based on symmetry, a framework might be created for the systematic harmonisation of mirror data within the EU, or other international organisations. Of course, this will have to start with careful

experimentation; human monitoring of the results will be needed for some time, if not always.

This paper does not detract from the need to improve the reporting procedures. Rather, apart from being a tool for harmonisation, the estimated accuracies might also indicate where improving the reporting procedures might be most urgent.

## REFERENCES

- [1] M.E. Tsigas and T.W. Hertel and J.K. Binkley, Estimates of systematic reporting biases in trade statistics, *Economic Systems Research* 4 (1992), 297–310.
- [2] A. ten Cate, Modelling the reporting discrepancies in bilateral data, CPB Memorandum 179, [www.cpb.nl](http://www.cpb.nl) (2007).
- [3] G. Gaulier and S. Zignago, BACI: International trade database at the product-level; 1994-2007 version, [www.cepii.fr](http://www.cepii.fr) (2010).
- [4] M. Poulain and L. Dal, Estimation of flows within the intra-EU migration matrix. G  DAP-UCL, [mimosa.gedap.be](http://mimosa.gedap.be) (2008).
- [5] J. de Beer and J. Raymer and R. van der Erf and L. van Wissen, Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *European Journal of Population* 26, (2010), 459–481.
- [6] J. Raymer and J.J. Forster and P.W. Smith and J. Bijak and A. Wisniowski, Integrated Modeling of European Migration, *Journal of the American Statistical Association* 108 (2013), 801-819.
- [7] M.J. Gehlhar, Reconciling bilateral trade data for use in GTAP, GTAP Technical Paper 10, [www.gtap.org](http://www.gtap.org) (1996).
- [8] Z. Wang and M. Gehlhar and S.Yao, Reconciling Trade Statistics from China, Hong Kong and Their Major Trading Partners--A Mathematical Programming Approach, GTAP Technical Paper 27, [www.gtap.org](http://www.gtap.org) (2007).
- [9] R. Stone and D.G. Champernowne and J.E. Meade, The precision of National Income estimates, *The Review of Economic Studies* 9 (1942), 111–135.
- [10] J. Nelder and R. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society Series A* 135 (1972), 370–384.
- [11] A. ten Cate, The identification of reporting accuracies from mirror data, CPB Discussion Paper 216, [www.cpb.nl](http://www.cpb.nl) (2012).
- [12] A. ten Cate, The identification of reporting accuracies from mirror data, *Journal of Economics and Statistics* 234 (2014), 71-84.

# Aligning estimates from different surveys using Empirical Likelihood methods

Ewa Kabzinska (ejk1g12@soton.ac.uk)<sup>1</sup>, Yves G. Berger (Y.G.Berger@soton.ac.uk)<sup>1</sup>

**Keyword:** Empirical Likelihood, survey estimation, aligning estimates, auxiliary variables

## 1. INTRODUCTION

It is often the case that several surveys carried out independently in the same population measure some common variables. The population level parameters associated with these common variables are often unknown. Whether the common variable is of interest itself or is treated as an auxiliary information for estimation of other parameters, it may be beneficial to combine information gathered separately in different surveys. Combining information will usually increase precision and ensure that estimates are consistent across surveys. By consistency we mean a requirement that both samples give the same point estimate for the unknown population level parameter associated with the common variable. Typically there are also other side variables measured in the surveys, for which population level parameters, such as totals or means, are known. These variables are used to create benchmark constraints.

Aligning estimates from two surveys in presence of benchmark constraints was first addressed by Zieschang [1] in relation to the American Consumer Expenditure Survey. His method was extended later on by Renssen and Nieuwenbroek [2]. These authors propose to estimate the unknown population totals of the common variables using a pooled sample from two surveys and then include them as additional regressors in a GREG-type estimator. One of the drawbacks of this estimator is an increased probability of obtaining negative weights, especially when the number of regressors is large, which may be inconvenient from the practical point of view. Use of GREG-type estimators to combine information from different surveys was also investigated by Merkouris [3].

Wu [4] used Pseudo Empirical Likelihood methods to combine information from two independent surveys and obtained an estimator for a mean which is asymptotically equivalent to a GREG-type estimator.

Berger and de La Riva Torres [5] proposed an Empirical Likelihood based approach that may be used to estimate more complex parameters than means and totals in complex sampling designs. They obtain confidence intervals which may be calculated without relying on variance estimation or on unknown population parameters such as the design effect or the population size.

We extend the approach presented by Berger and de La Riva Torres [5] so that it can be used to combine multiple samples and to ensure that the estimates based on the common variable are equal across samples. We propose a method to obtain point estimators and confidence intervals for a wide class of parameters which are defined by estimating equations. Our approach allows to easily incorporate constraints constructed around the common variables as well as benchmark constraints. It is relatively computationally simple and does not require the intermediate step of estimating the unknown population level parameters associated with the common variables. It also produces weights that are

---

<sup>1</sup> The University of Southampton

always positive. We measure the relative bias of the proposed estimator in a series of simulations on a real dataset as well as on some purposively created data.

## 2. METHODS

### 2.1. Empirical Likelihood approach

Empirical Likelihood (EL) is a non-parametric method that uses the likelihood ratio function for inference. In this section, we briefly present how we use the EL approach to obtain point estimators and confidence intervals for population level parameters.

Suppose that two surveys are carried out independently in the same population. In each survey  $t$  the following variables are measured: a study variable  $y_t$ , an auxiliary variable  $x_t$ , for which a population level parameter is known and a common variable  $z$ , for which no population level parameters are known. Suppose that we wish to estimate some fixed, unknown population level parameters of interest,  $\theta_1^N$  and  $\theta_2^N$ , solutions to the following estimating equations:

$$\sum_{i \in U} g_{1i}(y_{1i}, \theta_1) = 0, \sum_{i \in U} g_{2i}(y_{2i}, \theta_2) = 0. \quad (1)$$

Consider the following combined empirical log-likelihood function for two samples:

$$l(m_1, m_2) = \sum_{i \in s_1} \log(m_{1i}) + \sum_{j \in s_2} \log(m_{2j}), \quad (2)$$

where  $\mathbf{m}_t = (m_{t1}, m_{t2}, \dots, m_{tn_t})^T$ . The values  $m_{ti}$  are unknown positive scale loads which need to be estimated [5].

### 2.2. Estimation of scale loads

The scale loads  $m_{ti}$  are estimated by the values which maximize (2) under a set of constraints, including benchmark and consistency constraints as well as a requirement that the estimated scale loads are positive. The constraints incorporate also the inclusion probabilities. Adding the inclusion probabilities to the system of constraints rather than putting them in the likelihood function is a key difference between our estimator and the method proposed by Wu [4]. One of the benefits of our approach is that it makes it possible to obtain Empirical Likelihood confidence regions for the point estimator, as explained in section 2.4.

### 2.3. Point estimation

The point estimators for  $\theta_1^N$  and  $\theta_2^N$  are obtained as the values which maximise the following log likelihood ratio function

$$\hat{r}(\theta_1, \theta_2) = 2(\ell(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2) - \ell(\hat{\mathbf{m}}_1^*, \hat{\mathbf{m}}_2^*, \theta_1, \theta_2)), \quad (3)$$

where  $\ell(\hat{\mathbf{m}}_1^*, \hat{\mathbf{m}}_2^*, \theta_1, \theta_2) = \sum_i \log(\hat{m}_{1i}^*(\theta_1)) + \sum_i \log(\hat{m}_{2i}^*(\theta_2))$  and  $\hat{m}_{ti}^*(\theta_t)$  are the values which maximise (2) subject to the same constraints as those imposed on  $\hat{m}_{ti}$  and two additional constraints:

$$\sum_{i \in s_1} \hat{m}_{1i} g_{1i}(y_{1i}, \theta_1) = 0, \sum_{i \in s_2} \hat{m}_{2i} g_{2i}(y_{2i}, \theta_2) = 0, \quad (4)$$

for given values of  $\hat{\theta}_1$  and  $\hat{\theta}_2$ .

## 2.4. Confidence regions

Under some regularity conditions, the log likelihood ratio function (3) follows a  $\chi^2$  distribution asymptotically under  $H_0: \theta_1 = \theta_1^N, \theta_2 = \theta_2^N$ . This property allows to construct the  $(1-\alpha)$  Wilk type confidence regions for  $\theta_1$  and  $\theta_2$  by selecting the values  $(\theta_1, \theta_2)$  which satisfy the following condition:

$$\hat{r}(\theta_1, \theta_2) \leq \chi_{df=2, \alpha}^2. \quad (5)$$

## 3. RESULTS

Finite population performance of the proposed point estimator is compared with other existing methods: the GREG estimators of Zieschang [1] and Renssen and Nieuwenbroek [2] and the pseudo EL estimator presented by Wu [4]. We design two scenarios, one relying on artificial (generated) data and one using a real dataset. In the first scenario, we generate a dataset according to a model proposed by Wu and Rao [6]. In each of the samples, there is a different variable of interest, which follows a skewed distribution. We treat the generated dataset as a population. We select two independent samples and estimate parameters of interest using the proposed EL estimator, the GREG estimators of Zieschang [1] (ZG) and Renssen and Nieuwenbroek [2] (RN) and the pseudo EL estimator presented by Wu [4] (WU). Sampling and estimation is repeated 10 000 times. Samples are selected using random systematic sampling design. The relative bias is calculated for each estimator.

In the second set of simulations we use data from the 2006 British Expenditure and Food Survey [7]. The simulation process is the same as described above, i.e., in each of the 10 000 iterations, two independent samples are selected by systematic random sampling and estimates are calculated using the four estimators. In all the simulations, the number of people living in the household and the number of rooms in the household are used as auxiliary information with known population totals. Gross weekly income is the common variable with unknown population total. The study variables differ in each simulation. In simulation 7, the total gross expenditure is estimated from both samples. In simulation 8, the total expenditure on clothing and the total expenditure on housing are estimated from the first and the second samples respectively. In simulation 9, the total expenditure on clothing and the total expenditure on food are the parameters of interest. The following table shows relative biases of the estimators considered.

Table 1. Relative biases of the proposed Empirical Likelihood estimator (EL), Wu's Pseudo Empirical Likelihood estimator [3] (WU), GREG estimators proposed by Zieschang [1] (ZG) and Renssen and Nieuwenbroek [2] (RN).

	N	$n_1$	$n_2$	$\hat{\theta}_1^{(EL)}$	$\hat{\theta}_1^{(WU)}$	$\hat{\theta}_1^{(RN)}$	$\hat{\theta}_1^{(ZG)}$	$\hat{\theta}_2^{(EL)}$	$\hat{\theta}_2^{(WU)}$	$\hat{\theta}_2^{(RN)}$	$\hat{\theta}_2^{(ZG)}$
<b>Generated data</b>											
1	100000	1000	1000	0.01%	-0.02%	0.19%	-0.16%	-0.03%	-0.06%	-0.16%	-0.17%
2	100000	200	400	0.01%	0.01%	-0.99%	-0.76%	-0.01%	-0.11%	-0.37%	-0.53%
3	100000	200	200	0.01%	0.13%	-0.76%	-0.64%	0.02%	-0.06%	-0.62%	-0.68%
4	2500	160	160	0.00%	-0.04%	-1.14%	-0.98%	-0.02%	-0.12%	-0.97%	-1.09%
5	2500	140	260	-0.01%	0.15%	-1.28%	-0.98%	0.00%	-0.13%	-0.51%	-0.72%
6	2500	240	240	0.01%	0.13%	-0.76%	-0.64%	0.02%	-0.06%	-0.62%	-0.68%
<b>Expenditure and Food Survey data</b>											
7	6645	500	500	-0.11%	0.07%	-0.57%	-0.31%	-0.05%	0.21%	-0.56%	-0.20%



8	6645	500	500	0.38%	0.44%	-0.07%	0.03%	0.06%	0.06%	-0.38%	-0.35%
9	6645	500	500	0.07%	0.07%	-0.38%	-0.30%	0.01%	0.01%	-0.36%	-0.32%

The table presented above shows that in all scenarios, the relative bias of the proposed estimator is of an acceptable size. In most cases, the proposed estimator has smaller relative bias than the alternative estimators, especially the GREG estimators. Note that when the sample size is small, the GREG estimators show relative bias close to 1%, while the relative bias of the proposed EL estimator remains lower. We conclude that the EL point estimator is asymptotically unbiased.

The main advantage of the proposed method is not in the performance of the point estimator, but in the possibility of obtaining asymmetric EL confidence regions, defined by the shape of the log likelihood ratio function (3). An example of such a confidence region is presented in Figure 1. Note that in each survey there is a different parameter of interest.

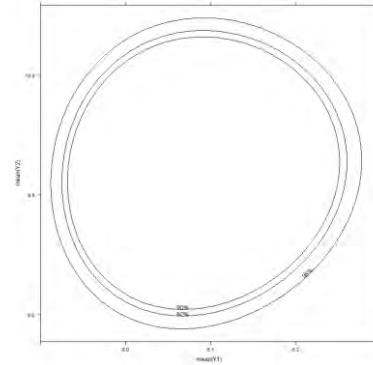


Figure 1. An example of a confidence region for two parameters  
Data generated according to a model proposed in [6]

#### 4. CONCLUSIONS

The proposed method allows to easily combine different datasets when common variables are measured in both of them and to ensure that the point estimates for the common variable are consistent across surveys. Additional benchmark constraints may also be incorporated. The method allows to obtain point estimators for a wide class of parameters which may be expressed as solutions to estimating equations, such as means, ratios or quantiles. The confidence regions are constructed using the  $\chi^2$  approximation of the log likelihood ratio function. Under the tested scenarios, the proposed point estimator shows satisfactory performance compared to the other available estimators in terms of relative bias.

#### REFERENCES

- [1] K. D. Zieschang, Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85(412), (1990), 986–1001.
- [2] R.H. Renssen and N.J. Nieuwenbroek. Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92(437), (1997), 368–374.
- [3] Takis Merkouris. Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468), (2004), 1131–1139.
- [4] Ch. Wu, Combining information from multiple surveys through the empirical likelihood method, *Canadian Journal of Statistics*, 32(1) (2004), 15–26.
- [5] Y.G. Berger and O. De La Riva Torres. Empirical likelihood confidence intervals for complex sampling designs. Southampton Statistical Sciences Research Institute, (S3RI Methodology Working Papers), (2012).

- [6] Ch. Wu and J.K. Rao, Pseudo Empirical Likelihood Ratio Confidence Intervals for Complex Surveys, *The Canadian Journal of Statistics*, 34, (2006), 359-375.
- [7] Office for National Statistics and Department for Environment, Food and Rural Affairs, Expenditure and Food Survey, 2006 [computer file]. 3rd Edition. Colchester, Essex: UK Data Archive [distributor], July 2009. SN: 5986.

# Informal Settlements in Egypt, 2011

## “The Case of Al-Duwika Zone”

**Keywords:** Slums Formation, Types of Slums, Properties of Slums Dwellers.

### 1. INTRODUCTION

Within the Egyptian context slums have been known as ‘Ashwa’iyyat’, which literally means ‘disordered’ or ‘haphazard’. It refers to informal areas and suffering from problems of accessibility, narrow streets, the absence of vacant land and open spaces, very high residential densities, insufficient infrastructure and services. (World Bank, 2008).

#### 1.1 Importance of The Study:

A set of problems arises in domains where the residents of informal areas cannot fill the government’s role and help themselves. They constraint these informal areas, growing up their location on agricultural land or in unsafe geographical areas which all led to several major problems in the quality of life for those who are living there. We can summarize the problem to be as a lack of support from the government towards its people and the failing in implementing of their rights to be as other citizens.

#### 1.2 Objectives of the Study:

When residents of informal areas compare their housing conditions to similar kinds of housing in formal areas, they feel it is unfair that the government is not taking care of them. Bad living conditions, along with the feeling of being unfairly treated lead to the frustration of many people residing in informal areas. So the study aims to understand well the magnitude of the problem by many steps as following:

- Focusing on the informal settlements in Greater Cairo Region **GCR** that takes many forms and types as:
  - Invasion on privately-owned agricultural land.
  - Squatter settlements on state-owned land Slums areas.
  - Cemeteries or Cities of the Dead.
- Determine the major demographic characteristics of the dwellers residing in these informal settlements concerning Al-Duwika zone as a case study.
- Evaluating the role of governmental programs and agencies in solving the problems of these informal settlements and suggesting some solutions.

#### 1.3 Data Sources:

This paper uses mainly data of a field survey of informal settlements obtained by **Central Agency for Public Mobilization & Statistics (CAPMAS)**, concerning slums in collaboration **Informal Settlement Development Facility (ISDF)** for the period of (2010-2011). In addition the study utilizes the data from Population and Housing Census 2006 (CAPMAS), and some estimated demographic variables obtained by (CAPMAS). The unit of analysis in this study is the demographic data available for dwellers of Manshiet-Naser section focusing on Al-Duwika zone which has been divided into zone 1 and 2 through the survey to be illustrated through spatial data maps as they represent here a case study of the survey.

### 2. Methodology of the Study

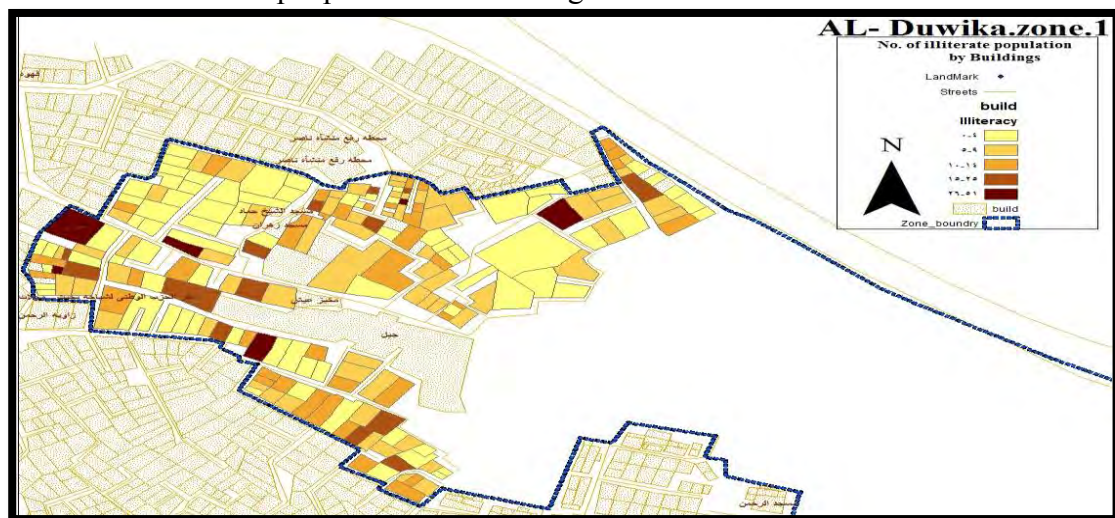
The study uses an analytical procedure of Geographic Information System or GIS data that provides a powerful medium for managing, visualizing, and communicating about our world, by using geographic location as a reference for each database record

and can be very powerful within cities for identifying spatial growth patterns, slum locations and by using analytical procedure of joining and relating data through attribute tables to show the demographic indicators through spatial data maps. This information can be combined with census and other data to determine the spatial dimensions of poverty and access within a slum zone. The modeling and visualization capability of GIS provides a means of testing alternatives and turning data into information, and subsequently into knowledge.

### 3. Findings and results of the study

**NOTE:** There are other maps represent zone.2 for every status but due to the capacity of the uploaded file I can't show them.

**Illiteracy Status:** represented below the gradation of blocks colors represents the densities of illiterate people in these buildings.



By analyzing data for both of the two zones it is shown that about 23.5% of the dwellers of Al-Duwika are completely illiterate distributed among 341 buildings out of 1548.

**Employment status:** The gradation of blocks colors represents the densities of unemployed population aged 15+ in these buildings.





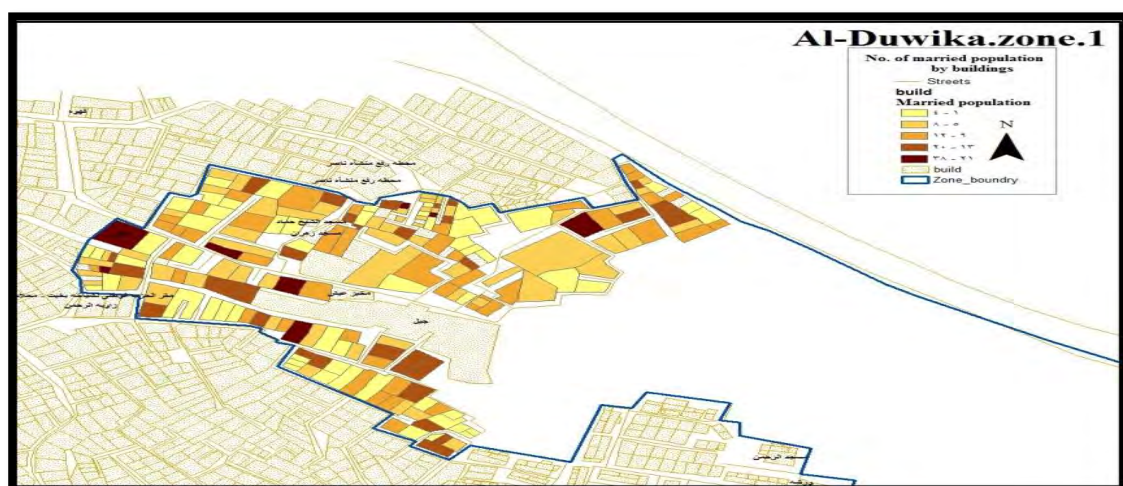
It is found that the percentage of the unemployed population aged 15+ is about 75% from the persons who expected to have work and distributed among 356 buildings out of 2935 for both zones.

**High Dependency ratio Status:** The gradation of blocks colors represents the densities of people who are expected to be in high dependency ratio in these buildings.



The percentage of males and females less than 15 and above 65 years In zone.1 about 69.5% of the dwellers are expected to be in high dependency ratio distributed among 190 out of 1387 building in this zone, by regarding zone.2 we found that about 36% of the population is are expected to in high dependency ratio distributed among 126 building out of 1548 building.

**Marital status:** The gradation of blocks colors represents the densities of no. of married population for both sexes in these buildings.



For Zone.1 no. of married population for both sexes represent about 86% from the population for this zone distributed among 197 building out of 1378 building. Concerning zone.2 the analysis indicates that about 44% from both sexes are married distributed among 134 building shown in the map.

#### **4. Conclusion**

**The analyzing of data** revealed that the informal settlements differ from any other areas with respect to their age structure, educational level, marital status, occupational composition and it's economically active population. **Illiteracy** is a major feature that informal settlements can be characterized with a high percentage of illiterate population and coupled with low employment status for those who considered being the economical active population. **There is a substantiated direction** for the relation between marriage, unemployment and illiteracy. There is a relation between unemployment status for the educated, skilled persons and marriage they are unlikely to get married because we suggest that the problem of unemployment in these areas is widespread among the relatively educated ones not for illiterate people. **The age structure** is certainly attributed to the differences in the function of fertility and mortality in these areas besides the effect of in-migration. Although in-migration is predominately single males, but they tend to marry shortly after settling in the area or bring their families from place of origin.

#### **REFERENCES:**

- **A survey done by Central Agency for Public Mobilization & Statistics (CAPMAS), concerning slums zones in collaboration Informal Settlement Development Facility (ISDF) 2010-2011.**
- **Cairo's informal areas between urban challenges and hidden potentials. GIZ. Egypt and Participatory Development Programme in Urban Areas (PDP) 2011.**
- **De Soto, H. (2000) The mystery of capital: Why Capitalism Triumphs in the West and Fails Everywhere Else. Basic Books, New York.**
- **Sims, D. (2003) The case of Cairo, Egypt. Understanding Slums: Case Studies for the Global Report on Human Settlements 2003. United Nations Human Settlements Programme (UN-Habitat).**
- **UNDP/INP (2011) Egypt's Social Contract: The Role of Civil Society. Egypt Human Development Report 2011, United Nations Development Programme, and Institute of National Planning, Egypt, New York and Cairo.**
- **UN-Habitat (2008) Global Campaign for Secure Tenure - Background. United Nations Settlement Programme, Nairobi.**
- **Viratkapan, V., & Perera, R. (2006). Slum relocation projects in Bangkok: what has contributed to their success or failure? Habitat International.**
- **World Bank - Sustainable Development Department, Middle East and North Africa Region (2008) Arab Republic of Egypt.**

# Visualisation of macroeconomic indicators in maps with R

Jan-Philipp Kolb ([Jan-Philipp.Kolb@gesis.org](mailto:Jan-Philipp.Kolb@gesis.org))<sup>1</sup>,

**Keywords:** Spatial visualisation, macroeconomic indicators, choropleth maps, R.

## 1. INTRODUCTION

The map is the most conventional way to visualize areal data. Maps help people to understand complex phenomena. The spatial perspective is very important for the visualization of macroeconomic indicators in a European context. Differences between regions are better understood with a regional visualisation.

For a long time shapefiles for regional entities have not been available free of charge. But this situation changed. The website [www.gadm.org](http://www.gadm.org) (Global Administrative Areas) offers shapefiles for many administrative levels. To visualize macroeconomic indicators the NUTS-levels (Nomenclature des unités territoriales statistiques) are particularly interesting. Shapefiles for the first three NUTS levels are available on the GADM-website.

With statistical programming language R it is possible to do the data editing, the statistical analysis and the visualisation of the results in one environment. Other Geographic Information Systems (GIS) are not necessary. With the data and R it is thus possible to produce interesting choropleth maps. R offers numerous possibilities to produce such maps. Many of these are described in [1] and [2].

## 2. METHODS

Different possibilities do exist to visualize information related to space. With R it is possible to apply many of these by using only one program. The methods available in the `maptools`-package [3] and the `sp`-package [4] are presented as well as data sources for macroeconomic indicators and shapefiles for administrative areas. Choropleth maps are used to visualize application examples. In these kind of maps every area is shaded pursuant to a discrete scale. The latest developments in spatial data analysis and visualisation are shown as well as possibilities to publish the resulting visualisations [5].

## 3. RESULTS

The target of the work is to present possibilities to draw maps in a clear and simple way. Ideally these maps should then help the observer to understand complex macroeconomic coherences better. Recommendations concerning the choice of colour are made.

## 4. CONCLUSIONS

Maps are a good possibility for visualisations. And the programming language R is a good tool to produce these spatial visualisations. An overview of the possibilities to produce choropleth maps is presented in the paper.

---

<sup>1</sup> Gesis – Leibniz Institute for the Social Sciences

## REFERENCES

- [1] M. M. Fischer, & J. Wang. Spatial Data Analysis (2011). New York: Springer. P. 16
- [2] R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio, Applied spatial data analysis with R (2008). New York: Springer.
- [3] N. J. Lewin-Koh, R. S. Bivand, E. J. Pebesma, E. Archer, A. Baddeley, H. Bibiko & R. Turner. maptools: Tools for reading and handling spatial objects (2011). *R package version 0.8-10*, URL <http://CRAN.R-project.org/package=maptools>.
- [4] E. Pebesma, R. S. Bivand, B. Rowlingson & V. Gomez-Rubio. sp: classes and methods for spatial data (2013). *R package version 1.0-15*, URL <http://CRAN.R-project.org/package=sp>.
- [5] I. Zavlavsky. A new technology for interactive online mapping with vector markup and XML (2000). *Cartographic Perspectives*, (37), 65-77.



# Towards better communication channels

Agnieszka Mróz, a.mroz@stat.gov.pl

**Keywords:** dissemination, communication, visualisation tools, programming, monitoring

## 1. INTRODUCTION

The way and quality of communicating information determine the recipients' perception of socio-economic reality and their attitude towards the official statistics. In the face of new technological advances, statistical institutions as official data producers face the challenge of improving communication channels with various groups of users. Attractive designs, intuitive functionalities, interactive applications, data visualisation are nowadays of crucial importance for statistical data recipients. Effective communication is key to mutual satisfaction of both data users and producers.

## 2. MAIN OBJECTIVE

With the aim to address this challenge, the Central Statistical Office of Poland has taken activities aiming at enhancing communication channels with users. Taking into account the growing public demand for more innovative data communication services, a special system (called STRATEG) has been created, which perfectly fits the requirements of a novel statistical product. It is a modern database application used for monitoring of development policies and reinforcement of social cohesion. Users will find here valid texts of all strategic documents (the Europe 2020 strategy in force at the EU level, and strategic documents binding on country, sub-regional and regional level) and details of the coordinating entities. Apart from extensive information resources in the database, the system enables **visualisation of indicators** in form of charts and maps, which considerably facilitates data analysis process. Visualisation modules are equipped with a range of functionalities enabling the creation of various graphic forms. Apart from static visualisation, it is also possible to view charts and maps in an animated form, with visible changes occurring in subsequent years. The function enabling export and visualisation for different formats makes it possible to use the system resources directly for publications or reports, which considerably improves the process of reporting on the implementation of strategic documents.

**DEVELOPMENT MONITORING SYSTEM**  
strateg.stat.gov.pl

Change contrast: A A A | Sitemap | About the system | Help | Contact | PL

strateg

Homepage -->>

STRATEGIES & PROGRAMMES - check progress made in the implementation of strategic goals

**Welcome to STRATEG**

STRATEG is a system created by the Central Statistical Office for programming and monitoring of development policy. It is a collection of data derived from different sources and used to monitor the implementation of strategies binding in Poland (at the national, supra-national and voivodship levels) as well as in the European Union (Europe 2020 strategy). Additionally, the system gathers statistical indicators relevant to the implementation of cohesion policy. Apart from numerical data, STRATEG also includes definitions of concepts, methodological explanations, reports and thematic analyses. Along with a highly developed database application, the STRATEG system offers its users functional tools supporting the analysis of socio-economic trends in form of charts and maps.

**MAPS AND CHARTS**

Foreign trade turnover per capita

Natural increase per 1000 inhabitants in 2013

**RECENTLY ADDED**

19.01.2015 - Updated

Please be advised that some data concerning following thematic areas have been updated: agriculture; public safety and the efficiency of the State; transport and communication; national accounts.

...see more

**DATABASE**

LOCAL DATA BANK

OTHER MONITORING SYSTEMS

**SEE ALSO**

CSO INFORMATION PORTAL

Copyrights © 2014 - Development monitoring system  
The portal is cofinanced from the European Social Fund under the Human Capital Operational Programme 2007 - 2013.

HUMAN CAPITAL NATIONAL COHESION STRATEGY

MINISTRY OF INFRASTRUCTURE AND DEVELOPMENT

STRATEGIC CENTRE FOR POLAND

ELIS

EUROPEAN UNION EUROPEAN SOCIAL FUND

Ensuring an accessible and attractive form of data presentation was one of the biggest challenges when determining the system construction principles. The first step was a review of information systems available throughout the world, including databases of international organisations, in order to identify good practices that could be applied in the developing system. The needs of future users of the system were crucial for determining its structure as well. The dialogue with users was at each stage of the project implementation. Particularly important were workshops during which the trial version of the system was presented. The meetings offered a great opportunity of showing users the structure of the system and its functionalities, which allowed for constructive discussion and identification of comments and suggestions to be considered in the project. Thus, the target users of the system had the possibility to co-create it and influence its final shape.

Owing to its rich content and transparent, attractive design, the STRATEG system is becoming incredibly popular among users. A special group of the system recipients are representatives of public administration (both at the central and regional levels). For the purpose of a successful promotion of the system and support for this user group, a series of training sessions has been organised to provide their participants with the necessary knowledge on the STRATEG system, its content and functionalities.

Besides STRATEG, Geostatistics Portal serves as a further modern channel of communicating georeferenced information using GIS technology. It is an advanced tool for interactive cartographic presentation and dissemination of data, which constitutes a comprehensive solution tailored to the European standards.

### **3. CONCLUSIONS**

While improving communication channels with users, their sustainability and visual attractiveness of services offered need to be taken into account.

A dialog with users at the early stage of a project is very productive for both sides – it allows to identify the users' needs and give them a opportunity to co-create the system. The training sessions/workshops, apart from the possibility of presenting the product to a larger group of users, also provide an opportunity to hold further discussions with them on the system functions.

# On estimation of Polish real estate market characteristics using Internet data sources

Maciej Beręsewicz ([maciej.beresewicz@ue.poznan.pl](mailto:maciej.beresewicz@ue.poznan.pl))<sup>1,2</sup>

**Keywords:** representativeness, real estate market, web scraping, Internet data sources.

## 1. INTRODUCTION

New data sources became an important issue in the Official Statistics as well as in statistics in general. This topic was raised in the context of usage of register data for deriving statistics but also as a source of auxiliary variables for small area estimation. However, assumed full coverage of population by registers/administrative records do not indicate that it reflects the current state nor fully covers the information needs of society.

Recently with rapid growth of new technologies (e.g. mobiles phones) or Internet coverage it becomes clear that new potential data source for statistics is rising [1]. Term big data appeared in the context of massive data sets that were time consuming to analyse with existing technology. Big data also describes the process how data is generated and this issue is more important for statisticians than what is the volume of the data sets.

Initial classification of big data was proposed in the UNECE project big data for Official Statistics [2] that includes human-sourced information, process-mediated data and machine-generated data (internet of things). Nonetheless there is a specific group of sources that could be classified as a Internet data sources which are defined [3] *as data collected and maintained by units external to statistical offices and administrative regulations, and are (mainly) available on the Internet (through web-based databases).*

From the methodological point there are questions that are crucial in the context of statistical data source – what is potential usefulness of use of this data? Is the data representative for general population? In what percent data source cover the general population? What statistics could be derived from the new data sources? What is the quality of data and furthermore what methods of estimation could be applied to receive sufficient statistics?

For example, Polish real estate market is only partially covered by official statistics that are published by National Bank of Poland with co-operation of Central Statistical Office. Reports are presented with delay (e.g. statistics on 2013 were published in the beginning of the October 2014) and in limited scope. On the other hand brokers and owners of properties in order to sell or rent need to publish information on the Internet to get to the potential buyers or reinters. Potential usefulness of the Internet data sources should be assessed.

The aim of the paper is to present data sources on real estate market in Poland and assesses the possibility of use of new data sources for deriving statistics on secondary real estate market. The problem of estimation of population size (secondary real estate market) will be presented and approach proposed by [4] will be applied.

---

<sup>1</sup> Department of Statistics, Poznan University of Economics

<sup>2</sup> Center for Small Area Estimation, Statistical Office in Poznan

## 2. PROPOSED APPROACH

For the purpose of the study web-scraping technique will be applied to obtain the data from the selected online portals on real estate market in Poland. Special program written in R using XML, RCurl and httr packages was developed and data for Poznań, Poland was collected. Data cleaning part included text processing and record linkage to impute and identified duplicated entries.

To estimate size of secondary real estate market in Poznań, Poland Dual System Estimator (DSE) will be discussed. Capture-Recapture estimators or DSE are widely used in the official statistics in order to estimate size of the population or evaluate the coverage of census/registers [4, 5]. In the context of many data sources that are on the Internet the problem of estimation of population size also appears. Nonetheless, there are several issues that differs the new data sources from the existing statistical sources (eg. census, registers) which one of them is coverage, in particular undercoverage and overcoverage errors. Recently [4] proposed an approach that includes in the dual system estimator coverage error with respect to registers and census data.

In order to estimate the size of secondary real estate market in Poznan approached proposed by [4] will be adopted. The approach applies log-linear model that takes into account overcoverage in the first source and assumes that second is free of coverage errors. For the purpose of the study similar situation will be assumed. Simulation study will be presented to assess the variance of proposed [4] estimator in the context of Internet data sources and secondary real estate market in Poznan, Poland. Results of the study could be further studied for real estate market research.

## 3. FINAL REMARKS

Internet data sources and big data were noticed by statisticians and official statistics in the context of statistical data source. Despite many advantages connected to low level of aggregation and timeliness the data sources should be studied in detail in order to assess the quality and potential usefulness for statistics. In the paper approach using the Internet data sources for estimation size of Polish secondary real estate market will be presented and discussed.

## REFERENCES

- [1] P. Daas, M. Roos, C. de Blois, R. Hoekstra, O. ten Bosch, Y. Ma, New data sources for statistics: experiences at Statistics Netherlands. The Hague/Herleen: Statistics Netherlands (2011)
- [2] UNECE Big data for Official Statistics (2014)  
<http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data>.
- [3] M. Beręsewicz, On representativeness and quality of Big Data for Real Estate Market in Poland, European Conference on Quality in Official Statistics (2014), [http://www.q2014.at/fileadmin/user\\_upload/BeresewiczMaciej\\_Quality.pdf](http://www.q2014.at/fileadmin/user_upload/BeresewiczMaciej_Quality.pdf)
- [4] L.-C. Zhang, On modelling register coverage errors. Journal of Official Statistics (to appear).

- [5] R.A. Griffin, Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020. *Journal of Official Statistics*, vol. 30 (2014) pp. 177-189.

# EMIR data from trade repositories as a new source of OTC CDS data

Grzegorz Skrzypczyński ([Grzegorz.Skrzypczynski@ecb.europa.eu](mailto:Grzegorz.Skrzypczynski@ecb.europa.eu))<sup>1</sup>

Linda Fache Rousová ([Linda.Fache\\_Rousova@ecb.europa.eu](mailto:Linda.Fache_Rousova@ecb.europa.eu))<sup>1</sup>

Małgorzata Osiewicz ([Malgorzata.Osiewicz@ecb.europa.eu](mailto:Malgorzata.Osiewicz@ecb.europa.eu))<sup>1</sup>

**Keywords:** European Market Infrastructure Reform (EMIR), derivatives, Credit Default Swaps (CDS), trade repositories

## 1. INTRODUCTION

According to the European Market Infrastructure Regulation (EMIR) [1], since 12th February 2014 counterparties located in the European Economic Area (EEA) that enter into a derivative contract have to report the details of the contract to one of several trade repositories (TR) recognized by the European Securities and Markets Authority (ESMA). As a result, the data are currently scattered among six trade repositories, which embraced different technical solutions for storing and representing the data. Moreover, the data are not standardised and suffer from data quality caveats. This decentralized and heterogeneous landscape poses significant challenges to the regulators in Europe accessing and analysing the data.

This paper elaborates on these challenges, putting particular emphasis on the lack of common standards and the need to match the data registered in different trade repositories, which in turn complicates any further work with the data. The main challenges are the following:

- (i) There is no single procedure for getting access to the data by the competent regulatory authorities, and the technical requirements for getting access to the data differ substantially between the TRs.
- (ii) Data are not standardized for all fields, i.e. in some cases, the reporting agents do not use common code lists for the information in the reported fields; in addition, the variables names vary from TR to TR.
- (iii) TRs are only obliged to present the reported trade data and aggregated positions. The data on the end-of-day outstanding amounts of the trades have to be calculated by the compiler for most TRs.
- (iv) Both counterparties to the transaction have the obligation to report their trades, therefore some trades may be duplicated and, as a result, the compiler have to develop a meaningful algorithm of reconciling the information between the duplicated trades, which can be scattered across the six TRs.
- (v) The reconciliation of trades relies on the use of a Unique Trade Identifier (UTI), but its definition is still under development. In the meantime, only tentative guidelines put forward by ESMA for its use exist.

---

<sup>1</sup> European Central Bank

## 2. METHODS

This paper tackles (to the extent possible) the above-mentioned caveats on the example of one specific type of derivatives contracts - the over-the-counter (OTC) credit default swaps (CDS). We focus on this derivatives type of contract, as it is expected to be the most standardised class of OTC derivatives, owing to previous voluntary reporting of these contracts to one global trade repository, the Depository Trust & Clearing Corporation (DTCC). Moreover, the interconnectedness of the CDS market may have a far-reaching impact on the stability of the financial system, which played a significant role in exacerbating the financial crisis of 2008-2009.

In the face of those developments it seems very important to collect broad and meaningful statistics on the development of the OTC CDS market and thus the EMIR initiative is very welcomed. The EMIR data are reported daily on a T+1 basis and the broad range of reported fields could allow the competent authorities to compile meaningful aggregates for the monitoring of the OTC CDS market, subject to the above-mentioned caveats of aggregating/reconciling the data. Hence, by outlining and by carrying out the necessary steps to overcome these caveats, this paper significantly contributes to the development of new statistics on the OTC CDS market.

## 3. RESULTS

Specifically, this paper shows that the majority of OTC CDS reporting under EMIR is concentrated in one of the six TRs in Europe, which in turns facilitates the aggregation/reconciliation of these contracts. Having derived the main aggregates for the OTC CDS market from the highly granular EMIR data, we compare those with the already established source of aggregate data on OTC CDS market, namely with the semi-annual survey conducted by the Bank for International Settlements (BIS) [2].

In particular, we elaborate on the conceptual differences between the two sources and explain how these seem to be reflected in the actual data. Despite some key methodological differences and different time periods (the latest available BIS data refer to end-2013), the preliminary results suggest that the OTC CDS aggregates obtained from the EMIR data are broadly in line with the aggregates available in BIS.

We also illustrate the added value of EMIR data in terms of the new breakdowns that up to now were not covered by any reporting scheme. In particular, we show how:

- (i) the geographical composition of the buyers and sellers of CDS, as well as the issuer of the underlying security may provide interesting insights into the process of credit risk flow in the euro area, and help to assess potential areas of systemic risk;
- (ii) the data on CDS contracts with non-euro-area countries (including with entities located in off-shore financial centres) and with Central Clearing Counterparties (CCPs – when the contract is cleared), may help to track the substantial changes in the market landscape in close-to-real time.

## 4. CONCLUSIONS

The new EMIR reporting supplies the competent authorities with a broad set of high-frequency data on the European derivatives market. The efficient use of this datasets



is full of challenges and will require a significant work with cleaning and aggregating the data. The first look at the data suggests, though, that processing this information is feasible (at least for some derivative classes at this stage) and that it can deliver informative outcomes, which will help to gauge the state of the derivative markets in Europe.

## **REFERENCES**

- [1] REGULATION (EU) No 648/2012 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 4 July 2012 on OTC derivatives, central counterparties and trade repositories (available [here](#))
- [2] Bank of International Settlement (BIS), May 2014, Statistical release, OTC derivatives statistics at end-December 2013 (available [here](#))

# Data Integration: an Application of a Spatially-Adjusted Regression Tree Model

Lisa Borsi ([borsi@uni-trier.de](mailto:borsi@uni-trier.de))<sup>1</sup>, Rebecca C. Steorts<sup>2</sup>, Ralf Münnich<sup>1</sup>

**Keywords:** BART, statistical matching, spatial model, MCMC

## 1. INTRODUCTION

Record linkage is the process of identifying and matching individual records from different databases that refer to the same entities. Statistical matching, on the other hand, aims at matching different databases not including the same populations. This can be done either on the micro level by producing a new micro dataset, or on the macro level where a summary statistic based on the different databases is estimated. We refer to all of such methods as data integration methods.

Our work is motivated by recent developments in German official statistics, where a new integrated household sample design will be implemented [1]. Within this setting, core variables are collected for all sampled households. In a second stage, some households are selected from the initial sample to participate in other surveys where more variables are collected. The application of data integration methods to raise estimation accuracy while simultaneously keeping response burden and costs low has become an increasingly popular area of research. Thus, through statistical matching, variables which initially were not jointly observed can now be analysed together. Consequently, under certain circumstances, these methods contribute to the reduction of response burden and costs by augmenting the scope of analysis and the number of observations needed for accurate estimates [2] [3] [4].

Since we lack access to German official statistics data, we use the AMELIA dataset which is a synthetic dataset based on the European SILC (Statistics on Income and Living Conditions) data [5]. We are interested in inference about the relationship of employment status and total disposable household income. From the AMELIA database we will draw two surveys, one similar to the SILC and the other similar to the LFS (Labour Force Survey), where income is collected in SILC but not in LFS and employment status is collected in LFS but not in SILC. Since the LFS is a much larger survey than the SILC, we want to impute household income from SILC to LFS. We use the approach of [6] to handle imputation and model estimation of employment status and household income.

## 2. METHODS

Zhang et al considered in [6] the problem of inference about the relationship of two variables reported in two different databases. Specifically, they considered inference on how some variable  $Z$  is affected by another variable  $Y$  when there exists no such database that collects  $Z$  and  $Y$  simultaneously. That is,  $Z$  is only reported in the first database, while  $Y$  is only reported in the second database. The data integration and regression problem was approached by spatially borrowing information from one database to supplement information that is missing in another database. This has been shown to be useful in medical studies (e.g. [7]) as well as census data [6]. Their approach

---

<sup>1</sup> Trier University

<sup>2</sup> Carnegie Mellon University

was a spatially-adjusted Bayesian additive regression trees (SBART) model, which imputes the missing variable in the first database based on individual-level covariates as well as geographic information. BART models an unknown function as a mixture of tree models, where each tree is a priori constrained to have a simple structure, where it only contributes to a small extent to the overall model [8]. SBART is an extension of BART since it additionally incorporates spatial random effects. Within this setting, correlation between neighboring areas is used to improve estimates. Imputation of the missing variable  $Y$  and inference about the relationship of  $Z$  and  $Y$  are obtained simultaneously using the posterior distribution. This is done by implementing a Markov Chain Monte Carlo simulation which also automatically accounts for imputation uncertainty.

## 2.1. Notation

Suppose we have  $I$  spatial units for the micro-data. In database  $D_1$  let  $m_i$  denote the number of subjects from area  $i$  and so, the sample size of  $D_1$  is  $m = \sum_i m_i$ . For the  $j$ -th person in area  $i$ , we are interested in the relationship between  $z_{ij,1}$  and  $y_{ij,1}$  where  $z_{ij,1}$  (but not  $y_{ij,1}$ ) is recorded in  $D_1$ . Let  $x_{ij,1}$  denote the vector of other individual-level covariates in  $D_1$ . In our application,  $D_1$  corresponds to the LFS data,  $z_{ij,1}$  is the employment status and  $y_{ij,1}$  is total household income which is missing in LFS. The variable that is missing in  $D_1$  is recorded on a different set of individuals in dataset  $D_2$  (corresponding to SILC in our application). Let  $y_{ij,2}$  denote the variable  $y_{ij}$  recorded in  $D_2$  rather than  $D_1$ . Also,  $x_{ij,2}$  denotes a vector of other individual-level covariates in  $D_2$  (assume that common variables to both datasets have been harmonised). Suppose  $n_i$  is the number of individuals from area  $i$ , hence,  $n = \sum_i n_i$  is the sample size of  $D_2$ . Let  $Z_1 = \{z_{ij,1}, i = 1, \dots, I, j = 1, \dots, m_i\}$ . Define  $Y_1, X_1, Y_2, X_2$  to denote the corresponding vectors of  $y_{ij,1}, x_{ij,1}, y_{ij,2}, x_{ij,2}$  respectively.

## 2.2. Sampling Model

We take the same model based approach as in [6], where we construct the model assuming sampling models for  $z_{ij,1}$  and  $y_{ij,1}$ . Assume a sampling model  $p(z_{ij,1} | y_{ij,1}, x_{ij,1}, \Phi)$  is known for  $z_{ij,1}$ . If  $z_{ij,1}$  is continuous, we can assume a linear regression model with  $z_{ij,1}$  being the dependent variable,  $y_{ij,1}$  and  $x_{ij,1}$  defining the design matrix, and  $\Phi$ , a parameter-vector including the regression coefficients and variance parameter. On the other hand, if  $y_{ij,1}$  is ordinal, a probit model can be used. We refer to [6] for examples. We provide illustrations of our own in this work. The model  $p(y_{ij,1} | x_{ij,1}, f, \theta, \sigma^2)$  describes the relationship between  $x_{ij,1}$  and  $y_{ij,1}$ . Specifically, we assume

$$y_{ij,1} | x_{ij,1}, f, \theta, \sigma^2 \sim N(f(x_{ij,1}) + \theta_i, \sigma^2). \quad (1)$$

Here,  $N$  denotes the normal distribution with mean  $f(x_{ij,1}) + \theta_i$  and variance  $\sigma^2$ .  $\theta = (\theta_1, \dots, \theta_I)^T$  is a vector of random spatial effects,  $f(x_{ij,1})$  is an unknown function associating  $Y_1$  with  $X_1$ , and  $\sigma^2$  is the residual variance. Note that we assume the same model holds for  $Y_2$  and  $X_2$  in  $D_2$ . We represent the mean function  $f(x_{ij,1})$  as a BART model. The extra random effects  $\theta$  introduce spatial correlation among neighboring areas. Thus, we refer to (1) as the spatially-adjusted Bayesian additive regression trees model.

## 2.3. The SBART Model

Recall, we wish to learn about the relationship between  $Z_1$  and  $Y_1$ , where  $Y_1$  is missing. Assume that  $(Y_1, X_1)$  and  $(Y_2, X_2)$  arise from some model  $M$ . Using the posterior from the parameters in  $M$ , we can obtain the conditional distribution of  $Y_2 | X_2$  to impute the missing values of  $Y_1 | X_1$ . Then the regression of  $Z_1$  on the imputed  $Y_1$  approximates the

relationship between  $Z_1$  and  $Y_1$ . The marginal posterior of the regression parameter  $\beta$  accounts for the variability induced by the imputation, which can be found by integrating out  $Y_1$ . For the learning process to work, we make the key assumption that  $Y_1$  and  $Y_2$  are independent samples from the same model. The described learning process is complicated by the need to specify a joint probability for  $(Z_1, Y_1, Y_2 \mid X_1, X_2)$ . For a review of BART, we refer to [8]. The prior models on the parameters are found in Section 2.2 of [6] and the Gibbs sampler is given in the Appendix of [6].

### 3. RESULTS

Our goal is to perform a thorough evaluation of the SBART and compare it to alternatives. First of all, we want to investigate the advantage of using SBART instead of BART and second, we want to compare it to a state of the art method in standard statistical matching. Furthermore, the impact of different sampling strategies, different overlap scenarios, and the consideration of different types of variables available within the matching process will be assessed. To do so, the AMELIA dataset represents the total population from which surveys will be sampled. This setting is very promising to explore the performance of new methods and to identify circumstances under which they perform well or weakly. The described evaluation is computationally very demanding and programs have still to be improved with respect to efficiency. Unfortunately, it was not possible to finish the simulations at this point in time. Nonetheless, current test runs show promising and interesting results.

### 4. CONCLUSIONS

Conclusions will be drawn as soon as results are available. The performance of SBART and BART and their favourable scenarios will be discussed and the added value of the implemented methodology will be presented. The assessment of the methodology under different circumstances serves as a basis to apply it in a real world setting. Furthermore, we want to apply it in future research to the estimation of poverty measures.

### ACKNOWLEDGEMENTS

One of the authors is supported by the Fonds National de la Recherche Luxembourg through a research grant. Furthermore, we would like to thank the German Federal Statistical Office for their support through RIFOSS (Research Innovation for Official and Survey Statistics).

### REFERENCES

- [1] T. Riede, Die Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken, in T. Riede, S. Bechtold, N. Ott, Weiterentwicklung der amtlichen Haushaltsstatistiken, SCIVERO 2013
- [2] S. Rässler, Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches, Lecture Notes in Statistics 168, Springer 2002
- [3] M. D'Orazio, M. Di Zio, M. Scanu, Statistical Matching: Theory and Practice, Wiley Series in Survey Methodology, Wiley 2006
- [4] A. Leulescu, M. Agafitei, Statistical matching: a model based approach for data integration, Eurostat Methodologies and Working Papers, 2013 Edition

- [5] A. Alfons, P. Filzmoser, B. Hülliger, J.-P. Kolb, S. Karft, R. Münnich, M. Templ, Deliverable 6.2 Synthetic Data Generation of SILC Data, Version 2011, AMELI, Advanced Methodology for European Laeken Indicators
- [6] S. Zhang, Y. Shih, P. Müller, A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets, Bayesian Analysis (2007) Vol. 2, No. 3, pp. 611-634
- [7] J. Mandelblatt, H. Andrews, J. Kerner, A. Zaubers, W. Burnett: Determinants of late stage diagnosis of breast and cervical cancer: the impact of age, race, social class, and hospital type. American Journal of Public Health (1991), Vol. 81, No. 5, pp. 646-649
- [8] H. A. Chipman, E. I. George, R. E. McCulloch, BART: Bayesian Additive Regression Trees, The Annals of Applied Statistics (2010) Vol. 4, No. 1, pp. 266-298

# Quality, analytic potential and accessibility of linked administrative, survey and publicly available data

Manfred Antoni ([Manfred.Antoni@iab.de](mailto:Manfred.Antoni@iab.de))<sup>1</sup>, Alexandra Schmucker ([Alexandra.Schmucker@iab.de](mailto:Alexandra.Schmucker@iab.de))

**Keywords:** administrative data, big data, data quality, record linkage, survey data

## 1. INTRODUCTION

Longitudinal data continuously gain in importance especially for life course studies. However, surveys increasingly face the problem of unit-nonresponse due to increasing data protection concerns and panel attrition or declining reachability and cooperation of respondents. Quality issues arise with item-nonresponse or misreporting, especially when recall error in retrospective interviews occurs. Particularly longitudinal interviews lead to high costs and response burden [1]. At the same time more and more process-produced data sources like big data or administrative data emerge and can be examined regarding their value for research [2].

In the social sciences there are usually three different kinds of potential data sources that can be linked. First, survey data on individuals, household, companies or establishments can be used. Second, administrative or register data e.g. from health, pension or unemployment insurances, routine health data or company registers are increasingly provided. Third, publicly available data can be a good supplement. Examples for this kind of data source are (commercial) business data from Bureau van Dijk, the European Business Register, Hoover's or data that are subject to Open Data Commons (ODC) or Open Database License (ODbL). In recent years data scraped from websites have emerged.

But each of these data sources has its specific advantages and disadvantages. Survey data have the outstanding advantage that they are specifically collected for certain research questions. Furthermore, we steadily improve our understanding of their data-generating process by using the total survey error framework [3]. Finally we can ask questions on behaviours and attitudes. On the other hand, surveys increasingly face problems like unit-nonresponse, item-nonresponse or panel attrition. Quality issues also arise with misreporting. Especially in retrospective interviews recall errors occur. Furthermore interviews have to deal with time restrictions. Finally all these problems lead to high costs.

Administrative data have several advantages: They usually cover long time periods and comprise precise and reliable information on the complete target population. Hence the typical problems within surveys do not occur. The main drawback of administrative data is the fact that they are primarily collected for administrative purposes. Research is only a secondary use. Furthermore, the data-generating process often is a black box for the researchers. In many cases we only know the theory and not how the data are generated in practice. Additionally, we have to take into account that changes in the data collection method and the recorded information happen without the consideration of research needs. Finally, there is often a remarkable time lag between the data collection and their provision for research purposes.

---

<sup>1</sup> Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

As publicly available data are very diverse, the described pros and cons are only applicable in some cases. The most obvious advantage is the less restrictive data access. Furthermore, the data may comprise information on the complete target population. Especially newly developed data sources provide innovative data on a timely basis. Similar to the administrative data, publicly available data may have a data structure that is not immediately suitable for empirical research. Additionally they often do not have longitudinal information. And finally even if access is not restricted to certain user groups, access may still be subject to considerable costs.

## **2. METHODS**

One potential remedy for quality and costs issues is the linkage of data from different sources. Using this approach we can balance the disadvantages of different data sources by combining their advantages. In our context linkage means combining micro data from different sources about the same observational unit. Our focus is not on the enrichment with aggregated statistics or statistical matching. Thus we can create new and more comprehensive datasets by linking survey data with administrative or publicly available data.

However, when conducting data linkages we also face challenges: first, we need to have an appropriate matching key for the data linkage. Often we have to deal with error-prone and non-unique matching variables within record linkage procedures (e.g. names, addresses, birth dates). Ridder and Moffitt describe a potential bias induced by imperfect linkage [4].

Second, we have to consider legal restrictions for data linkage and data access. This can potentially lead to lengthy process of getting permission from ethics boards or data protection commissioners.

## **3. RESULTS**

Data linkage thus potentially results in higher cost efficiency and data quality. Respondent burden can be reduced by shortening the questionnaires. Especially sensitive questions on income or questions on events from the past can be omitted. Linked data also provide higher analytic potential for substantive analyses than their separate parts, either by combining their sets of variables, by adding observational levels (e.g. employees within establishments within companies) or following respondents after their panel drop out. Moreover, research on the quality of either data source gets possible by applying validation, unit- or item-nonresponse analyses or by examining the selectivity of consent to and success of record linkage.

Our presentation will focus on the potential, quality and accessibility of linked data of the Research Data Centre of the German Federal Employment Agency. On the one hand they comprise administrative data on a daily basis with exact information on the income and receipt of benefit since 1975. On the other hand these data can be linked to several survey data sets on individuals, households or establishments which focus on fields of research like poverty, education, intergenerational mobility or lifelong learning. We describe the feasibility and results of data linkage on the basis of several examples.

## **4. CONCLUSIONS**

Data linkage allows a combination of traditional (so called designed) research data and process-produced data from various sources. These linked data may help researchers to

understand the data-generating process and to determine whether model assumptions are met. But there is still need for research. E.g. the total survey error framework has to be applied more thoroughly on process-produced data. Additionally improvements in the data access ways to linked data are necessary. As the increased richness of data also increases the risk of deanonymisation the ways of access to the single data sources may not be suitable for their combination.

## REFERENCES

- [1] Groves, R. M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly*, 75(5), 861-871.
- [2] Kreuter, F.; Peng, R. D. (2014): Extracting Information from Big Data: Issues of Measurement, Inference and Linkage. In: Lane, J.; Stodden, V.; Bender, S.; Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good. Frameworks for Engagement*, pp. 257-275. Cambridge: Cambridge University Press.
- [3] Groves, R. M.; Lyberg, L. (2010): Total Survey Error. In: *Public Opinion Quarterly* 74(5), 849-879.
- [4] Ridder, G.; Moffitt, R. (2007): The Econometrics of Data Combination. In: Heckman, J. J.; Leamer, E. E. (Eds.), *Handbook of Econometrics*, Vol. 6, Part B, pp. 5469-5547. Amsterdam: Elsevier.



## **Challenges of linking statistical data and phonetic pronunciation software.**

### **Case study: problem of regular statistics establishments' frames in Egypt.**

Nehall Ahmed Farouk Mohamed ([nehall\\_ahmed@capmas.gov.eg](mailto:nehall_ahmed@capmas.gov.eg) )

Research, computer, and sampling specialist in CAPMAS, Egypt.

### **Abstract**

**Key words:** Recent Neural Networks (RNN), International Phonetic Alphabet (IPA), Master aggregated frame, Automatic Speech Recognition (ASR).

## **1. INTRODUCTION**

Different types of statistical data are processed for various reasons to improve the statistical work and to provide new indicators. Some types of these data are measurable, comparable, and linkable but others are not. Statistical work might have a lot of challenges of mixing, comparing, and linking data, these challenges result from the nature of data type.

Many countries face the problem of linking data with each other because it's difficult to be compared. Methods, software, and techniques are implemented to solve this problem. The paper discusses the problem of linking certain case of data in Egypt and developing a methodology or technique on the basis of phonetic system for it. Also the paper discusses the nature of the Arabic language writing in Text To Speech software system (TTs). The case study problem appears in the implementation of the aggregation process for establishments' different frames in CAPMAS to create a master aggregated frame.

## 2. METHODS

2.1 First: finding out different Problems of linking statistical data that depends on the data type.

2.2 Second: understanding the basics of Phonetic pronunciation software systems and the latest achievements in Arabic language in both (TTS-ASR).

2.3 Third: The nature of Arabic language writing and its challenges for TTS software:

- a. Writing and pronunciation of Arabic are Very difficult.
- b. Arabic has a lot of problems to be implemented in TTS software.

Fourth: case study :( problem of regular statistics establishments frames' in Egypt).

Central Agency of Public Mobilization and Statistics (CAPMAS, EGYPT) conduct many different regular establishments surveys, each survey has its own frame. CAPMAS seeks to generate a main aggregated frame for all of the regular statistics establishments' frames. The total number of the related overlapped frames is 67 frames. The problem appears in the implementation of the aggregation process because there is no way to compare and link the same establishments. The paper discusses a new developing technique to link the data of the frames (establishments) and generate the aggregated frame.

The current situation of the frames is illustrated as following:

- Different frames of the establishments are overlapped and same establishment exists in different frames.
- All establishments have no unique ID number to be used in data linking.
- Disability of matching the same establishment in the related frames as it is not completely compatible in name but partially compatible because of the nature of writing in Arabic.
- Disability of matching the same establishment in different frames as it exists with different names (about 20% of the frames).

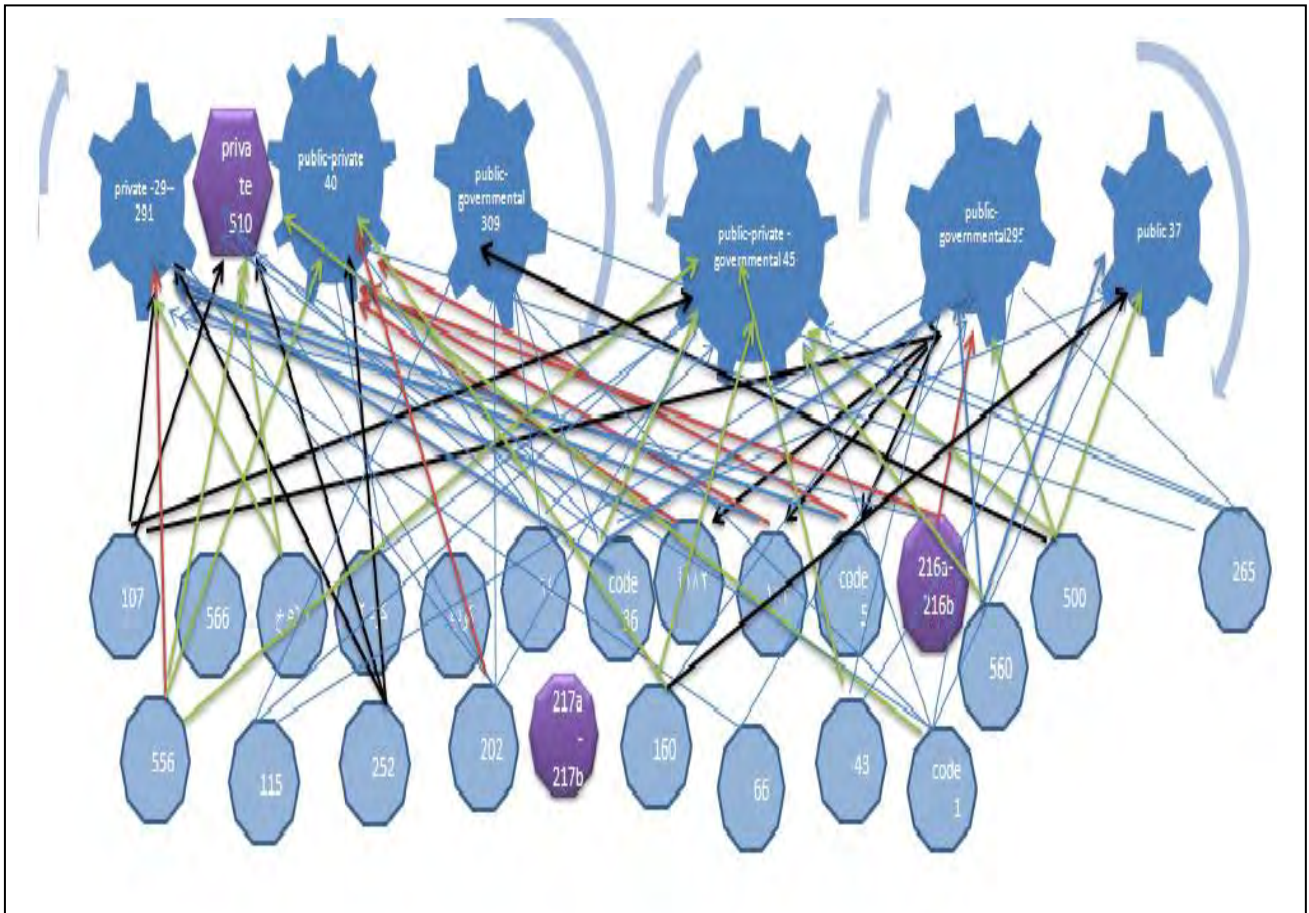


Figure 1. Relations between the frames

So frames aggregation and unification process is not accomplished due to lack of matching techniques. The idea of linking data here will depend on phonetic pronunciation software technique as a main part in the aggregation process to compare the data first and then linking it.

- The aggregation process will include main steps that are:
  1. Determining and collecting metadata about all of the overlapped related frames.
  2. Determining relationships and inter-relationships between the frames.
  3. Classifying the frames :
    - Relationship (master frames - related frames - independent frames )
    - sectoral activity(public /business sector – governmental sector – private / investment sector ).
  4. In parallel: (Creating a unique ID number- compare through the pronunciation phonetic system).
  5. Final aggregation process (matching through TTS software).

### **3. RESULTS**

The expected results of generating the master aggregated frame will have many effects in our statistical work, economic and technical systems like:

1. Data about one establishment will be collected once.
2. Reduce the fieldwork cost.
3. Helping in generating the administrative data for establishment.
4. Excluding some surveys and affects the total cost.

### **4. CONCLUSIONS**

Linking incomparable data can be achieved by the analysis of the data. The step of finding out relations between different files of data and how to compare then is the most important point to link data. So statisticians must study the nature of data and then think of how to use the most technological systems or methods to link it. Also that phonetic software are useful in comparing and linking data if the suitable software was developed.

### **REFERENCES**

- 1] Judith Rosenhouse, Arabic phonetics in the beginning of the third millennium.
- 2] Laura Mayfeild , Alan W Black , and Kevin A. Lenzo , Arabic in my hand: small –footprint synthesis of Egyptian Arabic.
- 3] yasser Hifny , Shady Quranny , Salah Hamid , Mohsen Rashwan , and other participation , ARABTALK – An implementation for Arabic Text To Speech system.
- 4] Bente Maegaard, The NEMLA R project on Arabic language sources.
- 5] Ali A. Sakr, An Arabic – English interactive system (Aeis), April 2012.
- 6] Khalid Choukri, Salah Hamid, with the cooperation of the NEMLAR partners, Specification of Arabic TTS speech corpus, May 2005.

# Estimation of Economic Performance Indicators for the Agricultural-Food Sector Through Integration of Surveys and Administrative Sources

Alfredo Cirianni<sup>1</sup>, Roberto Gismondi<sup>2</sup>, Paolo Righi<sup>3</sup>

**Keywords:** administrative data, agriculture, imputation, model based estimation, non response

## 1. INTRODUCTION

The paper deals with the methodology used to estimate a set of economic performance indicators for the Agricultural-Food sector (AF). This sector includes: 1) Agriculture, 2) Technical tools for agriculture; 3) Food products manufacturing; 4) Food products trade. In this context, the activities taken into account are: as regards the domain 2), pesticides production; fertilizers production; animal feeds production; as regards the domain 3): milk production; red meat slaughtering; poultry meat production. The AF performance indicators are new for ISTAT; they aim at measuring the productivity and the proficiency of agricultural firms.

The most important indicators are value added per employee, labour cost per employee and return on sales. The economic performance indicators are listed in the following table 1. These indicators are widely used in many economic performance analyses concerning industry and service sectors as well [1], while applications to the AF sector are scarce.

These indicators cannot be computed by using only Structural Business Statistics (SBS)<sup>4</sup>, mainly because the most part of activities carried out in the AF sector cannot be identified by univocal NACE Rev.2 codes and may be due to secondary activities, normally not covered by current business statistics. Moreover, the SBS sample surveys include only a small number of units (enterprises or *kaus*: kind of activity units) which belong to the AF domain; moreover, secondary activities are not taken into account.

Moreover, current surveys carried out in the agriculture sector – related to enterprises or agricultural holdings – do not produce economic performance indicators, while the Farm Accountancy Data Network survey (FADN) doesn't permit to estimate economic performance indicators for the AF sector, because it considers only the agricultural holdings and not the AF enterprises.

**Table 1. The list of set of economic performance indicators for the AF sector**

Indicators	ISTAT statistical source	Sub-domains concerned
1) Industrial costs/turnover	New indicators for ISTAT	-
2) Labour costs/value added	New indicators for ISTAT	-
3) Return on sales= gross operating surplus/turnover	New indicators for ISTAT	-
4) Industrial goods purchases/industrial costs	New indicators for ISTAT	-
5) Value added per employee	Structural Business Statistics (SBS)	Small-medium enterprises
6) Labour cost per employee	Structural Business Statistics (SBS)	Small-medium enterprises
7) Vertical integration indicator=value added/turnover	Farm Accountancy Data Network survey (FADN)	Only agricultural holdings, not used in this context
8) Industrial cost competitiveness = (value added per employee)/(labour cost per employee)	Farm accountancy data network survey (FADN)	Only agricultural holdings, not used in this context

<sup>1</sup> Researcher ISTAT.

<sup>2</sup> Head of Agriculture Statistics (ISTAT).

<sup>3</sup> Senior researcher ISTAT

<sup>4</sup> In Italy, the main SBS surveys are: a) Small-medium enterprises survey (the sample concerns firms with less than 100 persons employed); b) Census structural survey, concerning all enterprises with 100 or more persons employed.

For these reasons, estimation of economic performance indicators regarding the AF sector has been founded on integration of administrative data with existing surveys carried out by ISTAT. Use of administrative sources (fiscal and balance sheet data) led to the availability of an integrated database, which was the basis for estimating the indicators without the need of additional resources. Three basic issues to be tackled were the following ones: 1) identification of the reference population; 2) integration among sources through record linkage; 3) treatment of partial non responses, for units for which no fiscal or survey data were available after the integration process.

## 2. REFERENCE POPULATION AND DATA SOURCES INTEGRATION

A database was created for the years 2008-2011. The reference population concerning the AF sector was derived from the lists actually used in the frame of ISTAT agriculture statistics<sup>5</sup>. Overall, in 2011 the population included 2,205 firms for milk, 1,539 for red meat, 168 for poultry meat, 87 for pesticides, 704 for fertilizers and 729 for animal feeds. As regards the milk sector, about 500 enterprises derived from the Business Register have been added, since this sector is the only one in the AF sector identified by a specific NACE economic activity class (10.51). The sources were:

- ISTAT Business Register: it measures the population structure and provides economic variables (employment and turnover) needed to calculate economic performance indicators like labour cost per employee and value added per employee. The business register is updated every year.
- Balance sheet "IAS": updated by firms listed in stock exchange market on a voluntary basis.
- Balance sheet: it is updated by capital societies.
- Fiscal sector studies: they are filled by firms with less than 7.5 million of turnover, with the exception of firms with less than 30.000 euro of turnover.
- Small and Medium Enterprises survey: it is carried out each year by ISTAT on a sample of 11 firms for pesticides, 98 firms for milk sector, 85 firms for fertilizers, 12 firms for poultry meat, 71 firms for red meat in 2011 (overall, 277 firms).

The key source is given by balance sheets [1], because it is the source using the same definition of economic variables as requested by the SBS Regulation. The main gap is that only the capital societies fill the balance sheet data by law. The second best source is given by the sample used in the Small-medium enterprises survey by ISTAT, which is a part of SBS. This source adopts the same definitions requested by the SBS Regulation, but only a few firms are investigated. The third option is given by fiscal sector studies, for which definitions of economic variables are not always similar to those requested by the SBS Regulation. The census structural survey is not taken into account, because it is based on balance sheet data, already available from administrative sources.

The linking key was based on the fiscal code of enterprises. The fiscal code is unique and without errors; when the fiscal code was not available, the record linkage was done using the main identification variables (name and address).

## 3. THE IMPUTATION PROCEDURE

The estimation process follows a typical model based or predictive approach [2]. We assume to have a sample of units chosen not randomly. Then, we define a working super-population model fitted with the sampled units to predict the interest variables of the not sampled units that are, in our application, the not linked units. In this context the estimation procedure coincides with a method of imputation and the imputation is performed on the variables defining the eight indicators (table 1). Let  $y$  be the variable of interest and  $\mathbf{x} = (x_1, \dots, x_q, \dots, x_Q)'$  a vector of auxiliary variables. A general formulation of the working model be  $E(y_k | \mathbf{x}_k) = f(\mathbf{x}_k, \boldsymbol{\beta}) = \tilde{y}_k$  for unit  $k$  of population  $U$ , being  $\varepsilon_k$  the residual term, with  $Var(\varepsilon_k | \mathbf{x}_k) = \sigma^2 a(\tilde{y}_k)$  for some known function  $a(\tilde{y}_k)$  and that

<sup>5</sup> The agriculture production surveys concern: milk, slaughtering, fertilizers, pesticides, animal feeds.

$Cov(\varepsilon_k, \varepsilon_j | \mathbf{x}_k, \mathbf{x}_j) = 0$  for  $k \neq j$ . The parameters of interest are the unknown totals defining the indicators. The predictive estimator for the domain of interest  $U_d \supset U$  is  $\hat{Y}_d = \sum_{k \in s_d} y_k + \sum_{k \in \bar{s}_d} \tilde{y}_k$ , where  $s_d$  and  $\bar{s}_d$  are respectively the set of sampled and not sampled units in  $U_d$ . The basic hypothesis for obtaining accurate estimates is that, conditionally to the auxiliary variables known for all the units in  $U_d$ , the models generating the interest variables are identical for units in  $s_d$  and  $\bar{s}_d$ . The critical issue of the predictive approach is how the hypothesis tailors to the reality. Generally, model based estimates are biased since the identity condition of working and true model fails. Nevertheless, using model diagnostics we can have a degree of confidence on how the working model is far from the true super-population model and how much the unknown bias could be. Another important condition for the accuracy is the coverage of  $s_d$  in  $U_d$ . For instance, in the Milk production sector the coverage is of about 65-89% in terms of number of enterprises (Table 2) and more than 88% in term of turnover. In this case, high coverage preserves from model failure. We mainly analyzed two class of working model to predict the variables of interests for obtaining the estimates of totals at time  $t$ . The first is a general heteroschedastic regression model using all the auxiliary significant variables such as: milk production, turnover, employees, referred to time  $t-1$ . Because of different patterns of the auxiliary variables we fitted different models each one related to a given pattern. The prediction based on the general regression model is called *regression method*. The second class focuses on the ratio model  $y_k^t = \beta^* y_k^{t-1} + \varepsilon_k \sqrt{y_k^{t-1}}$  where the weighted least square estimate of  $\beta$  is given by  $\hat{\beta} = \sum_s y_k^t / \sum_s y_k^{t-1}$  being  $y_k^{t-1}$  either  $y_k^{t-1}$  or  $\tilde{y}_k^{t-1}$ . In this case the estimator can assume a complex expression since if  $\tilde{y}_k^{t-1}$  is used for some  $k$  the model incorporates the time  $t-2$  value and so backward. The estimation procedure is easy to implement but the complexity of the underling model increases the complexity of the variance estimates. We denote the prediction based on the ratio regression model as *robust ratio method*. Finally to deal with the outliers for the regression parameter estimations we used the least trimmed square method [3]. Then we have the *robust regression method* and the *robust ratio method*.

#### 4. MAIN RESULTS

The main results are presented in table 2 and table 3 (values are expressed in euro). The first table provides the estimates of value added per employee for milk production sector as derived from different imputation methods. The table 2 shows that different imputation methods give quite similar results, which however are very different from the ones obtained from raw data or only linked data. The table 2 includes also the ratio method, which is based on the same model of the robust ratio method, but the parameter estimation is carried out by the standard weighted least square technique. The performance of this imputation method seemed worse than the robust version with a wider range of year-to-year changes. The imputation technique used for producing final estimates was the robust ratio method, because it is the best method for outlier treatment.

**Table 2. Estimation of value added per employee for Milk production – Years 2008-2011**

Imputation techniques % share of linked enterprises	Years				% changes		
	2008	2009	2010	2011	2009	2010	2011
<b>Imputation techniques</b>							
robust regression method	61,398	59,504	66,651	67,568	-3.1	12	1.4
ratio method	60,730	60,319	71,147	65,278	-0.7	18	-8.2
robust ratio method	58,421	61,626	68,208	65,587	5.5	10.7	-3.8
raw data or only linked data	53,283	52,491	42,881	62,740	-1.5	-18.3	46.3
<b>% share of linked enterprises</b>	65.5	65.9	63.4	89.9	-	-	-



The table 3 shows the comparison between estimates (referred to 2,205 units and obtained with the robust ratio method) and SBS, calculated on a sample of 98 firms for small and medium enterprises and of 55 firms for the census survey concerning enterprises with 100 or more employees. The table 3 shows how, for both value added per employee and labour cost per employee, the largest difference respect to SBS is for 2010. The main reason may be due to the calibration approach used in SBS, which links sample estimates to the not fully business register structure.

**Table 3. Comparison between estimates and SBS for Milk production – Years 2008-2011**

Year	Value added per employee				Labour cost per employee		
	Estimate	SBS	% difference		Estimate	SBS	% difference
2008	58,421	53,620	9.0		36,348	38,054	-4.5
2009	61,626	67,487	-8.7		38,468	38,368	0.3
2010	68,208	59,813	14.0		44,992	39,588	13.7
2011	65,587	63,130	3.9		39,790	38,609	3.1

We briefly give some comments on the results on Milk production sector comparing the performances of prediction estimator based on the general regression model and ratio model. Decide what the best estimator was it is a critical issue. We considered as reasonable indicators the goodness of fit and the degree of trend smoothness of the two estimation time series. Without going in detail the two model had similar goodness of fit (about 90%). Furthermore, the estimators produced comparable trend. Differences were visible especially for (Labour costs)/(Value added) indicator, where the robust regression method showed a better trend, while for (Industrial goods purchases)/(Industrial costs), Vertical integration and Industrial cost competitiveness the estimates based on the robust ratio method seemed to better fulfill the smoothness condition. Finally we tried to compute the variability of the estimates (year 2011) when using the ratio model.

The estimation procedure hides a complex model due to the use of imputed values, so we performed a naïve variance estimator based on the jackknife technique. We estimated the indicators leaving one unit at time and computed the jackknife variance based on the variability of such estimated indicators. In the case of Milk production sector the results seemed quite good. Although we must consider these measures of variability as pseudo variance, the table 4 shows that only the domain denoted as secondary we obtained large pseudo coefficient of variations.

**Table 4. Pseudo coefficients of variation of Milk production estimates based on the robust ratio model - Year 2010**

NACE Rev.2 domains	Sample size	Performance indicators ( <i>description in table 1</i> )							
		1	2	3	4	5	6	7	8
Overall	1,941	0.04	0.26	1.46	0.06	0.52	0.48	0.27	0.24
10511	46	0.37	2.73	9.14	0.37	2.72	1.81	2.59	2.46
10512	1,209	0.03	0.35	1.25	0.03	0.47	0.39	0.34	0.32
Secondary	686	0.97	1.65	-13.32	1.80	19.28	21.40	5.15	1.61

## REFERENCES

- [1] ISTAT (2014), *Rapporto sulla competitività dei settori produttivi*, Roma: ISTAT.
- [2] Dezzani F., Pisoni P., Puddu L. (1996), *Il bilancio*, Milano: Giuffrè.
- [3] Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.
- [4] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American statistical association*, 79, 871-880.



# Estimation of the at-risk-of-poverty rate from interval data

Simon Lenau (s4silena@uni-trier.de)<sup>1</sup> and Ralf Münnich<sup>1</sup>

**Keywords:** At-risk-of-poverty rate, variance estimation

## 1. INTRODUCTION

Combating poverty is since many years one of the goals of the European Commission. In order to adequately measure poverty and social exclusion, the Statistics and Income and Living Conditions (SILC) was invented. One very popular poverty measure is the At-risk-of-poverty rate (ARPR) which was widely investigated within different research projects, such as NET-SILC (cf. [1]), AMELI [2], or SAMPLE [3].

In several countries, however, the ARPR is also estimated from other surveys, which do not necessarily have a continuous income variable available, but only income classes. When estimating the At-risk-of-poverty rate (ARPR) from categorized data, calculating quantiles or shares requires some form of assumptions on the distribution within the classes. For instance, German statistical offices use the interval-censored income measured in the German micro-census to provide regional ARPRs. This is done by applying linear interpolation techniques [4], assuming uniform distributed income in several categories. This works quite well for point estimators, but may lack in precision when deriving accuracy measures.

Although the ARPR is a non-linear estimator, which prevents closed form variance formula, approximate solutions using linearization techniques exist [5]. However, applying linearized variance formulae assumes continuous incomes and does not account for handling interval-censoring. Given the rising importance of accuracy measures [6], evaluations of de-categorizing approaches in terms of variance estimation are important.

## 2. METHODS

### 2.1. De-Categorization: Linear interpolation

In this setting, the critical income-classes are those containing the median-value and the ARPT-value. For respondents falling into these categories, a first approach used by German statistical offices is to equally distribute the income values between the class boundaries, which is linear interpolation of the cumulative distribution function [4].

### 2.2. De-Categorization: Further Approaches

Alternatively to classical de-categorization methods, non-parametric approximations can be used. Especially spline methods [7] seem convincing since the ARPR requires approximations for the poverty threshold and its 60% share which lay in

---

<sup>1</sup> University of Trier

interior classes rather than in the most upper class where approximation may become more sophisticated.

### 2.3. Linearization

The linearization of the ARPR, using a poverty threshold (ARPT) of 60% of the median [8], is given by

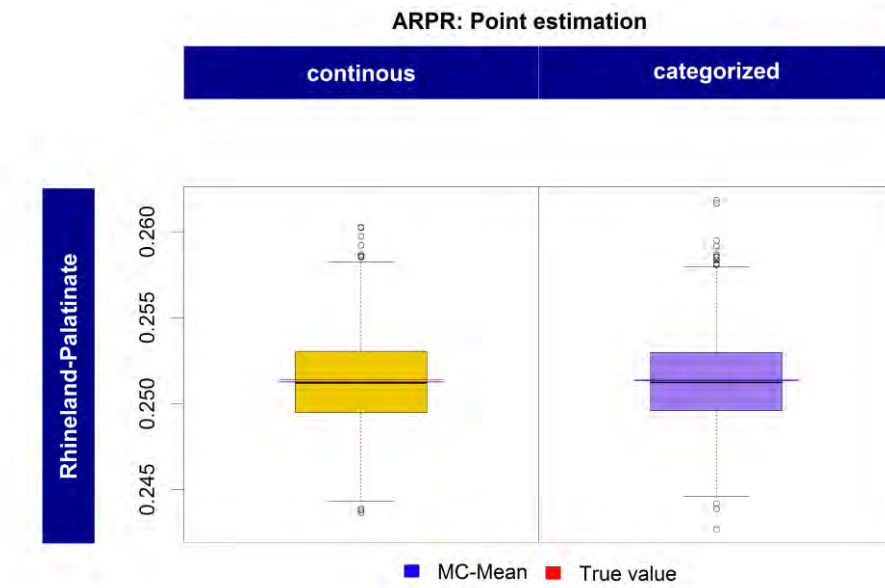
$$z_k = \frac{1}{N} \cdot \left( \mathbb{I}(y_k \leq ARPT) - ARPR \right) - \frac{0.6 \cdot F'(ARPT)}{N \cdot F'(F^{-1}(0.5))} \cdot \left( \mathbb{I}(y_k \leq F^{-1}(0.5)) - \frac{1}{2} \right) . \quad (1)$$

The variance of the ARPR can then be approximated by the variance of the estimated population total of  $z$ :

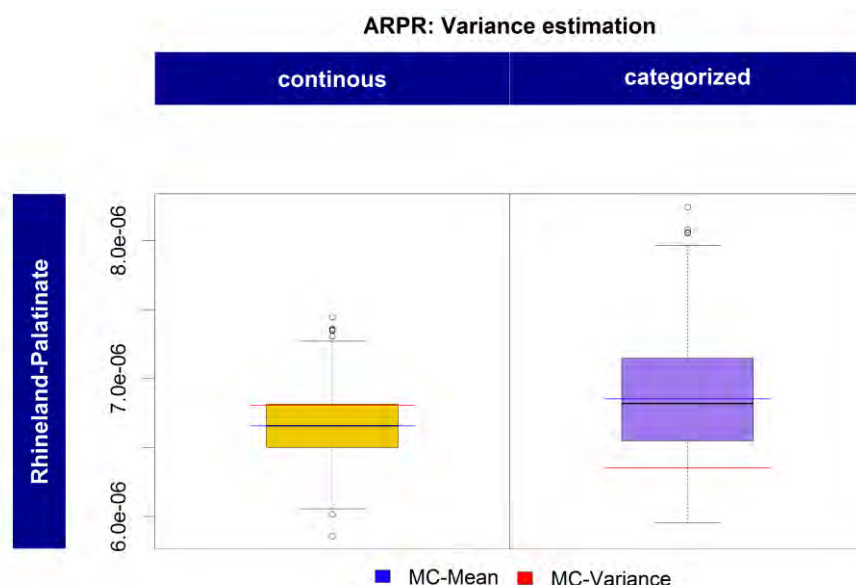
$$\hat{V}(\widehat{ARPR}) = \hat{V}(\hat{\tau}(z)) \quad . \quad (2)$$

## 3. RESULTS

*Preliminary* Monte-Carlo simulation results were computed using a synthetical income variable on census-data of the German federal state Rhineland-Palatinate. The sampling design is related to the micro-census design which is the largest German household sample drawing 1% of the population.



**Figure 1. Point estimation of the at-risk-of-poverty rate**



**Figure 2. Variance estimation of the at-risk-of-poverty rate**

While the point estimates (Fig. 1) do not reveal any clear bias due to interval-censoring, the variance estimation (Fig. 2) is visibly more biased when using categorized data, even if there is some small bias for continuous data, which is because of the approximate nature of the linearization [9].

#### 4. CONCLUSIONS

Even if linear interpolation allows for approximately unbiased point estimates, it induces a visible bias to variance estimation when used in linearization. Further approaches to deal with interval-censored income data as indicated in 2.2 may yield better results if not mere point estimates but also standard errors are to be reported.

## REFERENCES

- [1] <http://www.cros-portal.eu/content/second-network-analysis-eu-silc>
- [2] <http://ameli.surveystatistics.net/>
- [3] <http://www.sample-project.eu/>
- [4] Information und Technik Nordrhein-Westfalen (2009): Berechnung von Armutsgefährdungsquoten auf Basis des Mikrozensus, [http://www.amtliche-sozialberichterstattung.de/pdf/Berechnung%20von%20Armutsgefahrdungsquoten\\_090518.pdf](http://www.amtliche-sozialberichterstattung.de/pdf/Berechnung%20von%20Armutsgefahrdungsquoten_090518.pdf).
- [5] G. Osier (2009): Variance estimation for complex indicators of poverty and inequality using linearization techniques. Survey Research Methods, 3, pp. 167-195.
- [6] Eurostat (2012): European statistics code of practice. For the national and community statistical authorities.
- [7] J. Dai, I. Moral-Arce and S. Sperlich (2012): Calibrated estimation of a nonparametric income distribution from a few percentiles, <http://www.2013.isiproceedings.org/Files/CPS102-P14-S.pdf>
- [8] I. Dennis and A.-C. Guio (2003): Armut und soziale Ausgrenzung in der EU nach Laeken, Teil 1. Statistik kurz gefasst, pp. 1-8
- [9] C. Bruch, R. Münnich and S. Zins (2011): Variance Estimation for Complex Surveys. Advanced Methodology for European Laeken Indicators (AMELI), 3.1.

# Optimum allocation of variables in a modular survey architecture

Evangelos Ioannidis (eioannid@aueb.gr)<sup>1</sup>, Fernando Reis<sup>2</sup>, Cristina Calizzani<sup>2</sup>, Fabrice Gras<sup>2</sup>, Martin Karlberg<sup>2</sup>, Takis Merkouris<sup>1</sup>, Michalis Petrakos<sup>3</sup>, Photis Stavropoulos<sup>3</sup>, Li-Chun Zhang<sup>4</sup>

**Keywords:** Modularisation, integration, pooling, social survey, simulated annealing.

## 1. INTRODUCTION

### 1.1. A European toolbox for a modular design and pooled analysis of social surveys

The current context of the social surveys which are run by official statistics organisations is characterised by an increasing demand for the coverage of new topics while the budgets available are decreasing. While some new topics translate into new variables to be added to existing surveys or to brand new surveys, other are transversal, such as economic well-being, and require the analysis of the joint distributions of variables from separate surveys.

In order to deal with these challenges European Statistical System (ESS) agreed in 2011 on a so-called Wiesbaden Memorandum [1] which calls for the ESS to strengthen the capacity for reaction and adaptation. In the context of this memorandum, Eurostat launched a project on the streamlining and integration of the European social surveys. This project included the development of a methodological toolbox to deal with estimation and sampling issues in the context of an integrated system of surveys [2].

### 1.2. Optimum allocation of variables

Part of the purpose of an overall system of surveys is to deal with the decision of where to include the new variables that over time social surveys are requested to collect. The traditional approach is to include them in a survey which would be thematically suitable. However, some new topics, e.g. victimisation, are not suitable to be included in any existing survey, and moreover, the existing survey questionnaires are frequently already too large to accommodate additional variables. Creating completely new survey is often not an option as statistical production systems have reached their capacity limit. Other topics are transversal and in order to answer to the different users' needs, they should be included in surveys in different statistical domains. Finally, some needs do not involve the introduction of new variables in the surveys but the estimation of joint distributions of variables currently included in separate surveys.

Therefore, a more flexible approach to the allocation of the variables is needed. This allocation needs to comply with the outputs (estimations) which need to be produced and with their corresponding specifications, namely the precision required, the mandatory crossings (for estimations on joint distributions) and periodicity.

The toolbox developed by the previously mentioned project includes an algorithm to allocate variables as the solution to an optimisation problem, specified in the framework of a modular survey architecture.

---

<sup>1</sup> Athens University of Economics and Business

<sup>2</sup> European Commission (Eurostat)

<sup>3</sup> Agilis SA

<sup>4</sup> University of Southampton and Statistics Norway

## 2. THE FRAMEWORK: MODULAR SURVEY ARCHITECTURE

The basic building blocks of a modular survey architecture are the modules [3]. The modules are simply groups of variables which should be always kept together for analytical or data collection reasons. Each variable can be present in only one module.

Three types of modules can be distinguished: core modules, harmonised modules and specific modules. Core modules are groups of variables (most notably demographic variables) that are included in all data collections. Harmonised modules are groups of variables (such as education participation or health status) which are used in the surveys of different statistical domains, even if not in all surveys. Specific modules are used in only one survey. A large part of the variables of the current social surveys would fall into this latter category.

Modules are used to compose *instruments*, which then consist of sets of modules for which data is going to be obtained for the same individuals. Therefore, instruments allow the observation of joint distributions of variables present in different modules. A module can be present in more than one instrument.

For each instrument one sample is selected. The samples can be independent or coordinated. Data are then compiled for each individual for each module of the corresponding instrument. The compilation can be done via a tailor-made data collection or via the re-use of existing data, such as administrative records. In the existing systems of social surveys, instruments would often correspond to the surveys.

The estimation of population parameters will then be based on the pooling of the required input data from all the samples for which it has been collected. The pooling can be done with the methods proposed in [4] or in [5].

## 3. CODIFYING THE MODULAR ARCHITECTURE

### 3.1. Instrument composition

The composition of each instrument in terms of modules is codified via a composition matrix  $A$ . To accommodate longitudinal data collection, Instruments are grouped into *parent instruments* and *child instruments*. Each child instrument corresponds to each of the times the instrument is used to compile micro-data. For example, if an instrument has a quarterly periodicity and the programming period corresponds to one year, there would be 4 child instruments.

The composition matrix  $A$  has  $k$  columns, one for each child instrument  $I_{i,j}$  of every parent instrument  $I_i$ , and  $m$  rows, one for each module. Element  $a_{i,j}$  of matrix  $A$  is 1 if module  $i$  is included in instrument  $j$ , otherwise it is 0.

### 3.2. Output requirements

#### Precision requirements for each module

Precision requirements are specified as minimum sample sizes required for each module ( $n^*$ ).

$$A \cdot n \geq n^*$$

These are computed *a priori* based on the precision needed for the survey estimates of the various variables. The sample size of the instruments ( $n$ ) will be determined by the optimisation algorithm.

#### Mandatory crossings

Crossings between modules consist of the requirement that their data have to be compiled for the same individuals.

### *Admissibility conditions*

Crossings require that elements  $a_{i,j}$  of matrix  $A$  are jointly set to 1 for all modules of the crossing in at least one instrument. Therefore, admissibility conditions are set for matrix  $A$ . Only matrices  $A$  for which these conditions are observed are admissible as a solution to the allocation problem.

### *Additional rows in matrix $A$*

In order to allow estimations about the joint distributions of variables in different modules a minimum sample size for each crossing is specified as additional rows in matrix  $A$  and vector  $n^*$ .

### Periodicity

A periodicity is specified for each module. The period of a crossing is determined by the lowest of the periods of the modules participating in the crossing.

### *Admissibility conditions*

In order to guarantee that modules are collected with the required periodicity, only matrices  $A$  which comply with the following conditions are admissible:

- 1) Each module participates in at least one instrument with a frequency which is equal or higher than the frequency required for the module;
- 2) Modules are included in the children instruments of a particular parent instrument with a time interval equal to the periodicity of the module;

### *Additional rows in matrix $A$*

In order to allow modules to participate in crossings for which a longer period is desired than the one specified for the module, a low-frequency copy of the module is created and added as additional row of the matrix  $A$ . This copy of the module may then be added to the crossing.

### **3.3. Other constraints**

There are restrictions associated to the system of surveys which will constitute additional constraints in the admissibility of composition matrix  $A$ . The main examples are:

- 1) Core modules: modules which have to be included in all instruments;
- 2) Modules mandatory in some instruments;
- 3) Dependencies between modules resulting from routing;
- 4) Maximum size of instruments;

## **4. TWO STEP OPTIMISATION ACROSS INSTRUMENT COMPOSITIONS AND SAMPLE SIZES**

The allocation of modules to the appropriate instruments will be determined as the solution to an optimisation problem.

### **4.1. The cost function**

The total cost  $C$  of the system of surveys is the sum of the cost of compilation of the data for the  $k$  instruments. The cost for each instrument is the sum of a specific fixed cost  $A_j$  and a variable cost proportional by a factor  $\alpha_j$  to the sample size of the instrument  $n_j$ .

$$C = \sum_{j=1}^k (A_j + \alpha_j n_j)$$

## 4.2. The optimisation algorithm

In addition to the sample size of the instruments ( $n$ ), the total cost depends on the composition matrix  $A$ , in particular on how many and which instruments are used to allocate the modules.

$$C = C(A, n)$$

For any given instrument composition matrix  $A$ , the sample size of each instrument is determined by minimising the total cost. As the cost function is linear on the sample sizes  $n$  and the constraints are linear in  $n$ , it can be optimised via the **simplex algorithm**.

$$n^{opt} = n^{opt}(A)$$

However, when the instrument composition is not given, the first step has to be the identification of the instrument composition matrix  $A$  for which  $C$  is at its lowest, assuming  $n^{opt}(A)$  is employed.

$$\min_A C(A, n^{opt}(A)) \text{ under } A \text{ admissible and } A \cdot n^{opt}(A) \geq n^*$$

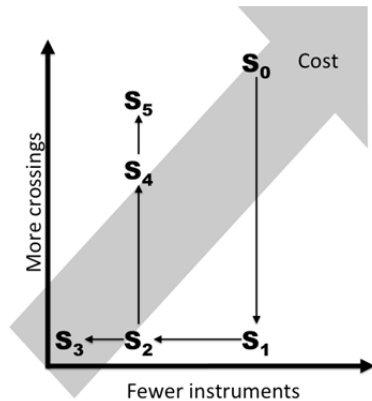
This problem is far more complex, as the cost function is highly non-linear in relation to  $A$  and the space of possible solutions is too large to be explored exhaustively. For this reason a stochastic search algorithm, **simulated annealing**, is used.

The algorithm starts with some arbitrary (but admissible) composition  $A_0$ . In each step  $t$ , a new candidate composition  $A_{t+1}$  is generated by randomly perturbing the current adopted solution,  $A_t$ . Via the simplex algorithm,  $n^{opt}$  is obtained for the new candidate composition  $A_{t+1}$ . If the cost of the candidate composition  $A_{t+1}$  is lower than the cost of  $A_t$  then it is taken as the new adopted solution. If the cost increases then the candidate may still be accepted with a certain probability which gets lower with each iteration. The algorithm is stopped when a maximum number of iterations is reached.

## 5. PILOTING THE ALGORITHM

The algorithm was tested in a series of scenarios  $S_i$  based on the specifications of three European social surveys, Labour Force Survey (EU-LFS), Statistics on Income and Living Conditions (EU-SILC) and the Adult Education Survey (AES).

The baseline scenario  $S_0$  was designed to be as close as possible to the specifications of the current three surveys considered. Five alternative scenarios ( $S_1$  to  $S_5$ ), with different design constraints, were then specified. Some of them ( $S_2, S_3, S_4, S_5$ ) allowed for a larger number of instruments in relation to today's surveys. All of the alternative scenarios imposed fewer requirements (in comparison to  $S_0$ ) concerning crossings. One scenario ( $S_5$ ) offers a greater analytical potential by including additional crossings for which needs were expressed in a user survey.



Run through these several scenarios the algorithm successfully found admissible compositions. Allowing some modules to leave their current instruments (i.e. surveys) and to be allocated to other instruments led to less costly compositions (from  $S_0$  to  $S_1$ ). Increasing the number of instruments allowed for a better sample allocation taking into account differing precision requirements between modules and consequently to less costly compositions (e.g. from  $S_1$  to  $S_2$ ). When faced with scenarios requiring a higher number of crossings (e.g. from  $S_2$  to  $S_4$ ) the algorithm found suitable



compositions. These compositions were more costly as one would expect.

## REFERENCES

- [1] European Statistical System Committee (2011), *Wiesbaden Memorandum*. Wiesbaden.
- [2] Stavropoulos, P. (2014), *Development of methods and scenarios for an integrated system of European Social Surveys – Final Report*. Produced by Agilis S.A. under Eurostat contract 61001.2011.005-2012.426.
- [3] Reis, F. (2013). *Links Between Centralisation of Data Collection and Survey Integration in the Context of the Industrialisation of Statistical Production*; WP2 presented at the UNECE Seminar on Statistical Data Collection.
- [4] Renssen, R. H. and Nieuwenbroek, N. J. (1997), “Aligning Estimates for Common Variables in Two or More Sample Surveys,” *JASA* 92, 368–375.
- [5] Merkouris, T. (2010) “An Estimation Method for Matrix Survey Sampling” *ASA, Proceedings of the Section on Survey Research Methods*, 4880–4886.

# ICT Tools for statistical linked open data:

## The OpenCube toolkit

Efthimios Tambouris (tambouris@uom.gr)<sup>1</sup>, Evangelos Kalampokis<sup>1</sup> and Konstantinos Tarabanis<sup>1</sup>

**Keywords:** Linked Data, multi-dimensional data, data cube, data analytics, visualization.

### 1. INTRODUCTION

The recent Open Data movement results in an increasing number of data offered via the Web. A significant part of these data concerns statistics. The ability to manage statistical data at a Web scale provides unprecedented analysis opportunities. Imagine if analysts could easily find statistical data coming from different sources all over the world, integrate, analyse and visualise them in any way they want.

Yet, we are currently lacking an overall understanding, including ICT tools, to enable us reaping the benefits of open statistical data. Nevertheless, at the technological level, linked data is an emerging technology with potential to overcome some of the current limitations.

In this abstract, we present the OpenCube toolkit [1] that aims at overcoming current limitations by (a) supporting the full life-cycle of statistical linked open data management from production to exploitation, and (b) providing open-source ICT tools to support advanced functionalities such as OLAP analysis, statistical analysis, data integration, and map visualizations based on linked data technologies.

### 2. STATISTICAL DATA AND LINKED DATA

Governments, organisations and companies are increasingly launching data portals that operate as single points of access for data they produce or collect [2]. A major part of these open data concerns statistics, such as population figures, economic and social indicators. For example, the vast majority of datasets published on the European Commission open data portal<sup>2</sup> are of statistical nature. In addition, many international organizations such as Eurostat<sup>3</sup>, World Bank<sup>4</sup>, OECD<sup>5</sup> and CIA's World Factbook<sup>6</sup> open up statistical data on the Web.

Statistical data is often organized in a multidimensional manner where a measured fact is described based on a number of dimensions, e.g. the unemployment rate can be described based on geographic area, time and gender. In this case, statistical data is compared to a data cube, where each cell contains a measure or a set of measures, and thus we onwards refer to statistical multidimensional data as *data cubes* or just *cubes*.

Linked data has been introduced as a promising paradigm for opening up data because it facilitates the integration of data across the Web. The term linked data refers to "*data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external*

---

<sup>1</sup> University of Macedonia and ITI-CERTH, Thessaloniki, Greece

<sup>2</sup> <http://open-data.europa.eu>

<sup>3</sup> <http://ec.europa.eu/eurostat/data/database>

<sup>4</sup> <http://data.worldbank.org>

<sup>5</sup> <http://www.oecd.org/statistics/>

<sup>6</sup> <https://www.cia.gov/library/publications/the-world-factbook/index.html>

*datasets*” [3]. Linked data is based on Semantic Web philosophy and technologies but in contrast to the full-fledged Semantic Web vision, it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inference.

In the case of cubes, linked data has the potential to realize the vision of integrating disparate and previously isolated data to perform analytics. A fundamental step towards this vision is the RDF data cube (QB) vocabulary, which enables modeling cubes as RDF graphs [4]. Recently, a few endeavors aimed at supporting data modeled according to the QB vocabulary. The resulting components and tools, however, present some limitations regarding (a) the functionalities they provide, (b) their licenses that hamper commercial exploitation, (c) their dependencies to specific platforms and environments, and (d) their ability to be used in complex scenarios in an integrated manner.

### 3. THE OPENCUBE LIFECYCLE

Exploiting statistical open data using linked data technologies calls for advances in relevant business processes. Figure 1 depicts a proposed general lifecycle that illustrates how linked data technologies can be integrated in organisations’ data management business processes [5]. The lifecycle steps are categorized in two broad phases (a) the *publish phase* that includes creating linked data cubes out of raw data, and (b) the *reuse phase* that includes exploiting linked data cubes in advanced analytics and visualizations.

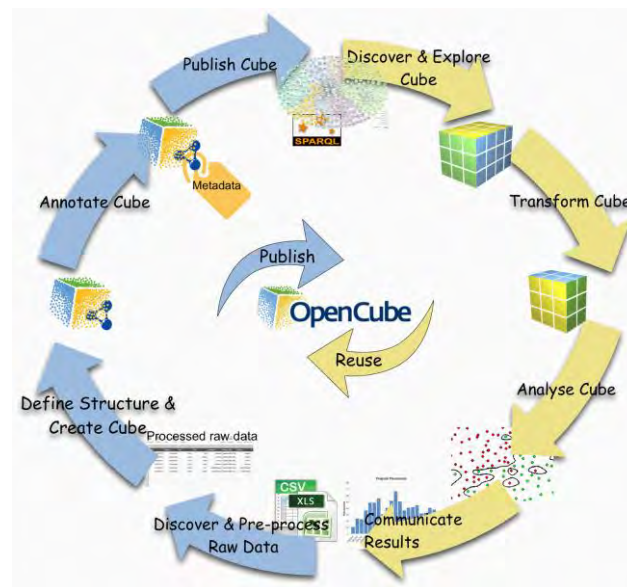


Figure 1. The OpenCube lifecycle

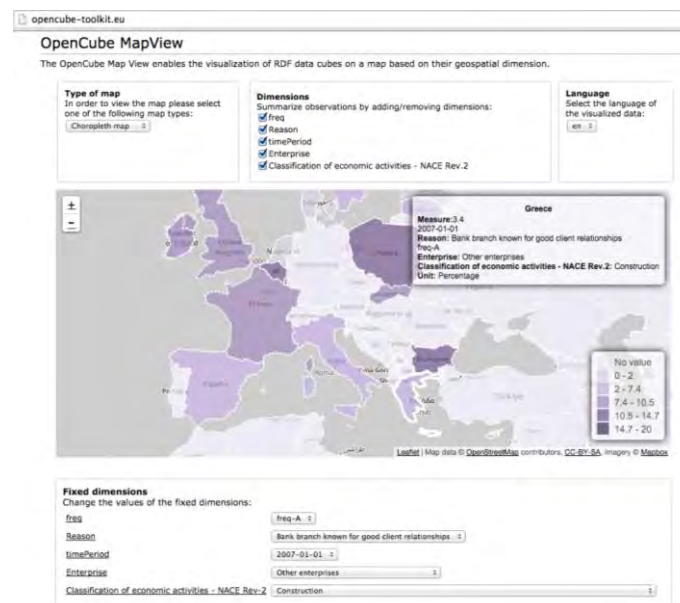
### 4. THE OPENCUBE TOOLKIT

The OpenCube toolkit includes ICT tools that support all steps in the relevant lifecycle. OpenCube tools that support the *publish phase* include:

- *TARQL extension for data cubes*: data conversion to RDF according to QB vocabulary from legacy tabular data, such as CSV/TSV files.
- *D2RQ extension for data cubes*: data conversion to RDF according to QB vocabulary from relational databases.
- *JSON-Stat to QB*: data conversion to RDF according to QB vocabulary from JSON-Stat files.

OpenCube tools that support the *reuse phase* include:

- *Data catalogue management* that provides user interface templates for managing metadata on linked data cubes and supporting search and discovery of cubes from various sources.
- *OpenCube Browser* that enables the exploration of a linked data cube by presenting two-dimensional slices of the cube in tabular form. This tool also supports OLAP-like operations such as dimension reduction.
- *OpenCube MapView* that enables the visualization of linked data cubes on a map based on their geospatial dimension. Currently, the MapView supports markers, bubbles and choropleth maps. It also supports OLAP-like operations to enable visualizing various views of a cube. For example, in Figure 2 a data cube is visualized using a choropleth heat map based on its geospatial dimension property.
- *Interactive chart visualization tool* that enables linked data cubes visualizations by summarizing data and creating charts.
- *OpenCube Aggregation tool* that pre-computes aggregations across dimensions in order to enable OLAP operations in the Browser and the MapView.
- *R statistical analysis tool* that enables implementing various statistical analysis methods on top of linked data cubes by integrating the R package in the underlying open source linked data management platform adopted by OpenCube.



**Figure 2. Visualization of a data cube that includes a geospatial dimension with the OpenCube MapView**

An important benefit of the OpenCube toolkit when compared to more traditional approaches is cubes integration. This enables expanding a cube by integrating it with a second (compatible) cube. We suggest a cube can be expanded if it is possible to increase the size of one of the sets that define a cube i.e. the set of measures, the set of objects of an attribute (level) of a dimension, the set of attributes of a dimension, or the set of dimensions. In particular, the toolkit enables (a) finding cubes on the Web of linked data that are compatible to expand an initial cube, and (b) creating and storing a new expanded linked data cube from the initial one. The expanded cube can be thereafter analyzed and visualized using the rest of the OpenCube tools.

From a technical point of view, the OpenCube toolkit is based on the Information Workbench community edition platform<sup>7</sup>, which is a linked data management platform. All the tools in the toolkit share access to a common RDF repository and can retrieve data by means of SPARQL queries. The user interface design is based on the use of wiki-based templates providing dedicated views for RDF resources.

## 5. CONCLUSIONS

A major part of open data concerns statistics that can be structured as multi-dimensional data cubes. Linked data technologies have the potential to realize the vision of finding, combining, analysing and visualizing previously isolated cubes at a Web scale. A fundamental step towards this vision is the RDF data cube (QB) vocabulary, which enables modeling cubes as RDF graphs. Current tools fall short to support the whole linked data cube lifecycle including the integration and analysis of multiple cubes. In this abstract, we presented the OpenCube Toolkit, a set of open source tools that cover the whole linked data cubes lifecycle in an integrated manner. The work reported here will continue by further improving existing tools, developing new ones and piloting them at various organizations across Europe.

## ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 611667

## REFERENCES

- [1] E. Kalampokis, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris, K. Tarabanis, Exploiting Linked Data Cubes with OpenCube Toolkit, Proc. of the ISWC 2014 Posters and Demos Track, a track within 13th International Semantic Web Conference (ISWC2014), 19-23 October 2014, Riva del Garda, Italy, CEUR-WS Vol.1272 (2014).
- [2] E. Kalampokis, E. Tambouris, K. Tarabanis, A Classification Scheme for Open Government Data: Towards Linking Decentralized Data. *International Journal of Web Engineering and Technology*, 6(3), (2011), 266-285.
- [3] C. Bizer, T. Heath and T. Berners-Lee, Linked Data—The Story So Far, *Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), (2009), 1-22.
- [4] R. Cyganiak and D. Reynolds, The RDF Data Cube vocabulary, <http://www.w3.org/TR/vocab-data-cube/> (2013)
- [5] E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris, K. Tarabanis (2014) Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services, Proc. of the 2nd International Workshop on Semantic Statistics (SemStats2014) in conjunction with the 13th International Semantic Web Conference (ISWC2014), 19-23 October 2014, Riva del Garda, Italy, CEUR-WS proceedings.

---

<sup>7</sup> [http://www.fluidops.com/en/company/training/open\\_source](http://www.fluidops.com/en/company/training/open_source)

# Official Statistics meets the Semantic Web:

## How SDMX and RDF can live together

Raffaella Maria Aracri<sup>1</sup>, Stefano De Francisci, Andrea Pagano, Monica Scannapieco  
{name.surname}@istat.it

**Keywords:** Semantic Web languages, Statistical data, Open Data, SDMX.

### 1. INTRODUCTION

In the Official Statistical (OS) domain, the issue of data interoperability has been present since decades: both National and International exchanges of data resulting from statistical processes are made possible only by adopting common metadata models and formats. Semantic Web technologies, and in particular the Linked Data initiative (see [linkeddata.org](http://linkeddata.org)) are more and more affirming as the principal mean for data interoperability, by permitting to create and interlink arbitrary volumes of structured data across the Web. In particular, the Linked Data initiative is made possible by the widespread adoption of Web standards for publishing data according to the Resource Description Framework (RDF) model. In this paper, we describe a project to integrate the SDMX [1] dissemination architecture with the Semantic Web Standards.

This work has been carried out by Istat (Italian National Institute of Statistics) within Eurostat grant “Horizontal and vertical integration: implementing technical and statistical standards in ESS”. In particular, the paper shows the design and implementation of an extension of the SDMX.Source.NET [2] to support the model and format translation of RDF. It is very much important to note that the translation step from SDMX to RDF is not only a “format” translation, but it involves a “model” translation, i.e. the specification of a set of rules to obtain RDF model’s constructs starting from SDMX model’s ones. Details of such a translation will be provided in the paper.

#### 1.1. Background

In this Section, some background information on used technologies is described.

**RDF, RDF Schema, RDF QB:** The RDF (Resource Description Framework) [3] is a standard W3C data model that has features facilitating data integration even if the underlying schemas differ. All objects are represented by URIs and URIs are linked by a simple subject-predicate-object (triple) structure. RDF data can be represented with one of the following serialization formats: (i) RDF/XML, (ii) N-triples [4]; (iii) Notation3 [5]; (iv) Turtle [6]. RDF Schema is an RDF standard [7] for ontology definition. The specialization of RDF protocol to represent statistical data is RDF Data Cube Vocabulary (RDF-QB) [8]. RDF-QB is based on the SDMX Information Model.

**SDMX – Reference Infrastructure (RI) and SDMXSource.NET:** SDMX-RI [9] has been developed by Eurostat and consists of software and tools that facilitate the production of SDMX data and their exposure via Web Services technologies. SDMX-RI features include: (i) Serving SDMX v2.0 v2.1 data and structural metadata (SOAP/REST

---

<sup>1</sup>Istat – Istituto Nazionale di Statistica

Web Services, Web Application, Windows Application); (ii) Off -line mappings to NSI's dissemination DBs (Mapping Assistant application, Mapping Store database, MS Sql Server, MySQL, Oracle, PC Axis); (iii) .NET and Java implementations; (iv) Open Source Software. SDMX.Source.NET is the .NET implementation of SDMX source on which RI tools are built.

**National Statistical Institutes Web Services:** NSI Web service (WS) is the component of SDMX-RI that provides data to connected consumers in various formats, according to their requirements. Using the HTTP Content Negotiation [10] mechanism, the client specifies the desired format and version of the resource using the Accept HTTP header. The NSI Web Service of SDMX-RI can make data available by: (i) SDMX-ML Generic Data Format, version 2.1; (ii) SDMX-ML Structure Specific Data Format, version 2.1; (iii) SDMX-ML Structure Format, version 2.1. The specific objectives of the present work are to: (i) design a translation mechanism from SDMX data model to RDF Data Cube Vocabulary, (ii) design and develop the RDF extension of the NSI WS, that supports RDF-XML Data Format, version 2.1 and RDF-XML Structure Format, version 2.1.

## 2. METHODS

In the first part of the work, we describe the design solution to map elements of the SDMX data model to elements of the RDF-QB data model [11]. In the second part, we describe the design and the development of the software to extend the SdmxSource.NET to the new RDF-QB output.

### 2.1. Modeling: from SDMX to RDF Data Cube Vocabulary

Given the space constraints, we cannot provide the detail of the rules, but we highlight the SDMX-RDF mapping complexity. Let us consider the component REF\_AREA stated in the key family section of the DSD related to “Separate collection indicators” of the Istat data on Environment and Energy Waste Figure 1. In general the component is more verbose when described in RDF-QB as it must be declared not only as a dimension of the Data Cube Vocabulary but also as an SDMX DimensionProperty and as an SDMX CodedProperty as it refers to a specific code list (in this case “territory” code list).

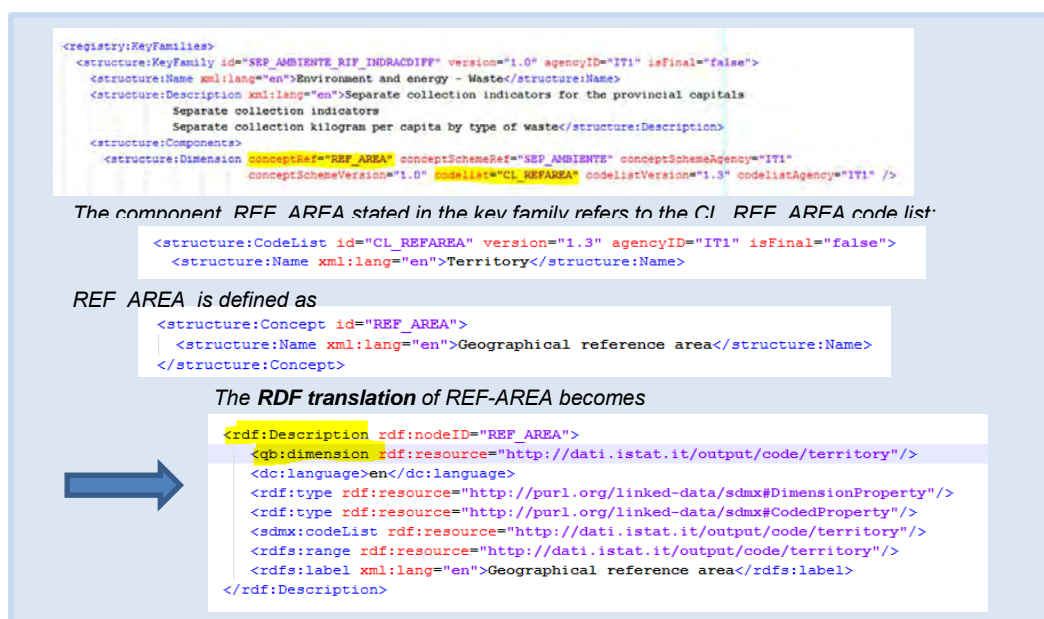
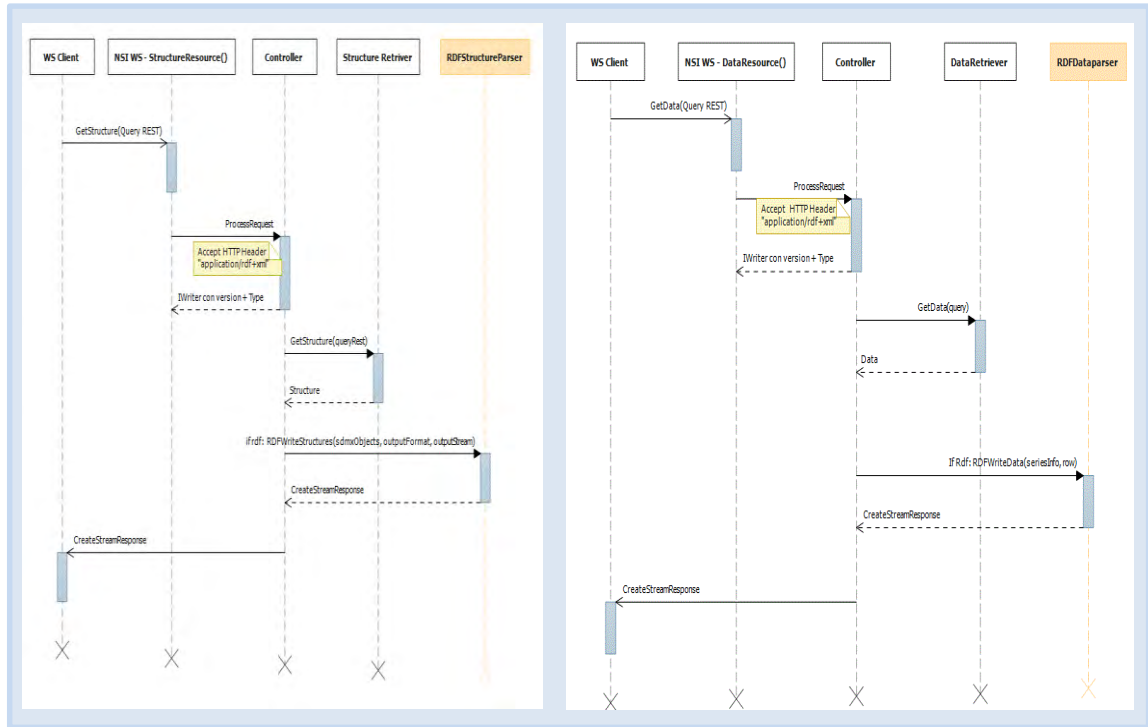


Figure 1. SDMX-RDF mapping of REF\_AREA component

## 2.2. Adding structure and data format to SdmxSource.NET

In this section we show how we design and develop the code to extend the SdmxSource.NET to the new RDF-QB output. In doing this we leverage the existing SdmxSource.NET source code wherever possible and we develop specific add-ons that can be included as separate libraries in future SdmxSource.NET releases.



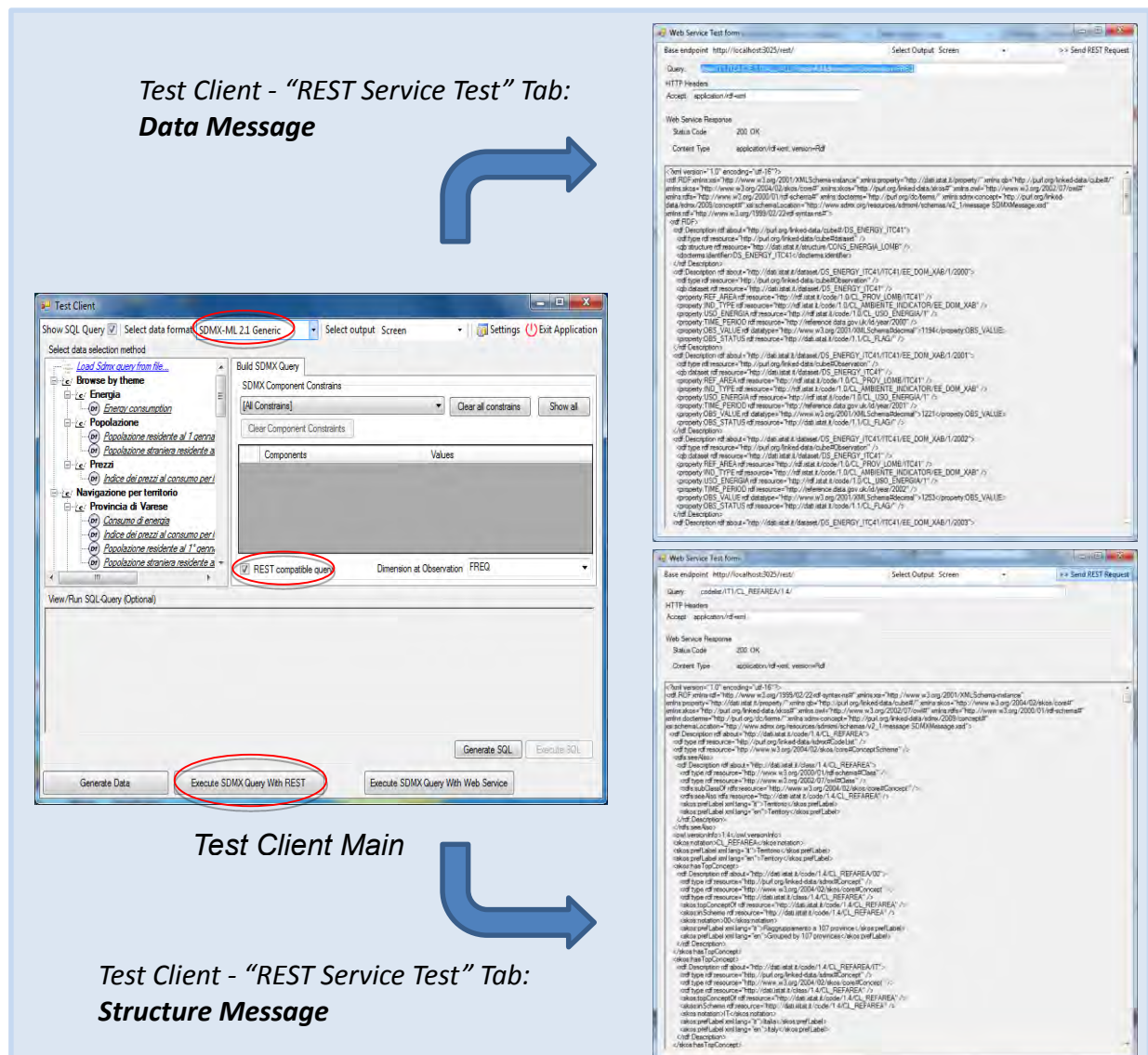
**Figure 2: Sequence diagrams for request RDF Structure (left) and Data (right)**

The sequence diagram in Fig.2 shows how the newly implemented “Writer” for structural metadata and data in RDF format works. As shown in the left picture when a client requests structural metadata in RDF format to the NSI WS a Controller module processes the request instantiating a Generic Writer containing the output format type and version (RDF V.2.1) the Structure Retrieve module connects to the Mapping Store DB and gets back the structural metadata. At this point the Controller specializes the Generic Writer into a RDF-Writer that returns the metadata into the required format to the client. Similar process is applied for data requests (right picture in Fig. 2).

## 3. RESULTS

To validate the results we leverage the “Sdmx-RI Test Client”. The Test Client is a tool to test the SDMX Reference Infrastructure building blocks and to expose/browse the dissemination environment of an NSI. First of all we deploy a new NSI WS on MS-IIS (Microsoft© Internet Information Services), then we use the Test Client (Fig. 3) to inquiry it and validate the RDF returned messages.





**Figure 3: Test Client: RDF Data Message and Structure Message**

In the Test Client the user can check the “REST compatible query” option in order to build a REST query for the SDMX-ML 2.1 version. Then the user clicks the “Execute SDMX Query with REST” button, to display the “REST Service Test” window: The application provides the REST query, starting from the selected DataFlow of the main form, and adding the http Header “application/rdf+xml”. Pressing “Send REST Request” button the Test Client invokes the NSI WS available at the “Base Endpoint” and returns the DataFlow in RDF format.

#### 4. CONCLUSIONS

The paper describes the solution implemented by Istat to integrate RDF-based technologies with the SDMX-based dissemination architecture promoted by Eurostat.

The major outcome of this work is the proof that the two world can live together: indeed, a direct translation from the SDMX NSI WS is possible. On the other side, we highlight that such a solution can be also integrated with a Linked Open Data dissemination channel based on a SPARQL Endpoint. Indeed, so far the URIs produced by the SDMX translation steps are not actual URIs: they need a “deployment” on a SPARQL endpoint. If such data were actually “deployed”, instead, the result of the SDMX translation could pass from the document/dataset state to the one of singly deployed data values.

## REFERENCES

- [1] SDMX: <http://sdmx.org/>
- [2] SDMX Source: <http://www.sdmxsource.org/>
- [3] RDF Specification: <http://www.w3.org/standards/techs/rdf/>
- [4] W3C, “N-Triples”, <http://www.w3.org/TR/n-triples/>
- [5] W3C, “Notation3”, <http://www.w3.org/DesignIssues/Notation3>
- [6] W3C, “Turtle”, <http://www.w3.org/TeamSubmission/turtle/>
- [7] RDF Schema: <http://www.w3.org/TR/rdf-schema/>
- [8] RDF Data Cube Vocabulary: <http://www.w3.org/TR/vocab-data-cube/>
- [9] SDMX-RI <https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp>
- [10] HTTPContent Negotiation: <http://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html>
- [11] Aracri R., De Francisci S., Pagano A., Scannapieco M., Tosco L., Valentino L.. :  
“Integrating Statistical Data with the Semantic Web: The ISTAT Experience” negli  
Atti della Congresso Nazionale AICA “Frontiere Digitali: dal Digital Divide alla  
Smart Society” 8-20 Settembre 2013

# Poverty mapping at local level with suitable modelling of income

Isabel Molina ([isabel.molina@uc3m.es](mailto:isabel.molina@uc3m.es))<sup>1</sup>, Monique Graf<sup>2</sup> and Juan Miguel Marín<sup>1</sup>

**Keywords:** empirical Bayes, linear mixed models, poverty indicators, small area estimation.

## 1. INTRODUCTION

Maps of poverty and other socioeconomic indicators support extremely important political decisions such as the allocation of regional development funds by governments and international organizations. As a matter of fact, the World Bank (WB) delivers poverty maps for many countries all over the world. In the U.S., the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program provides annual estimates of income and poverty statistics for all school districts, counties, and states, for the administration of federal programs and the allocation of federal funds to local jurisdictions. State and local programs also use the income and poverty estimates for distributing funds and managing programs. Political decisions should be based on the most accurate wellbeing statistical figures.

The problem arises because the official data sources used to produce poverty maps in many countries, such as the European Survey on Income and Living Conditions (EU-SILC), have a sample size that is planned to give estimates only for large regions. This limited sample size does not allow the production of estimates of desired precision at local level or for some risk population subgroups. The resulting sample sizes in some of these local regions or population subgroups (called here small areas in general) can be too small and, as a consequence, direct estimates, calculated using only the scarce area-specific data, can have unduly large sampling errors. This problem can be approached using small area estimation techniques, designed to improve the precision of estimates at local level. Since the total sample size of the surveys is typically large, finding some relationship among the areas in the form of a model that links all the areas often helps to obtain estimates with better precision. Small area estimation models can provide estimates with notably higher precision by combining the sample data from all the areas, see e.g. Rao (2003) for a comprehensive account of small area estimation techniques or Pfeiffermann (2014) for a recent review.

## 2. METHODS

Small area estimation models can be stated at the area level, using only the aggregated information for the areas, or at the unit level, making use of the unit-specific values of the target and auxiliary variables. We focus on the latter type of models because unit level information is typically much richer but area level auxiliary variables can also be included. Moreover, these models can be applied in a straightforward manner to estimate any non-linear parameter that is function of the (continuous) response variable in the model.

---

<sup>1</sup> Department of Statistics, Universidad Carlos III de Madrid

<sup>2</sup> Institut de Statistique, Université de Neuchâtel, Elpacos Statistics, la Neuveville, Switzerland

Most poverty indicators are non-linear functions of a welfare variable such as income or expenditure. The first method designed to estimate general non-linear parameters in small areas is ELL method (Elbers, Lanjouw and Lanjouw 2003), used by the WB to construct poverty maps at local level. This method assumes a (unit level) linear mixed model for the log income or other variable used to measure the wellbeing. Molina and Rao (2010) have shown that the poverty estimates obtained by the ELL method can have poor accuracy. The empirical best (EB) method of Molina and Rao (2010) gives an approximation to the best estimates in terms of mean squared error (MSE), provided that the log incomes (or other one-to-one transformation of the welfare variable) are normally distributed. However, the histogram of the log incomes for several European countries displays a left skewed distribution with a thin but long left tail. Thus, the normality assumption does not really hold for the log-income data. This left tail is caused by the presence of individuals with atypically small incomes. These extreme values are hardly explained by auxiliary variables and produce a bias in both EB and ELL estimates that might be significant for some regions.

As an alternative to the log-normal model for the incomes used in EB and ELL methods, we propose to consider a much more flexible distribution called generalized beta distribution of the second kind (GB2). The success of the GB2 distribution in modelling income data has been shown by many authors, see e.g. by McDonald (1984). Recently, Graf and Nedyalkova (2013) have used the GB2 to estimate poverty and social exclusion indicators under a large population framework. The four parameters of the GB2, with one parameter controlling each tail, give a lot of flexibility to accommodate distributions with different types of skewness. As particular cases, it includes distributions such as Fisk, also called log-logistic, Dagum and Singh-Maddala. A limiting case is the Generalized Gamma distribution.

A model for small area estimation is proposed based on a multivariate extension of the GB2 family of distributions. Maximum likelihood is used to fit this model. The best estimator of a general area parameter is given by the expectation of the parameter with respect the distribution of the non-sample given the sample data in that area. To calculate this expectation, we find the distributions of conditional random vectors of the form  $\mathbf{Y}_1|\mathbf{Y}_2$ , where  $(\mathbf{Y}_1', \mathbf{Y}_2')$  follows a multivariate GB2 distribution. Using that result, we obtain a Monte Carlo approximation of the conditional expectation defining the empirical best estimator. But this Monte Carlo approximation entails generating a large number of full populations (or censuses) of the target variable from the conditional distributions. For real life areas such as e.g. Spanish provinces, the dimension of the multivariate GB2 vectors to generate (population sizes of provinces) is huge and raw generation of these vectors many times is unfeasible in a reasonable time. Efficient computational algorithms are developed to generate full populations of the study variable and thus approximating the EB estimates based on the proposed GB2 model in reasonable time.

Assessing the reliability of the obtained EB estimates is indeed crucial, since these estimates are supposed to be more precise than direct estimates. To estimate mean squared errors, we develop bootstrap procedures under the proposed model, and the properties of the bootstrap methods will be studied.

### 3. RESULTS

Simulation results indicate that, when income follows the considered multivariate GB2 model, the EB method based on this model gives less biased and more efficient estimates of poverty indicators than the usual EB method based on the log-normal distribution for income. Moreover, if income follows the log-normal nested error model, then EB

estimates based on the multivariate GB2 model perform similarly as the corresponding EB estimates based on the log-normal model.

#### **4. CONCLUSIONS**

The log-normal distribution does not fit well the income distribution. Consequently, poverty estimates based on the usual nested error model for the log incomes are biased. We propose a multivariate GB2 model that fits better the distribution of income. We obtain EB estimates that are approximately optimal under the proposed model. Simulation results support the use of the proposed EB method based on the multivariate GB2 model.

#### **REFERENCES**

- [1] J.N.K. Rao, Small Area Estimation (2003), Hoboken, New Jersey: Wiley
- [2] D. Pfeffermann, New Important Developments in Small Area Estimation, *Statistical Science* (2013), 28, 40-68.
- [3] C. Elbers, J.O. Lanjouw, and P. Lanjouw. Micro-level estimation of poverty and inequality. *Econometrica* (2003), 71, 355-364.
- [4] I. Molina, and J.N.K. Rao. Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics* (2010) 38, 369-385.
- [5] J. McDonald. Some Generalized Functions for the Size Distribution of Income. *Econometrica* (1983), 52, 3, 647-663.
- [6] M. Graf and D. Nedyalkova. Modeling of income and indicators of poverty and social exclusion using the Generalized Beta Distribution of the Second Kind, *Review of Income and Wealth* (2013). DOI: 10.1111/roiw.1231.

# SMALL AREA ESTIMATES OF INCOME: MEANS, MEDIANS AND PERCENTILES

Alison Whitworth ([alison.whitworth@ons.gsi.gov.uk](mailto:alison.whitworth@ons.gsi.gov.uk))<sup>(1)</sup>, Kieran Martin<sup>(2)</sup>, Nikos Tzavidis<sup>(3)</sup>, Marie Cruddas, Christine Sexton, Alan Taylor

**Keywords:** Model based estimates of income, Average household income, Median household income, Quantiles for income, Empirical best predictor.

## 1. INTRODUCTION

The Office for National Statistics (ONS) currently produces model based estimates of mean household income and the proportion of households below the nationally defined poverty line for Middle Layer Super Output Areas (MSOAs) in England and Wales, (ONS, 2010<sup>[1]</sup>). The estimates have limitations in meeting user needs though because the mean income provides little information about the distribution across households and can be inflated by the relatively small number of households with very large incomes. Estimates of the median income (together with other quantiles) are considered to be more useful and would better inform user requirements.

The current methods used by ONS for the published small area income estimates, model survey data in terms of area level aggregates and cannot be easily modified to estimate median income. Recent advances in small area estimation, for example the methods developed by Molina and Rao (2010)<sup>[2]</sup>, provide a flexible approach by using simulation techniques based upon modelled parameters to obtain estimates for the whole population. The simulated estimates are then used to derive measures for the distribution.

This paper explores an application of the method provided by Molina and Rao, applying it to 2001 Census data from the North West and South East regions of England to provide estimates of medians, means and quantiles of income at MSOA level. The method is assessed by measuring the precision of the estimates, and comparing them against the published estimates as well as independent “proxy” data for income such as the indices of deprivation.

## 2. METHODS

### 2.1. Current approach for estimates of household income

The current approach described in ONS (2010), is to estimate the area (MSOA) level relationship between the survey variable and auxiliary variables by regressing individual responses from the Family Resource Survey (FRS) on area values of the covariates. The FRS is the survey with the largest sample that includes suitable questions on income and aims to interview all adults in a selected household. In 2001 a final sample size of 23,790 households was achieved. The auxiliary variables are generally average values of proportions relating to all individuals or households in the area and are based on administrative or census data with coverage in all the areas being modelled.

<sup>1</sup> Office for National Statistics, Segensworth Road, Titchfield, Hants

<sup>2</sup> This paper reports on part of an ongoing project within the Small Area Estimation Branch, ONS, to improve the current published estimates of income and poverty. The initial development and programming was undertaken by Dr Kieran Martin who has since left the team.

<sup>3</sup> Dept of Social Statistics & Demography, Southampton University, Hants

The model for income is:

$$\ln(y_{ir}) = \alpha + \beta \bar{X}_{k(ir)} + u_r + e_{ir};$$

$y_{ir}$  is weekly income for household  $i$  in postcode sector (PCS)  $r$ ;

$\bar{X}_{k(ir)}$  is the population mean for the covariate in MSOA  $k$  that household  $i$  in PCS  $r$  falls within;

$\alpha$  and  $\beta$  are the regression parameters for intercept and slope respectively;

$u_r$  is the area level residual assumed to have expectation 0 and variance  $\sigma_v^2$ ;

$e_{ir}$  is the individual within area residual, with expectation 0 and variance  $\sigma_e^2$ .

Once the model has been fitted the model parameters are applied to the covariate values for each area to obtain the target estimates. While the model is constructed only on responses from sampled areas, the relationships identified are assumed to apply nationally.

## 2.2. Empirical best predictor (EBP) approach

The EBP approach starts by fitting a standard mixed effects model which relates the observed survey household income to a set of covariates common to both the survey and census. The estimated parameters of the distribution of the out of sample income can then be obtained using the estimated parameters from the fitted model. The next step is to produce the EBP of the statistic we are interested in, for example MSOA household median income. For each *out of sample* household in the population a fixed number of estimates of household income are simulated through random sampling of the estimated conditional distribution, and the median household income is found for each MSOA for each of these simulations. The EBP of median household income for an MSOA is then obtained by averaging over all of the medians obtained from each of the simulations for that MSOA.

The final stage is to obtain an estimate of the variation associated with the estimated median. A parametric bootstrap sampling technique is applied which samples the distribution of the fitted model to produce a large number of bootstrap census populations where every household has an estimated income. The 'true' median household income is obtained for each MSOA in each bootstrap population. From each bootstrap population a new sample is drawn and the EBP estimation process applied to obtain a bootstrap estimate of the MSOA median household income. The MSE can then be calculated from the 'true' median household incomes and the estimated median household incomes.

Since estimates of income are obtained for individual households at each stage, it is relatively easy to obtain whatever small area statistic is required, for example mean household income, median household income, proportion of households in poverty, etc.

For the application presented in this paper the method was applied to produce estimates of household income at MSOA level using the FRS individual household level covariates, 2001 Census household level covariates and MSOA level covariates sourced from the 2001 Census, the DWP administrative data and other sources of administrative data. (Details of the covariates considered are included in the paper.) Household income was defined as net household weekly income (adjusted for household size and composition after housing costs).

### 3. RESULTS

The median weekly household income for MSOAs in the North West and South East of England was £299 in 2001. The median weekly income for the 25th percentile was £197; so on average the bottom 25% of households within MSOAs had a weekly household income of £197 or less. Conversely on average the top 25% of households within MSOAs had a weekly household income of £454 or more. The median weekly household income for MSOAs in the North West was lower than that for the South East (not shown).

**Table 1. Summary statistics for the MSOA percentiles of household income**

Percentile	Minimum	Median	Maximum
2.5	£56	£89	£154
25	£132	£197	£353
50	£204	£299	£548
75	£314	£454	£847
97.5	£708	£992	£1,905

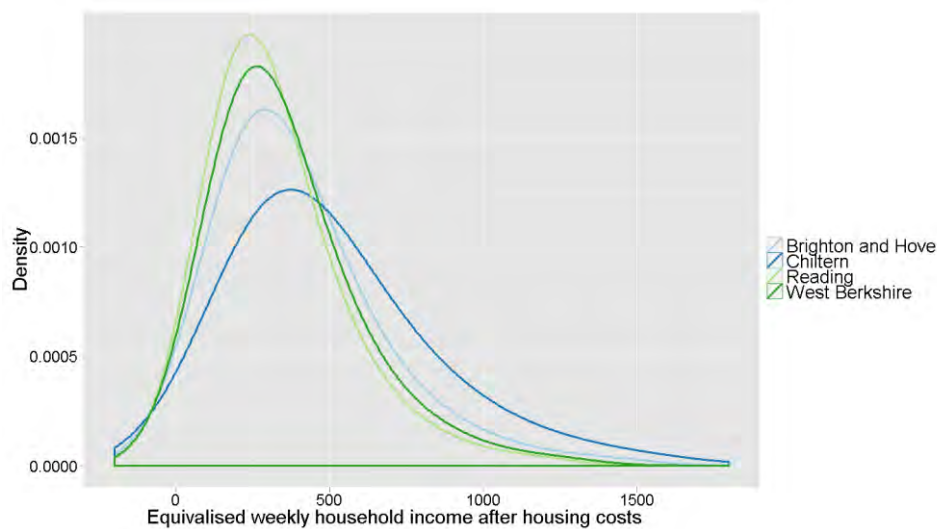
Table 2 gives summary statistics for the coefficients of variation (CVs). All CVs of variation remain small, indicating that a high degree of accuracy can be achieved. As might be anticipated, the percentiles with the most data, such as the median, have the lowest coefficient of variation, while for the 2.5th and 97.5th percentiles the coefficient of variation is larger.

**Table 2. Summary statistics for the coefficients of variation for the MSOA percentile income estimates**

Minimum	1 <sup>st</sup>	Median	3 <sup>rd</sup>	Maximum
0.05	0.07	0.07	0.07	0.09
0.05	0.06	0.06	0.07	0.08
0.05	0.06	0.06	0.07	0.08
0.05	0.06	0.06	0.07	0.08
0.06	0.07	0.07	0.07	0.09

If more percentiles are calculated they can be used to produce visualisations of the income distribution in an MSOA. Figure 1 shows, for example, that for the MSOA in Reading there was a high peak in the proportion of households with income at approximately £250 per week and a steep decline in the distribution for higher incomes (with virtually no households above £1,500 per week), whereas for the MSOA in Chiltern the peak in the distribution was much less acute and at a higher income at around £350 to £400 per week. For Chiltern there was a more gradual decline in the proportion of households with higher income and some have an income greater than £1,500 per week.





**Figure 1. Income distribution across four different MSOAs**

#### 4. CONCLUSIONS

This application provides a practical demonstration of the method and provides proof of concept that estimates have been derived for the mean, median and quantiles of income as well as for poverty, all under one estimation model. Assessment of the estimates was promising; the coefficients of variation indicated that the estimates are precise and the low income and poverty estimates were significantly different from the higher ones. Although not directly comparable, assessment of the EBP estimates against proxy estimates gave assurance that model bias is relatively small.

A limitation of the method however, is that individual level census data is required and so estimates can only be derived in census years. The usefulness of an individual set of estimates and how they can be used alongside the series of official estimates must be established. They will only provide a viable alternative to the current approach if the method can be extended to provide estimates between census years.

##### 4.1. Future work

The next step is to apply the method to all areas of England and Wales and use 2011/2012 data to provide updated estimates. Benchmarking the updated EBP estimates against direct estimates of income at region and country level for England and Wales (as undertaken for the official estimates) will ensure consistency in outputs across geographies and will facilitate a full assessment including more direct comparison between estimates produced using the EB method and the official estimates. The diagnostic plots for the model indicated that the underlying modelling assumptions may not be fully valid; residual analysis demonstrated a marked deviation away from the normal distribution. An alternative transformation of the response variable or an alternative model may prove more appropriate for the data.

Finally a feasibility exercise should be undertaken to assess the potential to extend the approach to update the Census covariate data used in the model so that updated estimates could be derived between census years. This would provide a means for a continuous series of estimates with the flexibility to meet user requirements.

## REFERENCES

- [1] ONS Validation report: Model-based estimates of households in poverty for Middle Layer Super Output Areas in England and Wales, ONS publication: <http://tinyurl.com/9emp8>, 2007/2008.
- [2] I. Molina, and J. N. K. Rao, Small area estimation of poverty indicators. The Canadian Journal of Statistics **38**, (2010), 369-385

# Comparing small area estimation methods for poverty indicators in the municipalities of Minas Gerais State

Solange Correa ([s.correa-onel@soton.ac.uk](mailto:s.correa-onel@soton.ac.uk))<sup>1</sup>, Debora Souza<sup>2</sup>, Nicia Brandolin<sup>2</sup>, Viviane Quintaes<sup>2</sup> and Djalma Pessoa<sup>2</sup>

**Keywords:** Small Area Estimation, Poverty Indicator, Inequality.

## 1. INTRODUCTION

Official statistical agencies have been facing growing demand for detailed and accurate information and, on the other hand, suffer from constant financial constraint in the production of sample surveys. In this sense, researchers have been studying methods for small domain estimation in order to provide estimates for small geographic areas or domains with controlled precision without increasing costs.

Methods to estimate poverty in small areas combine information collected from multipurpose household sample surveys with information from censuses with broad geographic coverage. The poverty map published in 2008 by the Brazilian Institute of Geographic and Statistics [1] adopted the method proposed by Elbers *et al.* [2], the so called ELL method, and aimed to give a detailed description of the spatial distribution of poverty in the country. More recently, Molina and Rao [3] have developed a methodology (MR approach) that uses as much sample information for estimation of poverty indicators in small areas as possible. According to the authors, this proposal results in lower mean square errors compared to the ELL approach.

The main objective of this work is to compare in terms of bias and mean square error (MSE) the methods proposed by Elbers *et al.* [2] and Molina and Rao [3]. A simulation study is conducted using data from the 2008-2009 Brazilian Budget and Expenditure Survey (POF) [4].

## 2. METHODS

Consider a finite population of size  $N$  partitioned into  $D$  areas of sizes  $N_1, \dots, N_D$ . Suppose that  $E_{dj}$  is an adequate measure of welfare, e.g. income or expense, for individual  $j$  in small area  $d$ . Let  $z$  be a given poverty line, i.e. the threshold for  $E_{dj}$  below which a person is considered poor. The *FGT* family of poverty indicators for each small area  $d$  [5] is given by

$$FGT_{cd} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left( \frac{z - E_{dj}}{z} \right)^c I(E_{dj} < z), \quad c = 0, 1, 2; \quad d = 1, \dots, D, \quad (1)$$

---

<sup>1</sup> University of Southampton, UK

<sup>2</sup> Brazilian Institute of Geography and Statistics (IBGE)

where  $I(E_{dj} < z) = 1$  if  $E_{dj} < z$  (person is considered poor) and  $I(E_{dj} < z) = 0$ , otherwise.

For  $c = 0$ , expression (1) becomes the poverty rate or poverty incidence in area  $d$ . When  $c = 1$ , the  $FGT$  measure is called the poverty gap, i.e. the average relative distance to non-poverty for individuals in area  $d$ . When  $c = 2$ , expression (1) is called poverty severity and emphasises the extreme poverty.

The ELL method consists in estimating social welfare for the same geographic level of the Demographic Census. This information is obtained directly from external sources or derived from household sample surveys. As a result, it is possible to use the geographically detailed information from the Census to estimate poverty indicators and corresponding variances for small domains. The idea of the MR approach is to combine household survey data to Census data and apply random intercept two-level models to predict the welfare variable.

To compare the ELL and MR approaches, samples are selected from the Demographic Census data using the same complex sampling design used in the 2008-2009 Brazilian Budget and Expenditure Survey (POF). For each sample, poverty incidence and poverty gap are estimated by both methods and compared with those calculated in the population. This study may provide subsidies to choose an appropriate methodology to produce a poverty map for the country.

### 3. RESULTS AND CONCLUSIONS

In the simulation study, 400 random samples are selected from the 2010 Demographic Census data for the Minas Gerais State in Brazil using the same sampling design adopted by POF. Both methodologies, ELL and MR are applied to each sample and estimates of poverty measures (poverty incidence and poverty gap) obtained. These estimates were then compared in terms of bias and MSE with the quantities values calculated directly from the Census data. The small area of interest in this study is the municipality and the variable of welfare considered is the *per capita* household income.

Table 1 and Table 2 show descriptive statistics for the relative MSE and relative bias of the poverty rate and poverty gap, respectively, for municipalities of Minas Gerais State. Results show that both the relative bias and the relative MSE a higher for municipalities with lower values of poverty incidence. In addition, both methods overestimate the poverty rates, with the MR approach producing higher relative bias and relative MSE. Similar pattern was observed for the poverty gap.

Table 1 Descriptive statistics for the relative MSE and relative bias of the poverty rate for municipalities of Minas Gerais State.

Statistic	Estimate (%)			Relative MSE (%)		Relative Bias (%)	
	Population	ELL	MR	ELL	MR	ELL	MR
Minimum	7.32	11.29	9.81	0.04	0.05	-24.04	-21.38
1st Quantile	18.02	22.64	25.68	0.46	1.06	3.13	8.15
Median	26.99	29.92	32.16	1.86	4.27	12.52	19.86
Mean	29.07	32.16	34.3	7.71	15.39	17.42	27.49
3rd Quantile	39.15	40.5	41.92	8.06	16.98	27.62	40.27
Maximum	65.25	65.46	66.39	186.24	318.55	133.93	176.49

Table 2 Descriptive statistics for the relative MSE and relative bias of the poverty gap for municipalities of Minas Gerais State.

Statistic	Estimate (%)			Relative MSE (%)		Relative Bias (%)	
	Population	ELL	MR	ELL	MR	ELL	MR
Minimum	1.93	3.46	3.22	0.12	0.07	-45.15	-36.11
1st Quantile	6.21	7.39	9.19	0.68	1.12	-7.99	3.82
Median	10.35	10.54	12.38	2.03	4.62	4.22	19.73
Mean	12.41	12.32	14.05	7.88	23.94	9.52	30
3rd Quantile	17.25	16.34	17.92	5.83	23.89	21.11	47.77
Maximum	40.26	35.78	36.81	350.13	713.31	182.83	261.6

This is work in progress and extensions include comparison of the variance estimates obtained under both approaches.

## REFERENCES

- [1] IBGE, Mapa de pobreza e desigualdade: municípios brasileiros 2003, Rio de Janeiro: IBGE (2008).
- [2] C. Elbers, J. Lanjouw and P. Lanjouw, Micro-level estimation of poverty and inequality, *Econometrica* 71 (2003), 355-364.
- [3] I. Molina and J. Rao, Small area estimation of poverty indicators, *Canadian Journal of Statistics* 38 (2010), 369-385.
- [4] | IBGE, Pesquisa de Orcamentos Familiares 2008-2009 Despesas, Rendimentos e Condições de Vida, Rio de Janeiro: IBGE (2010).
- [5] J. Foster, J. Greer and E. Thorbecke, A class of decomposable poverty measures, *Econometrica* 52 (1984), 761-766.

# Analysing whether sample survey data can be replaced by administrative data

Arnout van Delden<sup>1</sup> (adln@cbs.nl), Reinder Banning<sup>1</sup>, Arjen de Boer<sup>1</sup> and Jeroen Pannekoek<sup>1</sup>

**Keywords:** registers, administrative data, correspondence, quality indicators, official statistics

## 1. INTRODUCTION

For National Statistical Institutes, administrative data sources are an attractive alternative to survey sampling to produce official statistics. When we have an administrative data source in mind for a certain official statistic that is up till now based on survey sampling, the question may arise whether these data can replace the survey sampling as the source for the observations, thus reducing response burden. The answer to that question may depend, amongst other factors, on the level of correspondence between the variables in the administrative data source and the target variables.

We believe that in judging this level of correspondence two aspects should be taken into account: the *conceptual* as well as the *numerical* differences between the variables in the administrative data source and the target variable. A number of papers focus on one of the two aspects. Work by [1] studying international trade in goods and services, put most attention on identifying the *conceptual* differences between administrative and sample survey data. [2] giving a set of indicators and [3] studying employment-related variables concentrate on *numerical* correspondence.

In the current paper, we apply regression analyses where we look into both numerical and conceptual correspondence. We do so for a case study on value added tax (VAT). In the second half of 2011, Statistics Netherlands implemented a production system that uses VAT-data for the smaller units and sample survey data for the largest units to estimate turnover levels to be published by the structural business statistics (SBS) and changes in turnover for the short term statistics (STS) [4]. Before 2011, both small and large units received a sample survey. The question at the time was: in which of the domains of economic activity can we replace the STS survey sampling by the administrative data? At the time, we visually inspected regressions of VAT-turnover against turnover from survey data to decide on that question [4]. In the current paper we seek for a more objective approach by using indicators obtained from regression analyses.

## 2. CONCEPTUAL DIFFERENCES

We analysed conceptual differences between turnover derived from the Dutch VAT form based on VAT regulations and turnover according to the European SBS regulation. We found that VAT regulations depended on the economic activities. We grouped economic activities according to the NACE rev 2 classification into so-called “base cells”, the smallest grouping from which we can produce our output domains. Those base cells spanned the whole domain of economic activities of the SBS regulation.

---

<sup>1</sup> Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands

We found twelve different VAT-regulations. Experts related those regulations to the base cells. Within each base cell one or more regulations may be applicable simultaneously, leading to a total of 27 unique sets of regulations - including “no regulation” - among the base cells. The most frequently occurring sets of regulations, in terms of the number of base cells to which they apply, are shown in Table 1.

Table 1. Number of base cells for each unique (set of) regulation(s), sorted by descending number of base cells.

Number of base cells	[name] description of regulations	Expected effect
85	No regulations	-
64	[Not charged – foreign services, NC-FS]. Before 2010 the VAT of <i>foreign services</i> were not collected on a separate item of the VAT form, since 2010 it is.	VAT < target from 2010 onwards
35	[International Trade, IT]. All kinds of regulations to declare international transactions, but turnover should be derived correctly	Small effect
18	[Transfer, TR] Subcontractors may transfer the VAT payment to the main building contractor <i>as well as regulation</i>	Small to no effect
	[Not charged – Foreign Turnover NC-FT] Certain activities of turnover obtained in foreign countries are not charged by VAT	Small to medium effect VAT < target
17	[Not charged –Derogation (NC). Derogation from VAT for certain activities for instance ambulances	Large effect VAT << target
16	[Transfer, TR] Explained above	Small to no effect
89	‘Others sets’ (here not specified further)	
324	Total	

### 3. METHODOLOGY OF NUMERICAL ANALYSES

#### 3.1. Data sets for numerical analyses

For the sample survey we used turnover from SBS survey sampling, which has the same definition as the STS turnover: in both cases the target turnover is requested. For the years 2009 and 2010 we linked SBS enterprises and administrative VAT-units to our General Business Register (BR). We only included VAT units that we could link uniquely to a statistical unit and that reported turnover during the whole year, in both sources. As a result, we had data for 55 per cent of the enterprises in our BR.

Within the data set a number of implausible cases were found with a very small VAT turnover but a very large SBS turnover and vice versa. To reduce the risk of potential (linkage) errors, we excluded units with VAT or SBS turnover  $\leq 10,000$  euro, which correspond to very small yearly turnover values.

#### 3.2. Indicators from numerical analyses

Visual data inspection showed that SBS turnover was linearly related to VAT turnover in practically all base cells. We therefore applied a linear model for each base cell  $k$ . Let  $y_{ki}^t$  denote the SBS turnover of unit  $i$  of base cell  $k$  in year  $t$ , and let  $x_{ki}^t$  be the corresponding VAT turnover. Further, we use indicator variable  $\delta_{ki}^t$  with 0 for  $t=1$  (2009) and 1 for  $t=2$  (2010). We used a linear model for the combined data of both years, according to:

$$y_{ki}^t = \alpha_k + d\alpha_k \delta_{ki}^t + (\beta_k + d\beta_k \delta_{ki}^t) x_{ki}^t + \varepsilon_{ki}^t$$

where  $\alpha_k$  stands for the intercept at  $t=1$ ,  $\beta_k$  stands for its slope,  $d\alpha_k$  stands for the year-effect on the intercept ( $t=2$  minus  $t=1$ ) and similarly  $d\beta_k$  for the slope, with errors  $\varepsilon_{ki}^t$  that are assumed i.i.d. with mean 0 and variance  $(\sigma_k)^2$ .

We used a weighted least squares loss function. We included weights  $1/x_i$  to deal with heteroscedasticity and included calibration weights to estimate the regression for the population based on the sample data. Furthermore, we accounted for outliers by using a robust regression method: an M-estimator with a Huber weight function.

We aim to use the conceptual analyses and indicators from numerical analyses to sort the base cells into four ‘situations’, namely base cells (1) without conceptual differences, (2) with conceptual differences but only small numerical differences, (3) with conceptual differences and systematic numerical differences and (4) with conceptual differences and non-systematic numerical differences. Administrative data can be used to replace survey data in situation 1 and 2, not in situation 4. They could be used with a correction factor in situation 3, but only when the numerical and conceptual differences are in line with each other.

We used the following indicators:

- a weighted coefficient of determination,  $R^2(w)$ ;
- mean absolute percentage error between VAT and SBS turnover,  $M_k^{x,y}(w)$ ;
- mean absolute percentage error between y-fitted from regression and SBS turnover,  $M_k^{\hat{y},y}(w)$ ;
- size and  $p$ -values of the regression coefficients.

#### 4. RESULTS

For each indicator, we divided the base cells into classes and computed the fraction of base cells with a VAT-regulation for each class. When we sort the classes by increasing value of the indicator  $R^2(w)$  we found that all base cells up to  $R^2(w) < 90\%$  did have a VAT-regulation. This means that base cells with a (relatively) poor correspondence between VAT and SBS turnover are base cells with a VAT regulation. Also, we found that among all base cells with a very good correspondence,  $R^2(w) > 99.5\%$ , still a fraction of 0.58–0.70 did have a VAT-regulation, which means that in some base cells VAT-regulations had only a minor effect on the correlation between VAT and SBS turnover. The indicator  $M_k^{\hat{y},y}(w)$  gave similar results as  $R^2(w)$ , but was more sensitive to outliers. Results from  $M_k^{x,y}(w)$  were less useful, since that indicator does not correct VAT for systematic differences from SBS turnover (by the linear regression). We found that we could use the indicators  $R^2(w)$  and  $M_k^{\hat{y},y}(w)$  to distinguish situation 4 from 2 and 3.

Surprisingly, the fraction of base cells with a regulation was neither related to the value of the regression coefficients nor to their  $p$ -values. For instance, both values for the slope  $\beta_k$  smaller than 0.95 or larger than 1.05 were found in base cells where no regulation is known. Furthermore, we found that base cells without a VAT-regulations could have values for  $\beta_k$  close to one but with a very small  $p$ -value ( $p \leq 0,005$ ) for a test of the null-hypotheses  $\beta_k = 1$ . We therefore concluded that the  $p$ -values are not useful indicators to separate situation 2 from 3, since that could lead to meaningless corrections.

Instead, we computed a ‘control range’ for the indicator values that contained 95% of the values of the empirical frequency distribution from base cells where no regulation was



found. Base cells with a VAT-regulation for which the indicators  $R^2(w)$  and  $M_k^{\hat{y},y}(w)$  are *within* the control range are ‘candidates’ for situation 2 and 3, base cells outside that range and with a regulation fall into situation 4. Subsequently, base cells with the regression coefficients outside the control range separate situation 3 from 2. Unfortunately, not for all base cells in situation 3 did the value of the slope and/or intercept correspond to the expectations based on the VAT-regulations as given in the last column of Table 1. For those base cells alternative models may be needed to study the relationship between VAT and SBS turnover, see section 5.

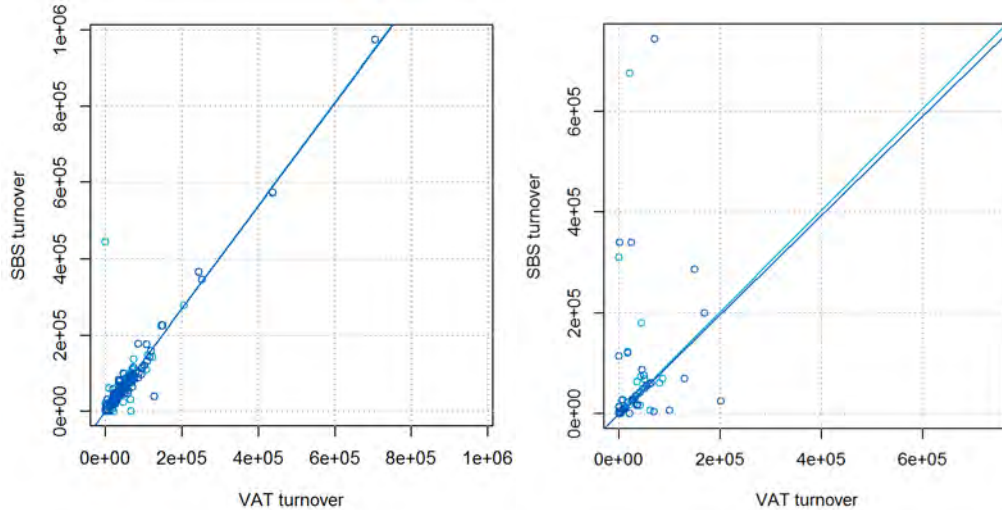


Figure 1. Example of a base cell in situation 3 (NACE 45.11.2 sale and repair of cars and light motor vehicles) and in situation 4 (NACE 50.200 sea and coastal freight water transport) (2009=light blue; 2010=dark blue).

## 5. CONCLUSIONS AND DISCUSSION

We presented a method to analyse the combination of numerical and conceptual correspondence in order to decide whether administrative data can be used to replace sample survey data, re-using already available data. We showed a first means to group the base cells into the four situations mentioned in section 3.2.

We found that the coefficients of linear regression could vary considerably even for base cells *without* a VAT-regulation. Of course, a possible explanation is that we overlooked some VAT-regulations. Nonetheless, it is known that both the administrative and survey sampling data are prone to errors [5], such as reporting errors and linkage errors. Those errors may affect the estimated relationship when they cause systematic effects. For base cells where the current approach is not sufficient to draw final conclusions, alternative models may be used, see [6], but these may require that additional data are collected.

## REFERENCES

- [1] Rich, S. and Burman, S. (2012). Use of VAT and VIES data for validation in International Trade in Goods and Services. Paper presented at the European Conference on Quality in Statistics (Q2012), 29 May – 1 June 2012 Athens, Greece.
- [2] Daas, P.J.H. and Ossen, S.J.L. (2011a). Report on methods preferred for the quality indicators of administrative data sources. Deliverable 4.2 of the BLUE Enterprise and Trade Statistics Project.
- [3] Vekeman, G. (2012). Confronting various administrative data sources to estimate employment variables. Deliverable for the ESSnet for Admin Data.
- [4] Van Delden, A and De Wolf, P.P. (2013). A production system for quarterly turnover levels and growth rates based on VAT data. Paper presented at the New Techniques and Technologies for Statistics (NTTS) conference, 5-7 March 2013, Brussels.

- [5] Groen, J.A. (2012). Sources of error in survey and administrative data: the importance of reporting procedures. *Journal of Official Statistics* 28, 173–198.
- [6] Bakker, B.F.M. (2012), Estimating the Validity of Administrative Variables. *Statistica Neerlandica* 66, 8–17.

# Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables

Sander Scholtus (s.scholtus@cbs.nl)<sup>1</sup>, Bart F.M. Bakker<sup>1,2</sup>, Arnout van Delden<sup>1</sup>

**Keywords:** administrative data, audit sample, structural equation model, VAT data

## 1. INTRODUCTION

Many NSIs are looking at ways to use administrative data to reduce – and ideally replace – their own data collection through surveys. Reasons for this include tighter budgets and a decreasing willingness of persons and businesses to participate in surveys. When examining the suitability of a given administrative source for statistical purposes, several questions need to be addressed [1]. In this paper, we will focus on issues related to measurement. In general, all data sources may contain errors. In the case of administrative data, a particular source of error arises from potential differences between the variable that is measured for administrative purposes and the variable that is needed for statistical purposes. As an example, the variable *turnover* as operationalised by the tax authorities may differ from *turnover* as defined in the short-term statistics (STS) regulation; e.g., some economic activities may be exempt from taxes. Assessing the measurement quality of administrative variables for statistical use is therefore important.

In the context of questionnaire design, there is a well-established tradition of using linear *structural equation models* (SEMs) to assess the measurement quality of survey variables [2]. Each observed variable is modelled as an imperfect measure of a latent (unobserved) variable. Repeated measurements are needed to identify such a model. The *validity* of an observed variable is defined as its correlation to the underlying latent variable. Applying this modelling approach to administrative data is not entirely straightforward. [3] suggested that repeated measurements may be obtained by linking an administrative data set to data from an independent sample survey. For this approach, one does not need to assume that either the administrative or the survey data are error-free.

In official statistics, population means or totals are often of interest. Therefore, it may be important to know whether any substantial measurement bias occurs in the levels of individual variables (*intercept bias*). This type of error is not captured by the above validity measure. In this paper, we show how the SEM approach may be extended to also evaluate bias. To illustrate, we discuss an application at Statistics Netherlands (SN) to assess the measurement quality of Value Added Tax (VAT) *turnover* for the Dutch STS.

## 2. METHOD: ESTIMATING VALIDITY AND INTERCEPT BIAS USING SEMS

In theory, an intercept bias can be assessed within the SEM framework by including an intercept term in each model equation. For our purposes, an SEM may be defined by:

$$\eta_i = \alpha_i + \sum_{j \neq i} \beta_{ij} \eta_j + \zeta_i, \quad (i = 1, \dots, m), \quad (1a)$$

$$y_k = \tau_k + \lambda_k \eta_{i_k} + \varepsilon_k, \quad (k = 1, \dots, n). \quad (1b)$$

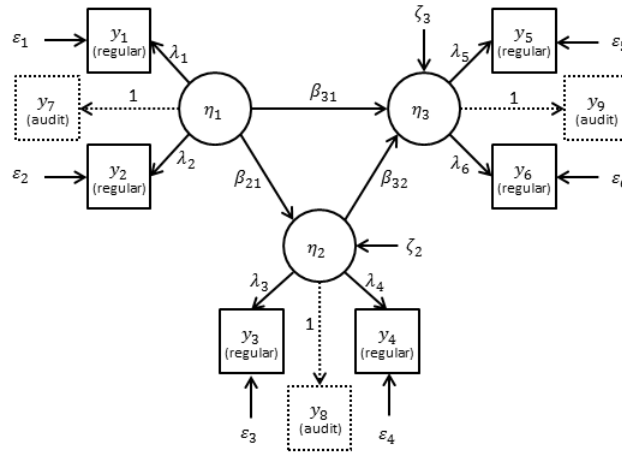
---

<sup>1</sup> Statistics Netherlands

<sup>2</sup> VU University Amsterdam

Here,  $\eta_1, \dots, \eta_m$  denote unobserved error-free variables of interest and  $y_1, \dots, y_n$  denote the corresponding observed variables that may be affected by measurement errors. Equation (1a) is a *structural equation* relating the unobserved variables to each other: the coefficient  $\beta_{ij}$  represents a direct effect of  $\eta_j$  on  $\eta_i$ ,  $\zeta_i$  represents a zero-mean disturbance term, and  $\alpha_i$  represents a structural intercept. Equation (1b) is a *measurement equation* relating an observed variable to an unobserved one in terms of a factor loading  $\lambda_k$ , a measurement intercept  $\tau_k$ , and a zero-mean measurement error  $\varepsilon_k$ . Note that we restrict attention here to SEMs in which each observed variable  $y_k$  loads on exactly one latent variable  $\eta_{i_k}$ . Provided that the model is identified, all parameters in system (1) can be estimated from the observed variance-covariance matrix and the observed vector of means. The absolute value of the standardised estimated factor loading  $\lambda_k$  may be used as a measure of the validity of  $y_k$  [3]. The intercept bias of  $y_k$  may be evaluated by comparing its observed mean  $E(y_k)$  to the estimated error-free mean  $E(\eta_{i_k})$ .

Identification of any SEM with latent variables requires that each latent variable be given a scale and, if the model contains intercept terms, that the origins of these scales be fixed as well. When one is interested only in estimating the validity, identification may be achieved by standardising the latent variables (mean 0, variance 1). However, this is not an option for evaluating the intercept bias. In fact, none of the standard identification procedures for SEMs are suitable in that context. A procedure for achieving meaningful model identification was suggested by [4]. The basic idea is to collect additional ‘gold standard’ data on each latent variable for a random subsample of the original data set. In practice, it is usually prohibitively expensive or otherwise inconvenient to collect ‘gold standard’ data for the entire population or even for a sizeable sample, but it may often be feasible to do this for a small subsample. Provided that this *audit sample* is chosen by random selection from the original data set, we can use it to assign a meaningful metric to the latent variables and identify the SEM. Figure 1 shows an example of a path diagram of such an SEM for the case of  $m = 3$  latent variables with two indicators (outside the audit sample). See [5] for details on how to estimate this model.



**Figure 1. Example of an SEM identified by means of an audit sample. Dashed lines indicate variables that are only observed in the audit sample.**

### 3. RESULTS

#### 3.1. Results on simulated data

To get an impression of the required size of an audit sample, some preliminary analyses were run on simulated data [5]. The simulations were based on the SEM of Figure 1. We

generated 100 random data sets of 1000 records each, by drawing from a multivariate normal distribution with the mean vector and variance-covariance matrix determined by a choice of SEM parameter values. All initial observed variables contained measurement error, including a substantial intercept bias. For a subsample of  $M$  records, we added indicators without measurement error, thus simulating an audit sample of size  $M$ .

**Table 1. Estimated latent means with model identification based on an audit sample; averages and standard deviations (in brackets) across 100 simulations.**

parameter	true value	estimates obtained with an audit sample of size $M$			
		$M = 25$	$M = 50$	$M = 100$	$M = 200$
$E(\eta_1)$	1.00	1.00 (0.06)	1.00 (0.05)	1.00 (0.04)	0.99 (0.04)
$E(\eta_2)$	2.40	2.40 (0.05)	2.40 (0.04)	2.40 (0.03)	2.39 (0.02)
$E(\eta_3)$	4.94	4.94 (0.06)	4.94 (0.05)	4.94 (0.04)	4.94 (0.03)

Table 1 shows selected results of this simulation study for various choices of  $M$ . These results suggest that an audit sample of  $M = 50$  or even  $M = 25$  units (5% and 2.5% of the original data set, respectively) leads to quite accurate estimates of the latent means. Indeed, investing in a larger audit sample may not be worthwhile, considering the marginal gain in accuracy. See [5] for more details and further results.

### 3.2. Application: Using VAT turnover for the Dutch quarterly STS<sup>3</sup>

Since 2011, SN publishes quarterly STS on *turnover* based on a combination of VAT data for small to medium-sized businesses and a census survey for the largest and/or most complex units [6]. The VAT data are obtained from tax declarations submitted to the Dutch tax authorities. The primary output of STS consists of estimated growth rates of *turnover* by publication cell (corresponding to sectors of the economy). Levels of total *turnover* are also estimated and used to calibrate the Dutch structural business statistics (SBS) and to weight the contribution of each sector to the national accounts. Given this secondary use of the STS estimates, it is vital that they do not suffer from intercept bias.

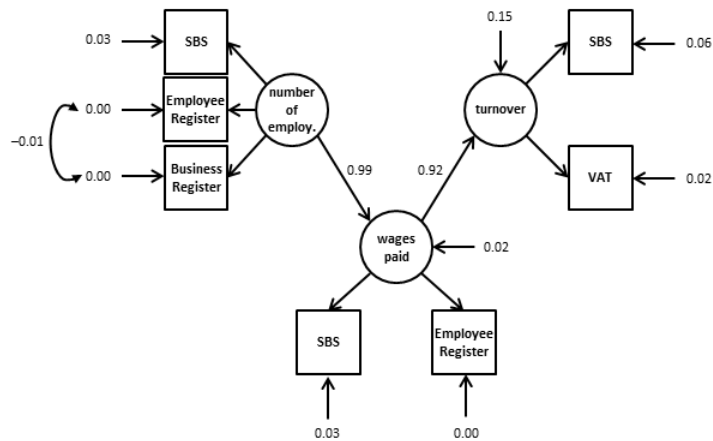
Previous analyses at SN of the measurement quality of VAT data were based on a linear regression of *turnover* as measured in the SBS survey on VAT *turnover* [7]. A drawback of this approach is that measurement errors in the SBS and VAT data are not explicitly taken into account. It is known that estimates of regression parameters may be biased in the presence of measurement errors. Therefore, we decided to do an alternative analysis using an SEM to account for potential measurement errors. We selected a number of publication cells for which the previous analysis was inconclusive. In this abstract, we focus on one such publication cell: “45.11.2 Sale and repair of passenger cars and light motor vehicles (no import of new cars)”. Besides *turnover*, we included *number of employees* and *wages paid* in the model. As *wages paid* is not relevant to businesses registered as natural persons with less than two employees, they were excluded from the present analysis. We are currently looking at ways to extend the analysis to these units.

For all concepts, we obtained one indicator from the SBS sample survey data and a second indicator from an administrative source. For *turnover*, the VAT data were used. Administrative values of *wages paid* were obtained from the Employee Register maintained by the Dutch Employee Insurance Agency. Finally, administrative values of *number of employees* could be obtained either from the General Business Register

<sup>3</sup> The presentation at the conference will contain more results of this application, including results on bias.

maintained by SN or from the Employee Register. We tested models containing either or both of these as administrative indicators for *number of employees*. In fact, the *number of employees* in the Business Register is derived from a variety of sources including the Employee Register. Therefore, when both indicators were included in the model, we also tested whether their measurement errors were correlated.

Figure 2 shows the final model with its standardised parameter estimates. The absolute standardised factor loadings may be computed as  $|\lambda_k| = \sqrt{1 - \text{var}(\varepsilon_k)}$ . It is seen that all observed variables have excellent validity ( $|\lambda| > 0.95$ ). The administrative data appear to have a slightly higher validity than the survey data. As far as the validity is concerned, VAT *turnover* could be used as a replacement for survey data in this publication cell.



**Figure 2. The final fitted model for publication cell 45.11.2.**

#### 4. CONCLUSIONS

In this paper, we discussed the possibility of using structural equation modelling to assess the measurement quality of administrative variables for statistical use. We specifically looked at validity and intercept bias. Estimating the intercept bias of an observed variable in a meaningful way requires the collection of additional ‘gold standard’ data for a random subsample of the original data. Results on simulated data suggest that the size of this audit sample may be small. Moreover, [5] discusses how the information collected from an audit sample may be re-used during multiple rounds of a repeated survey.

#### REFERENCES

- [1] Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica* **66**, 41–63.
- [2] Scherpenzeel, A.C. and Saris, W.E. (1997), The validity and reliability of survey questions: a meta-analysis of MTMM studies, *Sociological Methods & Research* **25**, 341–383.
- [3] Bakker, B. (2012), Estimating the validity of administrative variables, *Statistica Neerlandica* **66**, 8–17.
- [4] Sobel, M.E. and Arminger, G. (1986), Platonic and operational true scores in covariance structure analysis, *Sociological Methods & Research* **15**, 44–58.
- [5] Scholtus, S. (2014), Explicit and implicit calibration of covariance and mean structures. Discussion paper 2014-09, Statistics Netherlands.
- [6] Van Delden, A. and De Wolf, P.-P. (2013), A production system for quarterly turnover levels and growth rates based on VAT data. Paper presented at the NTTS-2013 conference, Brussels.

- [7] Van Delden, A., Banning, R., De Boer, A., and Pannekoek, J. (2015), Analysing whether sample survey data can be replaced by administrative data. Paper to be presented at the NTTS-2015 conference, Brussels.

# Measuring the quality of multisource statistics

Sorina Văju ([Sorina-carmen.vaju@ec.europa.eu](mailto: Sorina-carmen.vaju@ec.europa.eu))<sup>1</sup>, Mihaela Agafitei<sup>1</sup>, Fabrice Gras<sup>1</sup>, Wim Kloek<sup>1</sup>,  
Fernando Reis

**Keywords:** quality, multiple sources, integration, administrative sources, big data.

## 1. INTRODUCTION

Most EU Member States have been moving towards an increased use of administrative data sources for statistical purposes, as a substitution and/or as a complement to survey data. At the same time, the emergence of big data allows for a further increase of available sources for statistics. Statisticians are looking for new ways to combine sources and methods in order to accommodate new demands for statistics. As a result, statistical output is based on complex combinations of sources. Its quality depends on the quality of the primary sources and the ways they are combined. This paper analyses the appropriateness of the current set of quality measures for multiple source statistics, explains the need for improvement and outlines directions for further work.

## 2. QUALITY OF STATISTICS IN A MULTISOURCE ENVIRONMENT—GENERAL DISCUSSION

The ESS quality framework identifies five quality dimensions to describe output quality: (a) relevance (European Statistics meet the needs of users), (b) accuracy and reliability (statistical outputs accurately and reliably portray reality), (c) timeliness and punctuality (statistical outputs is released in a timely and punctual manner) (d) coherence (statistical outputs are consistent internally, over time and comparable between regions and countries) and comparability (it is possible to combine and make joint use of related data from different sources), (e) accessibility and clarity (statistical outputs are presented in a clear and understandable form, available and accessible on an impartial basis with supporting metadata and guidance) [1].

Some quality dimensions – relevance, accessibility and clarity – are not impacted by integrating multiple sources while others – timeliness and punctuality – may be impacted but the way we measure them is still appropriate. However, measuring other dimensions – accuracy and reliability, coherence and comparability – require incorporating the effect of sources and integration approach. More specifically, the first two groups describe the statistical product irrespective of the statistical process behind it (the choice of data sources to use, the statistical processing and the integration approach). The last group highly depends on the quality of sources and of the way they are combined, as it focuses on measuring the deviations from reality and on indicating the correct use of the statistical product. At each step of the production process [2], accuracy and comparability appear as the quality dimensions that are actually at stake.

Even if some accuracy measures (e.g. coverage rate, edit failure rate, imputation rate, average size of revisions, etc.) can apply to the several types of data sources it is very difficult to assess the sensitivity of the final statistical output to source specific errors and to the methods used to integrate them. Consequently, the accuracy and reliability dimension needs to be reconsidered in order to cover all methodological aspects and implications given by the combination of sources and methods. International comparability can be seriously affected when integrated statistics include administrative data coming from different national administrative systems and produced using different

---

<sup>1</sup> Eurostat, European Commission.



methodological approaches/combinations. At European level, this translates into a huge number of possible sources of lack of comparability, given by combinations of: (i) national legal and institutional environments, (ii) acceptable trade-off between quality dimensions at national level; (iii) appropriate trade-off between costs and benefits in terms of output data quality at national level, (iv) methodological choices to integrate the several data sources.

### 3. METHODS FOR QUALITY ASSESSMENT

There are three facets for which quality can be checked: input, process and output. Input assessment refers to the quality of raw data and should allow statisticians to decide whether and how a given data source – including big data and administrative sources – can be used on a regular basis to produce statistics. Process quality refers to intermediate steps; it describes or quantifies the transformations that the raw data has undergone through the statistical process (e.g. imputation, editing). Output quality refers to the final statistical product and it should provide to the user easy to understand information on the quality of the final data.

#### 3.1. Output quality assessment on the basis of input and process

The natural approach for identifying the possible impact on quality of combining several types of data sources in the statistical production process is to look at each step of the production process and assess the impact of such integration. The use of combined sources mainly impacts the way the accuracy measurement is made. The assessment of the other quality dimensions does not specifically depend on using combined sources, with the exception of the comparability dimension. Nevertheless, comparability assessment can be to a large extent reduced to structural error generated by the introduction of some possible statistical biases. This does not affect comparability over time, for which the break in time series and the outliers are the main threats. Possible outliers/breaks can be detected based upon existing methods; this will, as illustrated later, provide some first insights on how to assess the quality of data derived from multiple sources.

Table 1 gives an overview, for several statistical production activities, of the link between the risks of combining multiple data sources and the corresponding impacted quality dimensions and quality measurement. Accuracy assessment of the combination of sources should most likely focus on aggregating random mechanisms effects with the bias effects introduced by non-survey data. However, when using multiple sources, measuring final data accuracy via assessment of the data integration in the several statistical production activities appears not straightforward and even too complicated to be envisaged.

**Table 1. Risk and impacted quality dimension when combining multiple sources**

Statistical production activities	Risk	Impacted quality dimension	Error measurement
Linkage and determination of the target population	Missed link, wrong link: under/over coverage	Accuracy, comparability	Bias, confidence range of the target population
Concept/definition	Aggregation of different concept/definitions	Relevance, accuracy, comparability	Bias, Variance error, qualitative assessment

Imputation/estimation	Estimation error	Accuracy	Bias, variance error
Classification	Wrong classification	Relevance, accuracy, comparability below a certain level of aggregation	Bias, variance error

### 3.2. Direct output quality assessment

In this section we discuss possibilities to assess accuracy and comparability of statistical outputs without analysing the processes behind it. There are three options: direct assessment of the output quality on the basis of the output itself, assessment on the basis of a common reference source and methods involving mainly bootstrapping techniques.

There are several ways to assess the output quality on the basis of the output itself. Breaks in series are a direct indication of bias and show the impact of changes in sources and methods. The impact can be measured by keeping for a while a double production system or by extrapolation. Bias can be indicated by systematic corrections when doing revision of data when more information becomes available. In case the revisions show systematic corrections, this would be an indication of bias. Another example is applying outlier detection techniques to cross-sectional data. The main advantages of the direct assessment methods using solely the output are: (i) they require no knowledge about the sources and methods used in the statistical production process and (ii) they are fairly easy to implement. The major disadvantages are: (i) it is not always possible to distinguish between real differences, bias and variation and (ii) the method offers no clue on diagnosis and remedy.

As regards the assessment of output quality with a common reference data source, two cases are distinguished: the quality survey and any other reference source. The advantages of the quality surveys are: (i) the quality survey has a known variance and it is designed to have a low bias; (ii) it can have diagnostic value by identifying the weaknesses of the process steps; (iii) it is easy to summarise into an overall assessment. In practice costs and other practical considerations will probably prevent its full scale application. At a less ambitious scale it might be possible to assess specific elements where other information is lacking (e.g. under-coverage). Other reference sources might be other related statistics, administrative sources or big data sources with considerable conceptual harmonisation. The advantages are: (i) low additional costs and no additional burden; (ii) the separate production process. The main disadvantages are: (i) for an administrative reference source, or in the case of big data sources, we have no control over variance and bias and thus it will often require an assumption on the level of variation and on the stability/equal distribution of bias; (ii) usually it has no diagnostic value; (iii) the natural tendency to incorporate good sources into the production process, thus making them unavailable as independent reference source.

The ESSnet AdminData [3] proposed ways to adapt the bootstrap re-sampling methods in order to estimate the root mean square error (RMSE) that includes both sampling variance and bias due to non-sampling errors, incorporating thus the effect of interaction with administrative data. The reasoning behind is that bootstrap methods enable inserting randomness through the replication of samples [4]. Thus, replications of combined dataset are produced, either by simulating the distribution followed by the data or by

using existing samples for replication. The purpose is to simulate and/or replicate random behaviour of administrative data by undertaking statistical inference on administrative data. These methods can equally be applied to big data sources.

**Table 2. Bootstrap methods use by type of combination of administrative data with other sources**

Possible use	Remarks	Main practical problem
Replacement for primary and/or complementary data	Overlapping survey data can significantly increase the feasibility and relevance of the method	Inference on the distribution and/or generating process of the administrative data. Detection of break and outliers in time series.
Partial use for sample design or input for statistical registers	Uncertainty can be inserted by estimating false positive and negative probability	How to simulate the addition of a previously non selected unit in the replication of the sample
Additional variables for estimation; auxiliary information to support processing of primary data (editing, imputation, calibration)	Modelling on how random is channelled through the production process requires a good description of the production process	Simulation of the error caused by the imputation/estimation methods.

#### 4. CONCLUSIONS

Measuring output quality through input and process quality gets too complex in processes combining several sources, especially at the European level. Therefore, alternative solutions should be found. The paper lists three alternative approaches that do not depend on the design of statistical process: (a) direct output assessment; (b) a common reference source; (c) bootstrapping.

Information on output quality has internal use for monitoring and improving the statistical production process. The information also has an external role. The quality information should be summarised in such a way that data users can assess the accuracy and comparability. The alternative approaches contribute to the quality assessment for internal purposes, but a coherent external summary of information remains difficult. Assessing quality is not for free. Knowledge on quality is also required to allocate scarce resources between improving quality and measuring quality.

#### REFERENCES

- [1] Eurostat ESS Standard for Quality Reports, (2009).
- [2] ESSnet Data Integration Report on Work Package 1: State of the art on statistical methodologies for data integration (2011).
- [3] ESSnet Use of Administrative and Accounts Data in Business Statistics Deliverable 2.4: Guide to checking usefulness and quality of admin data (2013).
- [4] L. Kuijvenhoven, L. and S. Scholtus, Bootstrapping combined estimator based on register and sample survey data, Discussion paper 201123 of Statistics Netherlands, (2011).

# AN ONTOLOGY-BASED APPROACH TO ADMINISTRATIVE DATA SOURCES' DOCUMENTATION AND QUALITY EVALUATION

Giovanna D'Angiolini ([dangioli@istat.it](mailto:dangioli@istat.it)), Pierina De Salvo, Andrea Passacantilli, Edoardo Patruno, Teresa Saccoccio

**Keywords:** administrative data source, administrative data documentation, administrative data quality, ontology

## 1. INTRODUCTION

A large number of NSIs usually exploits administrative data for statistical purposes, in order to improve the quality of statistical outputs, to reduce the statistical burden on respondents and to minimize costs [1] [2]. Moreover the official statistical data production is not the only context in which administrative data are used for statistical purposes: in recent years more and more non statistical organizations have been implementing their own decision support systems for monitoring the context and the effects of the organization's strategy. Such systems exploit the organization's administrative data sources together with other data sources for drawing inferences about those phenomena which are involved in the organization's activities. This is a kind of statistical exploitation of administrative data even if the purpose is to satisfy the organization's own knowledge requirements, while NSIs produce statistical data for the public.

Any statistical usage of an administrative data source implies a source evaluation activity, whose goal is to ascertain if the source can be used for studying the phenomenon of interest. Such an evaluation activity includes two main distinguished phases:

- evaluating if the collective and the variables of interest can be derived from the observed administrative collectives and characteristics
- evaluating if the administrative data concerning the collective and the variables of interest exhibit a good quality from a statistical viewpoint, that is, they can be used as dependable measures of the underlying phenomenon.

Often the statistical users of the available administrative data sources perform such an evaluation activity from their particular viewpoint without any reference to standard procedures and shared methodological and documentation tools.

As to the first evaluation phase, in many situations the administrative data users are compelled to analyze the whole source's information content for determining if the source satisfies their particular information requirements. Moreover they apply different approaches and models and therefore they cannot share the produced documentation. This is a serious drawback: in fact analyzing the information content of an administrative data source may require advanced conceptual modelling competencies, due to the complexity of many administrative data sources' observed part of the real world. The statisticians would take advantage of the availability of standard and shared documentation of the administrative data sources' information content.

Similar remarks apply to the administrative data sources' quality evaluation. The main Frameworks of quality indicators for administrative data sources propose sets of very general indicators which aim at documenting the overall quality of the analyzed data sources, and are not well-suited to leading the quality evaluators in assessing the data source's quality with reference to their particular collectives and

variables of interest. In concrete situations the administrative data source's users perform more specific quality evaluation activities but they cannot share the results of such evaluation activities, because they generally apply empirical and not repeatable procedures.

In such a scenario the NSIs are required to play a new important role. They must devise and release guidelines, standard methods and tools for supporting any kind of user which need to evaluate the administrative data sources' information content and quality in order to exploit administrative data for acquiring knowledge about real world phenomena.

## **2. A STRATEGY FOR SUPPORTING THE STATISTICAL USERS OF ADMINISTRATIVE DATA SOURCES**

Our general strategy is aimed at:

- collecting information about the available administrative data sources, producing standard documentation about their information content and quality and disseminating such information to all the potential statistical users
- modifying, when possible, the content of the available administrative data sources through adopting standard statistical definitions, classifications and data management conventions.

It is implemented through several systematic documentation activities, which concern different kinds of administrative data sources:

- organizing surveys of the administrative data sources owned by local administration institutions
- establishing a procedure for allowing both central and local administration institutions to inform the NSI about any planned change in their managed administrative data sources, so as to allow it to give feedback and propose directions on the designed changes
- for the most important administrative data sources, performing a dedicated inquiry in order to collect comparable information about their information content and their quality, by means of interviewing experts
- for some selected administrative data sources, releasing in-depth quality documentation by means of a set of standard quality indicators.

In order to support such activities we have devised and implemented these information managing and methodological tools.

- a web-based metadata management system called DARCAP (Documenting Public Administration Archives) which:
  - provides the administrative data sources' owners with functionalities for informing Istat each time they plan a change
  - provides the NSI experts with several functionalities for documenting the administrative data sources' information content and quality
  - provides the statisticians with on-line public information about the available administrative data sources' information content and quality
- a Framework of quality indicators for administrative data sources which encompasses qualitative as well as quantitative indicators in order to support in-depth quality analyses.

### **3. AN ONTOLOGY-BASED APPROACH**

In our approach, the administrative data source's information content is documented by means of defining the data source's own ontology. An ontology of an administrative data source is a structured description of its information content, based on a standard conceptual model.

Our Framework of quality indicators for administrative data sources organizes the quality indicators according to the well-established quality model which has been proposed by Statistics Netherlands in 2009 [3]. However its distinguished feature is that it specifies a detailed set of indicators which are defined on the basis of the data source's ontology specification, particularly the quantitative indicators in the Data hyperdimension.

Our approach is innovative because the ontology-based description of the content of the available data sources is not yet a familiar documentation method among statisticians, despite the fact that today the ontology-based data documentation is a widespread practice in the database management field.

By means of anchoring the proposed indicators to the data sources' ontology we ensure a systematic specification of the indicators and we provide the quality evaluators with directions for choosing among indicators as well as for interpreting the calculated indicators, according to their particular requirements. Given that various distinguished factors influence the administrative data sources' quality, this is the only way for leading the quality evaluators in producing a standard assessment of the administrative data source's quality by applying standard and repeatable procedures.

### **4. OUR ADOPTED CONCEPTUAL MODEL**

In order to define our conceptual model, we have analyzed the life-cycle of the administrative data and singled out the different kinds of real-world objects to which they are referred, and we have put such objects into correspondence with those objects to which any statistic is currently referred, namely collectives and variables. Our conceptual model is oriented towards supporting the statistical exploitation of the administrative data sources, but it can be easily translated into other general-purpose conceptual models and languages for ontology specification. In the following we briefly introduce its main features.

Administrative data sources collect information about several kinds of real world objects in order to support administrative activities. First, any administrative activity entails collecting data about those entities which the activity addresses. Such entities are subsets of the two general populations of persons, on one side, and entities which perform economic activities, on the other side, or they are subsets of related populations such as households, territorial units. Moreover, information is collected about those particular sets of events which may involve these entities and are of interest for the purposes of the administrative activity. The observed populations and sets of events are linked by relationships. For both observed populations and sets of events proper information is collected about their characteristics, which may change in time. As an example, the Ministry for Public Education continuously collects information about the students, the schools and the universities with their characteristics as well as about sets of events such as the degree course enrolments, the examinations, the degree earnings with their characteristics.

Therefore inside an administrative data source's version we find two kind of linked collectives: populations and set of events. Populations are subsets of the two most general populations of persons on one side, and entities which perform economic activities on the other side, or subsets of their related

populations. Sets of events can be instantaneous (such as examination) or durable (such as degree course enrolment) and they may connect elements belonging to different populations, as an example any degree course enrolment event connects a student with a degree course. Each element of these collectives has qualitative or quantitative characteristics, such as date of birth, residence, date of the enrolment, examination score, as well as relationships with elements in other collectives.

According to a widespread ontology specification paradigm, in our conceptual model a qualitative or quantitative characteristic is regarded as a relation which links an element belonging to a collective with an item belonging to a proper classification, or with a number in a numerical domain respectively. From a statistical viewpoint, the quantitative characteristics and the qualitative characteristics together with their associated classifications are regarded as variables.

The administrative data source's information content documentation is produced according to such a conceptual model and stored into the DARCAP system.

The analysis of the administrative data source's quality is based on the data source's ontology specification. For each object in an ontology, namely a collective, a characteristic, or a relationship we can build belonging statements concerning observed elements, more precisely we can assert that a single observed element belongs to a set or that a couple of elements belongs to a characteristics or a relationship. In logical terms, the statements which concern populations and sets of events correspond to a single variable predicate, the statements which concern characteristics and relationships correspond to two variable predicates. Such statements will be true or false. The administrative data sources continuously collect and store data which are in fact proper combinations of such belonging statements. Each administrative data source has its own data collection procedure, which consists in accepting or not accepting such combinations of belonging statements inside the data source. As a result, at any time any administrative data source stores a collection of belonging statements for each collective, characteristic or relationship in its ontology. It may happen that some of these statements are false, and that some true statements are not in the data set. Our proposed indicators are based on singling out all the possible kinds of false statement which can be accepted by an administrative data source.

## 5. CONCLUSIONS

Our experience is highlighting the various sources of complexity which make the specification of the administrative data sources' ontology and the analysis of the administrative data sources' quality two difficult activities. In the future, the standardization of such activities will provide the basis for automating particular evaluation tasks such as the reasoning on the derivability of new information from the existing administrative data sources and the comparison of quality indicators.

## REFERENCES

- [1] G.J. Brackstone, Issues in the use of administrative records for statistical purposes, Survey methodology (1987)
- [2] United Nations Economic Commission for Europe (UNECE), Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices, United Nations Publication (2011)
- [3] R. Vis-Visschers, J. Arends-Tóth, Checklist for the Quality evaluation of Administrative Data Sources, Discussion paper by Statistics Netherlands (2009) [5] G. D'Angiolini, P. , De Salvo, A.

# Profiling Big Data sources to assess their selectivity

Piet Daas ([pjh.daas@cbs.nl](mailto:pjh.daas@cbs.nl))<sup>1</sup>, Joep Burger ([j.burger@cbs.nl](mailto:j.burger@cbs.nl))<sup>1</sup>

**Keywords:** Big Data, profiling, social media, feature extraction, selectivity.

## 1. INTRODUCTION

In our modern world more and more data are being created and remain to be stored. These kinds of data, generally referred to as Big Data, are very interesting sources of information. They, for instance, may reflect traces of human or economic activity and could possibly be used for official statistics [1]. However, extracting information from Big Data for such purposes is challenging for a number of reasons. First, not all data are relevant for the research question at hand, which requires one to find the signal in the noise [2]. Second, most Big Data available are composed of events [3] and usually provide very little or no information on the unit that generated the data. Third, if information is available on the creator of the data it may not be easily linked to a specific person or company. Fourth, not all units in the target population that the researcher envisaged may be included in Big Data and the ones that are included are not a random sample from the target population. All in all, these issues make it challenging, to say the least, to use Big Data for the creation of official statistics. In this paper we will mainly focus on the fourth challenge, i.e. how to assess the selectivity of a Big Data source.

### 1.1. Social media as an example

Let's illustrate the research question at hand with an example. Many people in the Netherlands are active on social media: 70% of the population posted messages according to a recent European study [4]. Compared with a probability sample this is an extremely high coverage rate. In contrast to a probability sample, however, we do not know to what extent these social media accounts represent our target population. Also quite a number of social media accounts actually reflect the activity of companies (even though they are created by humans). These are different target populations: for persons the target population are the persons included in the population register of the Netherlands, while for companies these are the units in the Dutch statistical business register. Also, hypothetically, cyber savvy and extravert people are more likely to be active on social media than computer novices and introverts. In addition, not all activity is publicly available as some social media messages are private only. However, these are not uncommon issues as selective non-response in sample surveys also causes a deviation from representativeness. Without correction for selectivity, estimates will be biased.

## 2. METHODS

A common method used to assess selectivity in sample surveys is by comparing the distribution of relevant background characteristics in the data source with their known distribution in the target population. In principle, the same approach could be applied to Big Data, although in practice this is not trivial [5]. In an ideal world, units are linked to a population register containing background characteristics. Our experiences on studying Big Data sources have revealed that many units hardly provide any information that could be used to deterministically link them to a population register. In a more realistic Big Data context, background characteristics will have to be derived from the Big Data source. The key question is how should this be done? We think that an approach called

---

<sup>1</sup> Statistics Netherlands, Heerlen, the Netherlands.



‘profiling’ is the best option. Here the term profiling refers to an approach from the field of information science. In this approach, large amounts of data are analysed with the aim of discovering patterns to discern groups of similar units [6]. In this abstract, we will discuss the topic of profiling units from a social media perspective, since social media are quite challenging in this respect and a concrete example helps to illustrate the challenges more clearly. Social media also have the advantage from an experimental point of view, since a lot of data is publically available and each unit in the population has a unique identifier: a user id. This in contrast to many other Big Data sources, of which the data may be owned by private companies or that may have computers or other electronic devices as units [1].

### **3. RESULTS**

From an earlier study performed at Statistics Netherlands in cooperation with Erasmus University [7] we obtained a list of 380 thousand Twitter usernames which were—at that point in time and according to the location information on their user profile—all identified as Dutch Twitter users. This list is the starting point for the studies described below. Based on this list we will describe ways that could be used to profile Dutch Twitter users. Some of the methods have already been tested and some have not. For these studies only data available on public Twitter user accounts are used, meaning that anyone with a PC, a browser and an internet connection could access the data we studied. Since we respect the privacy of the users, no information on or examples of individual users will be provided.

#### **3.1. Background characteristics**

Many surveys use a similar set of background characteristics that correlate with target variables. In social statistics commonly used variables are: gender, age, income, education, origin, urbanicity, and household composition. For companies often used characteristics are: number of employees (size class), turnover, type of economic activity, and legal form. Because both persons and companies are active on social media, the first distinction that needs to be made is if the owner of an account represents a person or a company. Note that company is used here as a general term to describe both businesses and other public organizations. The distinction between persons and companies could be made by studying the username, user profile, link to the (company’s) website on the profile, the profile photo, the tweet frequency and the tweet content. As it is to be expected that companies will preferably be active on social media with a highly similar or identical name, a list of Dutch company names could be used to quickly identify them. It is to be expected that accounts of self-employed people and small-companies will be more difficult to distinguish from non-commercial personal accounts.

If an account is used by a private person, how could one determine its gender? This can be obtained from the username, the profile (bio-)information, the profile picture, an associated account (such as an account on LinkedIn) or the combined set of public tweets of that person. Combing the information will likely increase the chance of success. The same holds for age, where some first names occur more frequently for persons born in a particular period. The types of words and abbreviations used in tweets also provide clues on the age of the person writing them. Urbanicity could be derived from the location information in the user profile or from the location related content in the messages written. Level of education could be obtained from the job description (if available), the content of the messages related to this topic or from an associated LinkedIn account. The latter may also provide clues on income as does the location information in the user

profile. Origin and household composition might additionally be derived by analysing the social network of a user.

For companies a lot of the required information could be obtained via the associated website. It is to be expected that companies provide such a URL in their official Twitter profile. On this website documents, such as company publications and links to annual reports, might be available from which more detailed information, such as turnover and number of employees, could be obtained. However, some companies might not be active on social media or someone could have created a ‘fake’ company account.

#### **4. CONCLUSIONS**

To check the viability of the above mentioned approaches, we start by manually profiling a sample of the Twitter database. This provides a dataset containing both input features (e.g. username, profile image, tweets) and outcome measurements (e.g. gender, age). The dataset is subsequently split (horizontally) into a training set and a test set. We can use these to train and test various approaches developed in the field of information sciences, such as artificial intelligence and machine learning methods. Here, the training set is used to train the algorithm the relation between input and output. The test set is used to compare predicted output with observed (manually profiled) output and to optimize the parameters of the algorithm. With these techniques one attempts to predict the background characteristics of the rest of the Twitter database using the input features. To enable the use of Big Data in official statistics it is important that successful profiling approaches are being developed.

#### **REFERENCES**

- [1] M. Glasson, J. Trepanier, V. Patruno, P. Daas, M. Skaliotis, A. Khan, What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, (2013).
- [2] N. Silver, The Signal and the Noise: Why So Many Predictions Fail—but Some Don't, Penguin Group, New York, USA, (2012).
- [3] P.J.H. Daas, M.J.H. Puts, B. Buelens, P.A.M. van den Hurk, Big Data and Official Statistics. Journal of Official Statistics, NTTS special issue, (2014), accepted for publication.
- [4] Eurostat, Internet access and use in 2012. Eurostat newsrelease, (2013), Located at: [http://epp.eurostat.ec.europa.eu/cache/ITY\\_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-18122012-AP/EN/4-18122012-AP-EN.PDF)
- [5] B. Buelens, P. Daas, J. Burger, M. Puts, J. van den Brakel, Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands, (2014).
- [6] M. Hildebrandt, S. Gutwirth, Profiling the European Citizen. Cross Disciplinary Perspectives. Springer, Dordrecht, (2013).
- [7] P.J.H. Daas, M. Roos, M. van de Ven, J. Neroni, Twitter as a potential data source for statistics. Discussion paper 201221, Statistics Netherlands, The Hague/Heerlen, The Netherlands, (2012).

# ANALYSIS OF THE POTENTIAL OF SELECTED BIG DATA REPOSITORIES AS DATA SOURCES FOR OFFICIAL STATISTICS

Michalis Petrakos (Michalis.Petrakos@agilis-sa.gr)<sup>1</sup>, Anais Santourian (Anais.Santourian@agilis-sa.gr)<sup>1</sup>, Gregory Farmakis<sup>1</sup>, Photis Stavropoulos<sup>1</sup>, Georgia Oikonomopoulou<sup>1</sup>, Eleni Ntakou<sup>1</sup>, Alexandra Trampeli<sup>1</sup> and Marina Koumaki<sup>1</sup>

**Keywords:** Big data, Official Statistics

## INTRODUCTION

The aim of this paper is an initial assessment of the feasibility of employing novel methodologies for producing high quality Official Statistics based on big data. Official Statistics are published by government agencies or other public bodies and are based mainly on survey, census, or administrative data that are carefully collected, processed and disseminated. On the other hand, big data challenge the way we think about data assets, the sources we collect them from and the way we analyse them. Paradigms are shifting and a reverse approach of designing statistics is being applied. With big data, it is fundamental to explore the vast amounts of data available first and then to decide on the quantities to be measured. Thus, inference techniques used for Official Statistics will need to make a shift too.

This paper examines the potential of using big data sources as input for Official Statistics. Five sources are examined: (a) the Automatic Identification System (AIS) records of vessel movement, (b) Internet classified advertisement sites concerning house sales and rentals, (c) Social network content, such as Facebook and Twitter (d) Credit card transaction data and (e) an open government portal publishing all government expenditure decisions. The work was carried out under the framework of a European Commission / Eurostat project.

## 1 METHODOLOGY

### 1.1 Criteria for selection of the sources

The five specific sources were selected so as to represent a representative range of situations across a number of dimensions.

Firstly, we created an inventory of data repositories among which we identified the most suitable cases in terms of broad quality criteria such as the cost, sustainability, accuracy, relevance, consistency, interpretability and timeliness of potential statistical outputs of these sources [1][2]. Another selection criterion was the policy interest concerning each domain. The second step was a thorough analysis of the factors affecting the suitability of the big data repositories for the production of Official Statistics. Based on the information collected in the inventory as well as following the European Statistical System's definition of quality<sup>2</sup> we evaluated the appropriateness of each data repository for use.

---

<sup>1</sup> Agilis SA Statistics and Informatics

<sup>2</sup> The quality dimensions according to the European Statistical System's definition are: a) relevance, b) accuracy and reliability, c) timeliness and punctuality, d) accessibility and clarity and e) coherence and comparability. The cost involved in the production process is also discussed.

## 1.2 Case studies

After the final selection of the five data repositories most relevant for use, we elaborated a scenario for the potential exploitation of the identified data. The steps that were followed are: a study of the background of the data source; a comparative analysis between the identified big data repositories and relevant Eurostat data collections; the examination of the correspondence between the datasets available in the selected repositories with datasets produced and disseminated by Eurostat in the relevant statistical domains; and finally the mapping between the variables already collected and/or disseminated by Eurostat with variables in the repositories.

Moreover, we examined their potential to fulfil existing needs for statistical information by either supplementing or completely replacing Official Statistical indicators produced by Eurostat. Finally, we examined whether the proposed big data-based indicators were feasible from a methodological and practical point of view. The potential of the five use cases, related to statistical domains, was assessed across the technical, organizational, methodological, cost-benefit, legal and socio-political dimensions.

## RESULTS

The cases presented in this paper, demonstrate that, in principle, big data can be used in a supplementary or complementary way for Official Statistics. Indicatively:

- AIS data can be used as a source for maritime transport emissions statistics or maritime transport statistics in general
- Real estate classified advertisements can be used as input for house price indices
- Social network content can be the basis for statistics about well-being
- VISA transaction data can be used for consumer expenditure statistics
- Government spending data are useful for national accounting purposes

As shown in the case of AIS data and real estate classified advertisements, big data offer a very fine geographical resolution and provide data about a far larger number of units than what any sample survey could offer. In terms of timeliness and frequency, statistics based on big data (e.g. VISA transaction based statistics) can supplement Official Statistics of very low frequency. Big data that are generated without the intervention of human reporting (e.g. AIS messages or VISA transaction data) reduce the burden imposed on individuals and enterprises. Finally, if a big data source has geographical coverage greater than a single country (e.g. the AIS messages have global coverage) this means that geographical comparability will be higher than that of survey or administrative data for the same countries.

On the other hand, big data sources are often implementing different concepts than those required by corresponding Official Statistics. For example, the real estate classified advertisements contain data on asking price but not on the final price at which each property is sold or let. Some adjustment is therefore needed. Moreover, the coverage of the intended target population may not be the desired one. For example, AIS messages cover vessels larger than 300 gross tons and only a voluntary subset of the smaller ones; expenditure data in Greek government's open portal omit some sensitive expenditure items. Therefore, either the target population of the statistics must be modified or the big data need adjustment or combination with additional sources. Some big data sets represent self-selected samples. For example, not all individuals have Facebook accounts, those that do have arguably choose what they want to post on them and moreover probably make public only a subset of it. Therefore, regular statistical

inference may not be correct without modifications, which is a research topic at present. Finally, access to big data sources may be very difficult. Some of them may be confidential (e.g. credit card transactions) and others may only be available via private intermediaries (e.g. Facebook status updates) who may charge for access. Sometimes, the cost of access to the data combined with cost for processing them may in fact offset the gains from not having to run a sample survey, therefore, cost efficiency should be established.

## **CONCLUSIONS AND FUTURE WORK**

The evaluation of the effect from the use of particular big data sets on the overall quality of the statistics produced reveals several potential benefits as well as disadvantages. Among all, big data are characterized by the (large) amount of information and the (high) frequency at which they are produced. Good geographical coverage and the accuracy of the data, are some of the pros. On the other hand, a major issue concerning big data is their availability due to privacy and confidentiality protection or other restrictions. In addition, the very high data volume further increases the processing needs. It is highlighted, that the quantities produced can be used as supplements to the respective Official Statistical indicators, while in most cases, it is not feasible to completely replace them. However, exploiting the vast amounts of data available in a methodologically sound way may enhance the timely production of low cost and high quality Official Statistics. That is, big data has great potential which, in order to be unleashed requires new, appropriately tailored methods to accelerate the analysis of large amounts of data. Moreover, the use of big data in Official Statistics presents many challenges [3] that need to be surpassed in the near future. Specific methodologies for ‘exploiting’ big data should be examined. Using big data as a potential source for Official Statistics triggers a need for new data treatment methods such as data mining techniques and statistical methods suited for large datasets. Also note that there is the concern that big data sets are not representative of the target population due the fact that they are selective by nature and then yield biased results. This issue is for example addressed in [4]. Not only the Official Statistics community but also other scientific fields, can benefit greatly from the possibilities offered by big data, but must invest in research and skills development [5].

## **REFERENCES**

- [1] UNECE (2014). How big is Big Data? Exploring the role of Big Data in Official Statistics. Draft for public review
- [2] UNECE (2014). The role of Big Data in the modernization of statistical production. Project proposal
- [3] UNECE (2013). What does ‘Big Data’ mean for Official Statistics for Official Statistics? Paper prepared on behalf of the High-Level Group for the Modernisation of Statistical Production and Services
- [4] Buelens B., Daas P., Burger J., Puts M., van den Hurk P.A.M. (2014). Selectivity of Big data. Statistics Netherlands
- [5] Daas P., Puts M.J., Buelens B., van den Hurk P.A.M. (2013). Big Data and Official Statistics. New Techniques and Technologies for Statistics conference, Brussels, Belgium.

# A Suggested Framework for National Statistical Offices for assessing the Quality of Big Data

Prepared by the Task Team on Big Data Quality, within the Big Data project overseen by the High-Level Group for the Modernisation of Statistical Production and Services<sup>1</sup>

**Keywords:** big data, input quality, throughput quality, output quality, hyperdimensions, quality indicators

## 1. Background

In April 2013, the UNECE Expert Group on the Management of Statistical Information Systems (MSIS) identified Big Data as a key challenge for official statistics, and called for the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to focus on the topic in its plans for future work [1]. As a consequence, the project *The Role of Big Data in the Modernisation of Statistical Production* was undertaken in 2014. The project comprised four ‘task teams’, addressing different aspects of Big Data issues relevant for official statistics: the Privacy Task Team, the Partnerships Task Team, the Sandbox Task Team, and the Quality Task Team.

This abstract summarises the outcome of the Big Data Quality Task Team, comprised of representatives from several national statistical offices all over the world. The Quality Task Team was asked to investigate the implications of Big Data for the quality of official statistics, and to develop a preliminary framework for National Statistical Offices (NSOs) to conceptualise Big Data quality.

The Task Team concluded that extensions to existing frameworks were needed in order to encompass the quality of Big Data. A preliminary framework was developed building on dimensions and concepts from existing frameworks, with additional dimensions and principles that were seen to be lacking in existing approaches.

The proposal from the Quality Task Team is going to be presented at the HLG meeting in November 2014. Development projects could be launched by the UNECE HLG to finalise the work undertaken by the Task Teams.

## 2. The Quality Framework: a general overview

The Big Data Quality framework, as developed by the UNECE Task Team, provides a structured view of quality at three macro-phases of the business process:

*input* – when the data is acquired, or in the process of being acquired;

---

<sup>1</sup> The task team was composed by the following members with affiliations: David Dufty, Australian Bureau of Statistics (chair); Hélène Bérard and Laurie Reedman, Statistics Canada; Sylvie Lefranc, INSEE, France; Marina Signore, Istat, Italy; Juan Munoz and Enrique Ordaz, INEGI, Mexico; Peter Struijs, Statistics Netherlands; Jacek Maślankowski and Dominik Rozkrut, Central Statistical Office of Poland; Boro Nikic, Statistical Office of Slovenia; Ronald Jansen and Karoly Kovacs, UNSD; and Matjaz Jug, UNECE.

*throughput* – any point in the business process in which data is transformed, analysed or manipulated. This might also be referred to as ‘process quality’;

*output* – the reporting of quality with statistical outputs derived from big data sources.

In terms of the General Statistical Business Process Model [2], *input* can be thought of as the “design” and “collect” stages of the GSBPM, *throughput* is equivalent to the “process” and “analyse” stages of the GSBPM, and *output* is equivalent to the “disseminate” stage of the GSBPM.

The proposed framework is using a hierarchical structure composed of three hyperdimensions with quality dimensions nested within each hyperdimension. The three hyperdimensions are the *source*, the *metadata* and the *data*. The concept of hyperdimensions has been borrowed from the administrative data quality framework developed by Statistics Netherlands [3].

*source* – this hyperdimension relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated.

*metadata* – it is the hyperdimension which refers to information available to describe the concepts, the contents of the file, the processes and the outputs.

*data* – the hyperdimension which relates to the various sources of error that could affect the data.

These same hyperdimensions are used throughout the framework, for the *input*, *throughput* and *output* phase. For each phase and hyperdimension, a set of quality dimensions and indicators have been defined. The quality dimensions can vary for each phase.

### **3. The Quality framework: specific dimensions to Big Data**

Starting from consolidated supranational and national frameworks, e.g. [4], [5] and [6], the Task Team focused on the specific quality requirements and challenges for the use of Big Data in official statistics. To this purpose, additional quality dimensions were considered, allowing for assessing error sources, error types, or possible limitations specific to the nature and origin of Big Data.

#### **3.1 Input phase**

At the *input* phase of the business process, a National Statistical Office should engage in a detailed quality evaluation of a Big Data source both before acquiring the data (known as the *discovery* component of the *input* phase), and after (i.e. the *acquisition* component).

The hyperdimensions *source* and *metadata* provide the opportunity for assessment before the data is obtained, i.e. in the *discovery* phase, and can be undertaken, for example, to decide whether to acquire the data at all, or determine what uses the data might be put to, or how much effort should be expended in acquiring it. The *Data* hyperdimension, on the other hand, can only be assessed once the data is actually acquired.

In some cases, the NSI requirements and the intended use of the data will be known prior to the start of the data quality evaluation. In other cases, potential use of the data will be discovered as the data is explored further. For both cases, at the onset, or as the evaluation

progresses, the Task Team recommended that the intended use be clearly documented as early as possible.

In addition to dimensions used to assess administrative data [3], the Task Team suggested the use of new dimensions for an NSO to employ, including *privacy* and *confidentiality* (a thorough assessment of whether the data meets privacy requirements of the NSO), *complexity* (the degree to which the data is hierarchical, nested, and comprises multiple standards), *completeness* (of metadata) and *linkability* (the ease with which the data can be linked with other data).

### 3.2 Throughput phase

For the *throughput* phase of the business process, four principles of processing are proposed:

1. *System Independence*: The result of processing the data should be independent of the hardware and software systems used to process it;
2. *Application of quality dimensions*: in monitoring process quality, quality dimensions as articulated here and in other frameworks should be used as a guide for quality evaluation;
3. *Steady states*: that the data be processed through a series of ‘steady states’, which are stable versions of the data that can be referenced by future processes and by multiple parts of the organisation (See [7] for a discussion of steady states in the context of statistical production systems);
4. *Quality gates*: that the NSO employ quality gates [8] as a quality control business process.

### 3.3 Output phase

For the *output* phase of the business process, the Task Team used the Australian Bureau of Statistics Data Quality Framework [5] as a starting point together with European framework [4]. At this level of the business process, the quality framework should be applicable to reporting, dissemination, and transparency. It is information about the quality of the statistical product that a consumer of that product would ideally have.

Additionally to existing frameworks, the Task Team recommended new output dimensions to consider: i) *secondary sources* and *confidentiality* for the hyperdimension *source*; ii) *complexity* for the hyperdimension *metadata* and iii) *selectivity* and *predictive power* for the hyperdimension *data*.

The dimension of *secondary sources* reflects the fact that Big Data products often involve the synthesis and transformation of data from a wide variety of data sources. The Task Team points out that quality reporting should allow for transparency about the nature of the acquisition, and the institutional credibility and trustworthiness of third parties.

With regard to accuracy, the dimension of *selectivity* or representativeness is to be considered when the use of Big Data source(s) does not allow for applying methodologies based on traditional sampling theory for deriving estimates and evaluating quality. This dimension may inform on potential bias when using BD. When BD source are sensors/meters,



representativeness of the target population might be affected also by the spatial distribution of the instrument (e.g the measurement grid) as well as by the periodicity of observations.

The dimension of *predictive power* is a necessary addition as new sources of data and types of statistical products become possible. With novel, diverse sources of data, sampling theory may not be an appropriate metric for evaluating the utility of derived metrics. Instead, the *predictive power* of the metric may need to be measured directly. *Predictive power* considers the ability of a metric derived from some data source to predict a variable of interest. The word ‘prediction’ here is being used in the statistical sense of providing some kind of empirical estimation; it does not necessarily refer to predicting future population characteristics.

#### 4. Conclusion

Rather than focusing quality efforts on statistical outputs, in regard to Big Data, NSOs need a series of quality frameworks across the business process. The UNECE Quality Task Team has recommended some principles as well as dimensions that would be useful for an NSO to evaluate Big Data sources and products.

#### References

- [1] UNECE (2013a). Final project proposal: The Role of Big Data in the Modernisation of Statistical Production. UNECE, November 2013.  
<http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production>
- [2] UNECE (2013b). GSBPM v5.0.  
<http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0>
- [3] Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Toth, J. (2009), *Checklist for the Quality evaluation of Administrative Data Sources*. Statistics Netherlands, The Hague/Heerlen
- [4] Eurostat (2011). European Statistics Code of Practice.  
[http://epp.eurostat.ec.europa.eu/cache/ITY\\_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF)
- [5] Australian Bureau of Statistics (2009) The ABS Data Quality Framework.  
<http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0>
- [6] Statistics Canada (2002). Statistics Canada’s Quality Assurance Framework.  
[http://www5.statcan.gc.ca/access\\_acces/alternative\\_alternatif.action?l=eng&loc=http://www.statcan.gc.ca/pub/12-586-x/12-586-x2002001-eng.pdf&t=Statistics%20Canada's%20Quality%20Assurance%20Framework](http://www5.statcan.gc.ca/access_acces/alternative_alternatif.action?l=eng&loc=http://www.statcan.gc.ca/pub/12-586-x/12-586-x2002001-eng.pdf&t=Statistics%20Canada's%20Quality%20Assurance%20Framework)
- [7] Struijs, Peter, et al. "Redesign of Statistics Production within an Architectural Framework: The Dutch Experience." *Journal of Official Statistics* 29.1 (2013): 49-71.
- [8] Quality Management of Statistical Processes Using Quality Gates, Dec 2010, cat. no. 1540.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1540.0>

# Proposal for an accreditation procedure for big data source

Albrecht Wirthmann ([Albrecht.Wirthmann@ec.europa.eu](mailto:Albrecht.Wirthmann@ec.europa.eu))<sup>1</sup>, Photis Stavropoulos ([photis.stavropoulos@agilis-sa.gr](mailto:photis.stavropoulos@agilis-sa.gr))<sup>2</sup>, Michalis Petrakos ([michalis.petrakos@agilis-sa.gr](mailto:michalis.petrakos@agilis-sa.gr))<sup>2</sup>, George Petrakos ([george.petrakos@agilis-sa.gr](mailto:george.petrakos@agilis-sa.gr))<sup>2,3</sup>

**Keywords:** accreditation, big data, secondary data sources, quality

## 1. INTRODUCTION

The exponential growth of semiconductors, storage and network capacities induced the digitization of increasingly more processes in businesses and in private life. This development has been creating massively increasing amounts of digital data, often referred to as big data, which can be exploited as a source of statistical information on social, economic and environmental phenomena. In general these data are collected by private and public entities outside the statistical system. However, they have high potential to be utilised for producing statistical data. Use of big data sources involves a paradigm shift for official statistics. Statisticians find themselves in the role of data customers instead of data producers who have to design statistical products from existing data sources. The production of official statistics is driven by high quality standards and principles. To be labelled as official statistics statistical products produced from big data sources have to meet these quality standards. The process of quality assessment of big data sources could be referred to accreditation. The paper proposes a possible accreditation procedure that producers of official statistics can use to for assessing big data sources.

## 2. METHODS

The authors reviewed sets of principles and quality frameworks adopted by producers of official statistics, in order to lay the background of the proposed accreditation procedure. Furthermore, the expanding literature on quality assessment of administrative and other secondary data as input for the production of official statistics was studied.

In a two-step approach, a set of principles were first proposed. Subsequently the accreditation procedure itself was specified. It is hoped that the principles will serve as guides for potential modifications of the proposed procedure or development of alternative ones.

## 3. RESULTS

### 3.1. Foundational principles

The following set of principles is proposed as a basis for any procedure of accreditation of non-official data sources. They are ‘designed’ to remain stable over time, even if changing conditions necessitate modifications in the accreditation procedure itself. The principles ensure that the procedure is fit for use in a statistical system, efficient, cost

---

<sup>1</sup> European Commission - Eurostat.

<sup>2</sup> Agilis S.A. Statistics and Informatics, Akadimias 98-100, 10677, Athens, Greece.

<sup>3</sup> The authors gratefully acknowledge their fruitful discussions with and useful input from Dr. George Sciadas, Statistics Canada.

effective and reliable and that it takes into account all possible impacts of the adoption of a new data source on the producer of official statistics.

**Principle 1:** Accreditation must be fully compliant with the well-established principles and quality frameworks that guide the world of official statistics and consistent with quality assurance practices embedded deeply in the work of statistical offices.

**Principle 2:** Any accreditation procedure must be flexible in a way that does not unduly prejudice or rule out new opportunities without serious examination.

**Principle 3:** An accreditation procedure should include sequential decision-making based on a pragmatic step-wise approach, so that new data sources that will not work are spotted early on, while investment in those that will work is not jeopardized.

**Principle 4:** The accreditation procedure must contain an empirical assessment with real data and it must be carried out by statistical offices directly. It cannot be delegated to filling out questionnaires by the source owners.

**Principle 5:** A systematic accreditation procedure must assess the quality of the statistical inputs (including the source and metadata), of the statistical outputs, as well as of the statistical processes involved.

**Principle 6:** The final decision for the accreditation of a new data source must also incorporate a combination of corporate criteria, broader than strict data quality. The accreditation procedure must compile adequate supporting documentation, including measurements.

### **3.2. Proposed accreditation procedure**

Consistent with the preceding principles, the proposed accreditation procedure evolves in a step-wise fashion with gradual assessments involving indicators<sup>4</sup> measured through scales and hard data, which in turn lead to recommendations associated with six decision points, one at the end of each stage as well as one more after the first phase of stage 3.

#### **Stage 1: Initial examination of source, data and metadata**

In order for a statistical office to even contemplate acquiring and using an external data source some knowledge of it is a necessary condition. At this stage, an early assessment of the data, the metadata and the source is needed. Anything that can be gauged from the outside or through limited and rather unofficial interaction with the working level at the source organisation should be collected, shared internally, and examined. The material should cover the *raison d'être* of the source's owner and as many aspects of content, data and metadata as possible.

The overarching question is potential usefulness. Detailed questions can examine the population coverage, units of measurement, variables, timeliness, frequency, as well as provide some information on the organisation. They should also include possible uses of the data.

---

<sup>4</sup> The indicators and illustrative examples as well as a flowchart of the procedure will be presented in the full paper.

There should be no concern with the feasibility of acquiring the data. Similarly, the quality of eventual outputs should not enter the picture, not even the quality of the data themselves yet.

## **Stage 2: Acquisition of data and assessment**

This stage entails negotiations with the source with a view to acquire a set of data adequate for rigorous testing. The primary objective is to clarify whether the source is willing and able to deliver files or extractions at the record level, as well as keep open a communication channel during the testing process. A number of issues must be discussed in a professional manner with the source, albeit without formalizing a legal agreement yet - which is more demanding. These include specifications of files or file extractions, time and method of transmission, as many metadata as possible, and any particular conditions that must be known.

## **Stage 3: Forensic investigation**

This represents a critical step and requires a fair amount of work by the statistical office. It can be sub-divided in four distinct phases:

i) Data editing: all the known steps taken for the processing of collection files apply. Duplicate records will be identified and removed, specifications for various kinds of edits will be developed in a way that will correct erroneous, inconsistent or contradictory entries, outliers will be dealt with, and documentation will be kept.

ii) Production of aggregate statistics: In this phase, we produce actual aggregate statistics, which are then analysed and compared with any existing data.

iii) Production of outputs: The data are used in the production of actual statistical products, which can be standalone outputs or parts of one or more existing outputs. This must be accompanied by detailed analyses of the impacts on the quality of existing outputs.

iv) Assessment of the capacity of existing statistical tools to handle the new data: In this phase we assess whether the available statistical tools of the statistical office can adequately deal with the examined data. Issues of storage and processing must be examined explicitly, as the amounts of data may be vast and conceivably may require special software, analytics tools and special skills.

## **Stage 4: Statistical office decision**

This stage is dedicated to the assessments necessary for a corporate decision to be made based on as much information and knowledge as possible.

As a first step, an account of the outputs and the indicators quantified during the previous stage is required. What is needed is an itemisation of the exact uses of the new data and their impacts. What specific new output/s can be produced that will expand the statistical office's offerings, which output/s can benefit, to what extent, how, and what would be the implications and trade-offs?

A second step entails a top-level cost-benefit analysis, which focuses on the financial picture. The suggested indicators are consistent with those listed in [1].

The third step places the emphasis on the risks that need to be undertaken and managed by the statistical office. How vulnerable will be the outputs involved, and by consequence the reputation of the statistical office, to factors outside its control? What will be the mitigation strategies?

A final step before making a decision at this point involves an assessment of the feasibility of incorporating the new source into the statistical office's operations from a legislative and socio-political point of view. Such issues are dealt with in [2].

#### **Stage 5: Formal agreement with source**

This final stage involves high-level negotiations with the source to secure cooperation and arrive at a formal and comprehensive agreement. The statistical office is now well equipped with the information it needs for such deliberations. This is one of the reasons why it is more prudent to start with the data. Information needed for the next level will be known, and there will be a clear identification of trade-offs to guide and facilitate negotiations.

At the outset a good understanding is needed that willingness to cooperate is not an abstract notion but matched by deeds. The early implications of this translate to obligations by the source to commit needed resources and by the statistical office to respect lines that the source may not want crossed. Much will depend on the type of the source – public or private, statistically inclined or not, stage of advancement etc.

Then issues of reciprocity involved in a fair deal must be explicitly clarified. Terms and conditions of the agreement will be discussed in detail, supported by accompanying documentation from the working teams. At the end, the issue of governance needs to be articulated, including change management and a dispute resolution mechanism.

This stage in the accreditation procedure can also be subject to quantifiable indicators as they emerge both from knowledge of what is involved and attitudes. They will serve well in subsequent rounds, complemented of course with the experience accumulated at that time.

#### **4. CONCLUSIONS**

In today's corporate environment that calls for ever greater cost-effectiveness on the one hand and faced with an increasing demand for reliable statistical information on the other hand, statistical offices simply cannot afford to ignore big data. Indeed, they are turning their attention to such data as potential sources for official statistics. On the other hand, statistical offices cannot lower their quality standards and professional integrity without risking losing their reputation as independent providers of reliable statistical information.

Any potential new data source needs therefore to be scrutinized in an unbiased, professional manner with a rigorous accreditation procedure which encompasses aspects beyond purely 'statistical' quality. The procedure proposed in this paper is flexible enough to accommodate the needs of different statistical offices vis-à-vis a multitude of potential sources. It is hoped that it can serve as the backbone of a harmonized approach for the assessment of big data in the European Statistical System and beyond.

## REFERENCES

- [1] P. Daas and S. Ossen, Report on methods preferred for the quality indicators of administrative data sources, Deliverable 4.2 of project BLUE – ETS (2011).
- [2] L. Angelis, D. Kalogeras, M. Petrakos, T. Priftis, V. Sotiropoulos, P. Stavropoulos, M. Vafopoulos, Results of the feasibility analysis, Deliverable D2 of Eurostat contract 50721.2013.002-2013.169 (2014).

# Theoretical and practical aspects of mapping poverty in Poland using small area estimation methods

Marcin Szymkowiak ([m.szymkowiak@stat.gov.pl](mailto:m.szymkowiak@stat.gov.pl), [m.szymkowiak@ue.poznan.pl](mailto:m.szymkowiak@ue.poznan.pl))

**Keywords:** small area estimation, mapping poverty, EU-SILC

## 1. INTRODUCTION

The European Survey on Income and Living Conditions (EU-SILC) is the basic source of information published by GUS (Central Statistical Office in Poland) about the relative poverty indicator both for the country as a whole and at the regional level. This also applies to other countries facing a growing demand for good poverty maps. In order to follow appropriate social, which is consistent with the guidelines of the cohesion policy, one needs to measure poverty and provide information about this phenomenon at lower levels of spatial aggregation. In this context poverty maps are used to support decisions concerning important political issues, such as allocation of development funds by governments, National Ministries of Infrastructure and Development or international organizations, such as the World Bank. Those decisions should be based on the most accurate poverty indicators, estimates or numbers and should be delivered at the lowest level of spatial aggregation.

However, given the small sample size in the relevant cross classifications of the EU-SILC survey, it is necessary to use the latest techniques of indirect estimation draw on alternative data sources to estimate the parameters of interest at low levels of spatial aggregation with acceptable precision. Since the EU-ILC survey does not cover adequately all the specific areas or population subgroups, the required information can only be obtained using small area estimation techniques based on the idea of ‘‘borrowing strength’’. In Poland, for instance, EU-SILC data are only sufficient to publish the at-risk-of-poverty rate at the level of the whole country and at the regional level (NUTS 1). Owing to small sample sizes and low precision of estimation, adequate estimates at lower level of spatial aggregation cannot be delivered.

## 2. METHODS

Small area estimation (SAE) methodology has been developed to produce reliable estimates of different characteristics of interest, such as means, count, quantiles or ratios for domains for which only small samples are available [1]. For specific domains there may even be no samples available. SAE methodology also deals with the problem of how to assess the precision of estimation given small sample sizes in specific domains, when the precision of obtained direct estimates is rather low. As a consequence, indicators or figures published by the official statistical system cannot be based on direct estimation including the Horvitz-Thompson approach.

SAE methodology is used by different National Statistical Institutes in different areas, especially to estimate quantities related to the labour market, agriculture or business statistics. It is also useful in mapping poverty. For instance, the World Bank has used the SAE methodology to prepare poverty maps for more than 60 countries all over the world.

In the field of poverty mapping there are different approaches to choose from. In particular, common SAE-based poverty mapping methods may include [2]:

- direct estimates, which are in generally inefficient,
- Fay-Herriot estimates, which enable aggregation, specific modelling, specification of sampling variances,
- ELL estimates, which are used by the World Bank and may be poorly efficient when auxiliary variables do not explain the entire between-area variation,
- the EB approach based on a nested-error model, which is very efficient under normality,
- the HB approach based on a nested-error model, which is similar to the EB approach but is less computationally demanding,
- M-quantile methods, which are less sensitive to outliers.

All of these methods may be used to tackle the problem of estimating different poverty parameters. In particular, these methods enable poverty mapping at lower levels of spatial aggregation. It refers to situations when EU-SILC samples are too small to produce reliable estimates and it is necessary to “borrow strength” from other statistical data sources, such as censuses, administrative registers or sample surveys.

### **3. RESULTS**

In Poland SAE methodology has so far been used in the area of the labour market, agriculture and business statistics. In 2013 the Center for Small Area Estimation, which is a special unit at the Statistical Office in Poznań, in cooperation with GUS and the World Bank prepared a poverty map of Poland at the level of subregions (NUTS 3) using the Fay-Herriot approach [3]. By implementing the Fay-Herriot area level model it was possible to produce estimates of the at-risk-of-poverty rate in Poland at the level of subregions, i.e. at a lower level of aggregation than the direct estimates published by official statistics so far. This has increased the scope of information about poverty: it is now available at the level of 66 subregions. A preliminary analysis of the poverty map created using the SAE methodology has revealed a difference between Central and Eastern Poland (with a higher poverty rate) and Western Poland, characterised by a lower at-risk-of-poverty rate. Given the growing demand for information about the at-risk-of-poverty rate at lower levels of spatial aggregation (NUTS 4 – *powiats* in Polish), there is a pressing need to take advantage of appropriate small area estimation techniques and data from different statistical sources (EU-SILC, census or administrative registers).

### **4. CONCLUSIONS**

The main aim of this presentation is to show theoretical aspects of small area estimation methods in the context of poverty mapping. Theoretical considerations will be illustrated with practical applications of SAE techniques in Poland at lower levels of spatial aggregation than those published by the Central Statistical Office to date.

### **REFERENCES**

- [1] Rao J.N.K (2003), Small Area Estimation, Wiley Series in Survey Methodology, A John Wiley & Sons, Inc., Publication.



- [2] Molina I., Rao J.N.K., Nandram B., Marin J.M., Graf M. (2014), Sae in poverty mapping, presentation on the SAE 2014 conference<sup>1</sup>.
- [3] Statistical Office in Poznan, Center for Small Area Estimation (2013), Poverty Maps at the Subregional Level in Poland Based on Indirect Estimation, Poznan<sup>2</sup>.

---

<sup>1</sup> [http://sae2014.ue.poznan.pl/presentations/SAE2014\\_Isabel\\_Molina\\_5479b14c2c.pdf](http://sae2014.ue.poznan.pl/presentations/SAE2014_Isabel_Molina_5479b14c2c.pdf)

<sup>2</sup> In Polish [file:///C:/Users/KS\\_MS/Downloads/mapowanie\\_ubostwa\\_prace\\_studialne%20\(1\).pdf](file:///C:/Users/KS_MS/Downloads/mapowanie_ubostwa_prace_studialne%20(1).pdf)

# Estimation of poverty rate for small areas by model calibration and "hybrid" calibration methods

Risto Lehtonen (risto.lehtonen@helsinki.fi)<sup>1</sup>, Ari Veijanen<sup>2</sup>

**Keywords:** Assisting model, Calibration techniques, Design-based methods, Generalized regression estimation, Unit-level auxiliary data

## 1. INTRODUCTION

We examine in the paper the statistical properties (design bias and accuracy) of certain calibration estimators for the estimation of finite population parameters for population subgroups or domains and small areas. Methods considered include the traditional model-free calibration and model calibration methods, and a combination of these two methods called hybrid calibration. We compare the methods with generalized regression (GREG) estimators. Quality of the estimators is assessed by design-based simulation experiments using register-based population data maintained by Statistics Finland.

Our interest is in the estimation of the at-risk-of poverty rate (poverty rate for short) for the domains of interest. Poverty rate is one of the components of the so-called combined AROPE indicator (at risk of poverty or social exclusion). The AROPE rate is the key indicator in monitoring the poverty target in the EU 2020 Strategy [1]. Data for our study variable comes from a sample survey (e.g. an income survey). We assume an access to register-based auxiliary information covering the target population, and an option to link the sample survey data with the register data at the unit level. This offers a flexible framework for the construction of the poverty rate estimators for the domains of interest. There is an increasing number of statistical infrastructures, where such a framework has been developed for official statistics production, notably in Europe (Denmark, Finland, Norway and Sweden as forerunners; see [2]).

## 2. METHODS

Calibration techniques [3], [4] are popular in design-based estimation of finite population parameters such as totals and means. In model-free (or linear) calibration, weights are calibrated to reproduce the known population totals of the auxiliary variables (the so-called coherence criterion). Aggregate-level auxiliary data are used in the construction of the calibration equations. No explicit model statement is needed in the method. The same set of calibrated weights can be supplied to the diverse study variables of a survey. In official statistics production, these properties are often considered a benefit. From statistical point of view, model-free calibration is a natural choice for continuous study variables whose relationship to the explanatory auxiliary variables can be described by a linear model.

Model calibration (MC) [5], [6] represents a model-assisted technique, where an assisting model is explicitly stated. The weights are calibrated to reproduce the population total of the predictions derived via the specified model. Access to unit-level auxiliary data is assumed, and a considerable modelling effort is often needed. In a model calibration procedure, the method is applied separately for each study variable. Coherence of the estimated totals or means with published statistics is not guaranteed. A benefit of model

---

<sup>1</sup> University of Helsinki

<sup>2</sup> Statistics Finland

calibration is that models beyond the linear model can be specified, such as logistic models and other members of the generalized linear models family. This option offers a flexible treatment of different types of study variables, such as binary, polytomous and count variables, as well as continuous variables whose relationship to the explanatory variables is non-linear. Model calibration was investigated for domain estimation in Lehtonen and Veijanen [7], [8].

Montanari and Ranalli [9] introduced a version of calibration called multiple model calibration. In this method, the coherence property of estimates with published statistics is retained for a specified subset of auxiliary variables, and another set of auxiliary variables is treated with model calibration. Lehtonen and Veijanen [10] call this method "hybrid" calibration, because the method aims at combining some of the favourable properties of model-free calibration and model calibration. For example, a part of the set of auxiliary variables (whose unit-level population data are available) is incorporated in the assisting model (to improve accuracy) and another part (variables whose domain-level population totals are available) is incorporated in the model-free calibration procedure (for coherence with published statistics). The two sets of auxiliary variables can be distinct or they may overlap.

In our first experiment, we compare the semi-indirect variant of model calibration with direct Horvitz-Thompson (HT) type estimator and indirect model-assisted logistic generalized regression (GREG) estimators [11], [12]. In the second experiment, we compare hybrid calibration with model calibration and model-free calibration methods. Our study variable (poverty indicator) describing poverty is binary and is modelled by logistic fixed-effects and mixed models. Design bias and accuracy of the methods are assessed by design-based simulation experiments using real unit-level register data maintained by Statistics Finland.

### 3. SIMULATION EXPERIMENTS

For design-based simulation experiments we constructed a unit-level population of one million persons in 36 NUTS4 regions in Western Finland. The equivalized income, age class (0-15, 16-24, 25-49, 50-64, or at least 65 years) and gender were obtained from registers, whereas the labour force status and the socio-economic status were obtained from a household survey for the household head and imputed for the other members of each household. Our binary study variable (poverty indicator) was constructed using register data on equivalized income. In the simulations, a number of independent SRSWOR samples were drawn from the fixed population. The domains of interest were of unplanned type. We calculated domain estimates of poverty rate for the  $D=36$  NUTS4 regions. The quality of an estimator of domain total over the simulations in a domain was assessed by absolute relative bias (ARB; absolute value of the difference of mean of an estimator over simulations to the true value, relative to the true value) and relative root mean squared error (RRMSE; square root of the squared difference of mean of an estimator over simulations to the true value, relative to the true value).

In the first experiment, we compared GREG and model calibration (MC) estimators with a direct HT-type method that does not incorporate auxiliary information. To study the effect of the type of an assisting model we fitted two types of models: (a) fixed-effects logistic common model without domain-specific terms and (b) logistic mixed model with random intercepts associated with the 36 NUTS4 regions. The models included class indicators corresponding to main effects and interactions of age class (5 classes) with gender, labour force status (3 classes) and the socio-economic status (5 classes). GREG

estimators and MC estimators were used with models (a) and (b). The MC estimator was of semi-indirect type. The direct HT type estimator serves as a reference.

The sample size in the  $K = 1000$  simulations was  $n = 1000$  persons. The domains were classified by expected domain sample size into three classes (5-12, 12-25, 25-151 units). Results are in Table 1. All estimators were nearly design unbiased as expected. This property also holds for the model-assisted estimators, irrespective of model choice. All model-assisted methods outperformed the HT-estimator in accuracy. In model-assisted methods, semi-indirect MC estimator yielded slightly smaller RRMSE than the corresponding LGREG or MLGREG estimator. The best results were obtained with the mixed model, and model calibration yielded slightly more accurate estimates than the GREG estimator assisted with the logistic mixed model (MLGREG).

**Table 1** Mean absolute relative bias (ARB) (%) and mean relative root men squared error (RRMSE) (%) of Horvitz-Thompson, logistic GREG and model calibration estimators of poverty rate over domain size classes for 36 small regional areas.

Estimator	Mean ARB (%)			Mean RRMSE (%)		
	Expected domain sample size			Expected domain sample size		
	5-12	12-25	25-151	5-12	12-25	25-151
HT	1.7	2.2	0.9	83.7	60.1	38.9
<i>Logistic fixed-effects model</i>						
LGREG	2.1	1.7	0.9	72.8	55.5	37.1
Semi-indirect MC	2.0	1.9	1.1	72.5	55.3	37.0
<i>Logistic mixed model</i>						
MLGREG	2.0	1.8	0.9	72.4	55.0	36.8
Semi-indirect MC	1.9	1.8	0.8	72.1	54.8	36.9

For hybrid calibration, we used gender, age group and labour force status as explanatory variables in the logistic fixed-effects models. Main effects of the x-variables were included in the models, and there were no domain-specific parameters. In model-free calibration we calibrated to domain totals of gender, age group and labour force status. In model calibration, we calibrated to domain totals of predictions from the model with all three variables (gender, age group, labour force status) as the x-variables. In hybrid calibration, we included sex and age group in the model calibration part. Labour force status was included in the model-free calibration part. Thus, the overall calibration procedure was to the domain totals of the fitted values from the model and to the known domain totals of labour force status. In the simulations,  $K = 500$  SRSWOR samples of  $n = 2500$  were drawn from the population. The domains were classified by expected domain sample size into three classes (<25, 25-50, >50 units). Results are in Table 2.

**Table 2.** Mean absolute relative bias (ARB) (%) and mean relative root mean squared error (RRMSE) (%) of model-free calibration, model calibration and hybrid calibration estimators of poverty rate over domain size classes for 36 small regional areas.

Estimator	Expected domain sample size			
	<25	25-50	>50	All
<i>Mean ARB (%)</i>				
Model-free calibration	2.0	2.2	1.2	1.8
Model calibration	2.4	1.9	1.1	1.7
Hybrid calibration	2.3	1.9	1.1	1.7
<i>Mean RRMSE (%)</i>				
Model-free calibration	51.5	39.2	25.9	36.1
Model calibration	51.9	38.8	25.5	35.8
Hybrid calibration	52.0	38.9	25.6	35.9

Again, all methods were nearly design unbiased. All methods were quite comparable in accuracy. In overall accuracy, model calibration and hybrid calibration slightly outperformed model-free calibration, but the differences were small. The results indicate that a feasible combination of traditional model-free calibration and more recent model calibration methods can provide a compromise method partly fulfilling the goals of flexible modelling, accuracy improvement and the coherence property. Relative properties of the methods are investigated in further research.

## REFERENCES

- [1] European Commission (2013) Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy. Publications Office of the European Union, Luxembourg.
- [2] Wallgren, A. and Wallgren, B. (2014) Register-based Statistics: Statistical Methods for Administrative Data, 2nd Edition. John Wiley & Sons, Ltd, Chichester.
- [3] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.
- [4] Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99–119.
- [5] Wu, C. and Sitter, R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185–193
- [6] Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling, *Journal of the American Statistical Association*, 100, 1429–1442.
- [7] Lehtonen, R. and Veijanen, A. (2014). Small area estimation of poverty rate by model calibration and "hybrid" calibration. NORDSTAT 2014 Conference, Turku, June 2014.
- [8] Lehtonen, R. and Veijanen, A. (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics*, 66, 125–133.
- [9] Montanari, G.E. and Ranalli, M.G. (2009). Multiple and ridge model calibration. *Proceedings of Workshop on Calibration and Estimation in Surveys 2009*. Statistics Canada.
- [10] Lehtonen, R. and Veijanen, A. (2014). Model-assisted methods to small area estimation of poverty indicators. In Pratesi, M. (Ed.) *Analysis of Poverty Data by Small Area Estimation*. Chichester: Wiley. (Forthcoming)
- [11] Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33–44.
- [12] Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.) *Handbook of Statistics, Vol. 29B, Sample Surveys. Inference and Analysis*. Amsterdam: Elsevier, 219–249.

# Towards an integrated Consumer Expenditure Survey -- Combining Multi-mode Data Collection and Big Data Extracts

Gustav Haraldsen ([gha@ssb.no](mailto:gha@ssb.no)), Sverre Amdam<sup>1</sup>, Li-Chun Zhang<sup>2</sup>

**Keywords:** Multi-mode, Transaction data, Big data, Response burden, Consumer Expenditure

## 1. INTRODUCTION

The Consumer Expenditure/Household Budget Survey is traditionally based on meticulous diary reports over a 14 days period, combined with an initial and concluding interview. The survey has suffered a high response burden and low response rate; typically about 50%. Interview costs and coding based on hand written specification or enclosed receipts have also made the survey expensive. Several redesign options have been suggested for this kind of surveys (e.g. Cantor, Mathiowetz et al. 2013). Our principal approach in the 2017-survey will be to combine digital transaction data with different kinds of questionnaires. In addition we will try to extract consumer patterns from databases kept by grocery retailers and link these results to the survey sample, e.g. to counteract bias because of nonresponse or to replace the most burdensome part of the diary.

## 2. METHODS

### 2.1. Multi-mode Survey Design

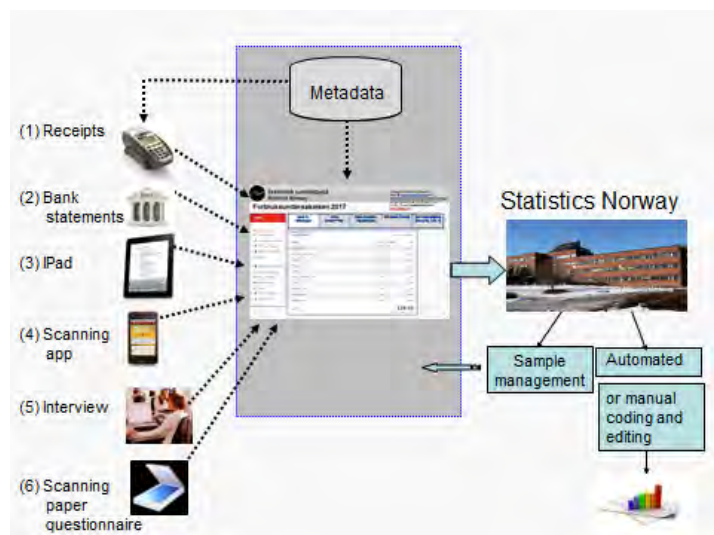
Figure 1 depicts a multi-mode design where data from different inputs are collected into a common self-completion web questionnaire and, when completed, data are transferred to the statistical office for statistical processing. The core element in this design is a common metadata-base that both define the response categories in the questionnaire and the deliveries from different inputs. The most noticeable input source in this design is e-receipts from debit and credit card transactions that are picked up from the shops payment systems and electronically transferred into the respondents' diary of daily expenditures (1). The technology for this is developed and will soon be tested in a pilot study. In a similar way we will try to collect e-bank statements covering invoice payments (2).

---

<sup>1</sup> Statistics Norway

<sup>2</sup> University of Southampton/Statistics Norway

Figure 1: Multi-mode Design for Consumer Expenditure Survey



We also envisage that some diary data can be entered by a tablet questionnaire (3) or scanned from smart phones (4). In addition to these input sources, some data may be collected by interviewers (5) or from paper diaries and questionnaires that are scanned by delivery (6). For respondents who are unable or unwilling to conduct the survey using digital inputs 1-4, input 5-6 (telephone and paper diary) will replace these.

## 2.2. Big Data Extracts

The multi-mode design will reduce the response burden for those who allow us to automatically collect transactions. These records will also be more accurate. Self-completion and digitalized data will reduce costs. But for those respondents who cannot or do not want to give us access to electronic transactions, the consumer expenditure surveys will still be a heavy burden. And there will still be unit and item nonresponse. We will try to meet these challenges with an alternative approach to sample-based statistics. Instead of generalizing from a small number of observations, as we do in surveys, we will try to extract expenditure consumption patterns from large data files and link these results to the same groups as those generated from the survey data.

The most burdensome part of the diary completion is to report daily grocery shopping. In Norway most of the grocery market is split between three major grocery retailer chains (GRC). Two of these, covering approximately 75 % of the grocery market, offer loyalty programs which record what members buy in their shops. When accumulated these files form a rich source of data. By linking social characteristics to those present in the retailer's database, the statistical office can produce statistics analogue to that produced from survey samples. The data can be accumulated over time to avoid seasonal variation associated with sample diary report. The large amount of available data also means that sampling uncertainty is a lesser error in this source. The key quality issues in this kind of data are rather selectivity and potential coverage errors. The loyal GRC customers is unlikely a simple random sample of the target population of the Expenditure Survey. Not all the household members of the loyal GRC customers may be found in the same database. Retailers at GRC may not cover the whole range of retail expenditure. Methodological issues to be explored are alternative possibilities of detecting and adjusting for such errors. The principal alternatives can be distinguished by whether they require linkage between the Expenditure Sample and the GRC databases or not.

Both methods have strengths and weaknesses which have to be dealt with, but potentially survey data and big data extracts supplement each other.

### **3. FUNDAMENTALS**

Receipts electronically captured from shops are already offered to customers of a digital post box in Norway, and Statistics Norway is able to use the same technology. One important constraint, however, is the number of shops which have taken the system in use. In particular, the coverage of GRC's is still low.

Only just above 4 % of the expenditures of Norwegian households are in cash (Norges Bank 2014). To leave this out of a consumer expenditure survey seems to be a minor problem. We are, however, dependent on an informed consent from the respondents in order to download their transaction data. The number of respondents who are willing to give us access to digital transaction data, opposed to those that want to conduct the survey in the original diary/interview form have a lot to say for both the business case of multi-mode design, as well as the quality of the data.

We have surveyed individual's willingness to let Statistics Norway access debit/credit card data, as well as data from bank accounts in a limited reporting period. The purpose of the study was to get a better understanding of people's attitudes towards disclosing this kind of personal financial information for statistical purposes, as well as understanding what affects the decision to participate. The survey was included as a part of Statistic Norway's omnibus survey in July/August 2014, which had a net sample of 1082 respondents.

The results showed that 40 % of those asked said they were willing to give SSB access to bank transactions. Only age had a significant impact on the willingness, where younger age groups were more willing to disclose debit/credit card information, as well as bank account transactions. Surprisingly people's internet experience did not seem to have a significant impact on willingness in the survey. An Integrated Expenditure Survey (IES), which also makes use of GRC database extracts, will have important differences from a modernised mixed-mode expenditure survey (MES) across the production processes, including sampling design, questionnaire design, editing and estimation. For instance, loyal GRC customers may be over-represented in the sample. The questionnaire for these sample households will focus more on supplementary retail data, and other information related to potential coverage errors. The editing and estimation may no longer be based on sample weighting alone, and statistical model-based methods will be necessary.

### **4. SUMMARY**

The technology and the widespread use of credit/debit card in the population form a basis for a multi-mode consumer expenditure survey primarily based on digital transactions. The main challenges are to sell in the technology to shops, particularly GRCs, as well as sell in the transaction option to those sampled. As an alternative we are therefore looking into ways to extract consumptions patterns from data files established by customers' loyalty programs.

### **REFERENCES**

Cantor, D., N. Mathiowetz, et al. (2013). Redesign Options for the Consumer Expenditure Survey. Washington DC, Bureau of Labour Statistics



Norges Bank (2014): Utviklingstrekk i kunderetta betalingsformidling - 2013 (Trends in customer-oriented payments). Oslo, Norges Bank Memo no 1 2014.

# Integrating the Web Mode in the Austrian Household Budget Survey 2014/15 – First Experiences with a New IT System for Data Collection in a Mixed Mode Design

Marc Plate (marc.plate@statistik.gv.at)<sup>1</sup>. Romana Riegler (romana.riegler@statistik.gv.at)<sup>2</sup>

**Keywords:** Web Survey; Mixed-Mode Design; Integrated System for Data Collection; Paradata; Total Survey Error; Quality during the field phase

## 1. INTRODUCTION

Modern surveys in official social statistics require fast data collection combined with high quality and a minimum of cost and burden for respondents. One popular strategy to meet these demands is the adaptation of classical face-to-face based survey designs to allow for mixed modes of data collection. In recent years this amounted in particular to the integration of innovative modes like CAWI. A second strategy is the modularization of questionnaires, in order to conduct efficient multipurpose surveys on varying levels of aggregation. Both strategies represent significant management challenges to the organizational and technical system of data collection within one institution. Existing data collection software such as Blaise can only partly meet these challenges. For this reason Statistics Austria has begun to develop its own integrated software. The tool has been designed to meet the following demands:

1. Enable cost-efficient case-management: centralize the management of all questions, cases and the field in one system.
2. Enable multipurpose surveys: make use of the concept of modularity and organize survey questions by topic in modules and submodules.
3. Enable mixed-mode designs: make use of “classic” and “new” ways of data collection (CATI, PAPI, CAPI, CAWI) and supply the possibility of combining modes in a concurrent, sequential or longitudinal mixed-mode design.

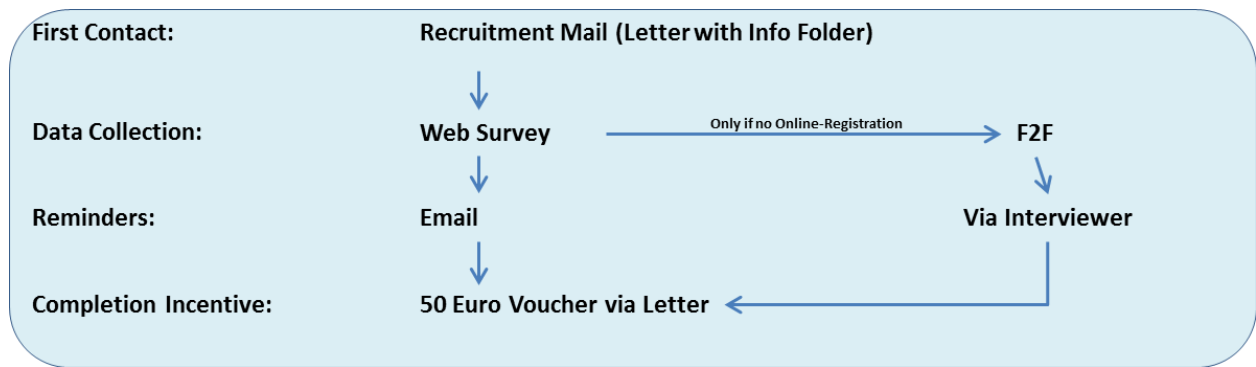
The new software tool is implemented gradually in social surveys conducted at Statistics Austria in order to accommodate upcoming needs whenever they arise. This paper presents a preliminary assessment of the first field experiences.

The Austrian Household Budget Survey 2014/15 was the first survey that has been selected to implement and test the new IT system. At the present stage of the development the tool is used primarily to support a web survey in the following design:

---

<sup>1</sup> Statistics Austria

<sup>2</sup> Statistics Austria



**Figure 1. Mixed-Mode Survey Design of the HBS 2014/15**

The introduction of new survey technology, and in particular the implementation of new data collection modes presents a strong case to systematically evaluate the quality of the survey, including aspects of the survey instrument and the field phase.

In our planned contribution we follow the Total Survey Error paradigm. We will present the current evidence and intend to discuss the potential impact of design changes on each source of error. In particular, we aim to demonstrate the added value of systematic integration of auxiliary information from paradata and administrative data. We will conclude with implications for future use of mixed mode designs in general and web survey when surveying the general population.

## **2. METHODS**

The assessment of the new survey design is based on all data available by end of February 2015, including survey responses from approximately 300 web interviews and 500 F2F interviews, as well as paradata and administrative data. We aim to construct indicators for each dimension of Total Survey Error to further adapt questionnaire designs and the fieldwork process.

During the fieldwork, the IT-System collects very useful paradata, that is information about the process of response. For the calculation of quality indicators the following paradata is used:

- Which devices, with which resolutions are respondents using
- Which buttons (Help, Back, Forward...) are being clicked and when
- Different durations: per question, per questionnaire, per explanatory note ...
- Which plausibility checks were triggered
- Changes of answers

## **3. EXPECTED RESULTS**

We want to concentrate on three dimensions of the TSE: Nonresponse Errors, Measurement Errors and Processing Errors. The field phase of the Austrian HBS has started only in September 2014. Consequently, there are only preliminary results available.

### **3.1. Non Response Error**

Indicator 1: Response Rate as calculated by the American Association of Public Opinion research (AAPOR).

- Expected results: we anticipate lower response rates for web interviews than for CAPI. The main question of interest is: which are the means of contacting that successfully raise participation in CAWI? Until the conference date we will have tried to contact households with a long letter and a short letter. Furthermore, experiments with additional reminder postcards and pre-incentives are scheduled for March 2015.

Indicator 2: Mode Selection Bias as made possible by the use of administrative data.

- Expected results: The socio-demographic structure for web respondents is likely to differ from that of CAPI respondents. The main question of interest is: to which degree do these two samples differ and what does this mean to the overall representativity?

### **3.2. Measurement Error**

Indicator 1: Answer durations as revealed through paradata.

- Expected results: Respondents will be likely to take less time to answer the questions than in the presence of an interviewer. The main question of interest is: Are certain types of questions or groups of respondents more affected than others, and what do different answer durations mean?

Indicator 2: Use of explanatory notes available in the web-questionnaire.

- Expected results: Respondents may be reluctant to read additional explanations. The main question of interest is: How can we ensure full comprehension by respondents?

### **3.3. Processing Error**

Indicator 1: Change of answers by respondents as recorded by paradata.

- Expected results: Respondents may tend to edit their own answers for vague questions. The main question of interest is: can the frequency of changes of responses be used to improve question wording?

## **4. CONCLUSIONS**

In the usual situation conclusions on new survey modes are only available after completion of fieldwork. By contrast, the present exercise aims to monitor ongoing fieldwork in order to take timely counter action. To optimize survey design it is important to jointly consider all the aspects which determine survey error. The total survey error framework therefore appears essential, especially when administrative and paradata are available.

## REFERENCES

- [1] Groves et. al (2009): Survey Methodology. Wiley Series in Survey Methodology. NJ, USA.

# Adapting Labour Force Survey questions from interviewer-administered modes for web self-completion in a mixed-mode design

Peter Betts ([peter.betts@ons.gsi.gov.uk](mailto:peter.betts@ons.gsi.gov.uk))<sup>1</sup>, Ben Cubbon ([ben.cubbon@ons.gsi.gov.uk](mailto:ben.cubbon@ons.gsi.gov.uk))<sup>2</sup>

**Keywords:** Labour Force Survey (LFS), web survey, mixed mode survey design, mode effects, collaborative research, qualitative research

## 1. INTRODUCTION

In common with many National Statistical Institutes the United Kingdom Office for National Statistics (ONS) is developing methods, processes and systems for online social surveys.

Questions on the Labour Force Survey (LFS) were originally designed to be interviewer administered. Interviewers perform important functions such as: motivation of respondents to take part in a voluntary survey; consistent administration of questions and accompanying guidance; clarification and explanation of questions and definitions to respondents. This helps to reduce respondent burden and ensure quality data. As social surveys move towards web data collection, respondents will be responsible for administering questionnaires themselves without assistance.

We cannot replace the existing interviewer-administered modes entirely, but can aim to employ a mixed mode design. This has implications for the design of the questionnaire instrument. As far as possible the data collected by different modes should be of equivalent quality, consistent within the mode and mode effects on measurement error be minimised. It is also desirable to minimise discontinuity in time series, particularly on such an important survey as the LFS where key outputs are important in monitoring the economy and informing government policy and planning. These needs have to be balanced against other important considerations such as questionnaire length, respondent burden and user experience with the aim of maximising the take-up of internet mode among the sample, reducing fieldwork costs and minimising attrition and response variability across waves. It is not simply a matter of copying face-to-face/telephone mode questions into a web instrument. We therefore developed a collaborative development process.

## 2. DEVELOPMENT METHODS AND ISSUES FOR CONSIDERATION

In this presentation we will discuss our ongoing qualitative work to adapt Labour Force Survey (LFS) questions from face-to-face and telephone modes into web mode. (We will not consider coverage, sampling, selection, platform architecture or survey management).

The process of adaptation has consisted of the following steps.

1. Consideration of the relative importance of the overall objectives and different design drivers to inform question design principles and development of a research

---

<sup>1</sup> Office for National Statistics, United Kingdom

<sup>2</sup> Office for National Statistics, United Kingdom

programme. The potential drivers include: reducing data collection costs; maximising web data quality by optimising web questionnaire design regardless of other considerations; enhancing the respondent experience; and maintaining comparability by minimising mode effects between face to face, telephone and web and discontinuities to time series.

- Options ranged from unimode design where questions are as similar as possible, through to optimising for web design without being constrained by existing designs and downstream systems.
- 2. Prioritisation of the numerous (approximately 600) LFS questions. Not every question needs to or can have the same level of work committed to it, therefore questions have to be ranked by their importance to published outputs. It is only possible to take a limited number of variables through the development process at a time.
- 3. Conducting a desk review of LFS questions by data collection methodologists to identify issues and propose designs.
  - The desk review compares the face-to face/telephone question and accompanying guidance against a version copied into a web instrument with little or no adaptation. The review covers many aspects of a question, including: the question stem; the answer categories; response format (open field, radio button, check box, drop down list, look up coding frame etc); question type (open, closed etc); instructions and guidance; editing and validation checks; how to permit 'don't know' and 'refusal' answers without their prevalence increasing
  - Proposals for changes to the design are made and issues for discussion documented. An assessment is made of potential effects on data compared with interview modes.
- 4. Collaborative workshops involving data collection methodologists, the programme managers overseeing the transition to mixed mode surveys, LFS managers/subject matter experts, social survey researchers, software programmers, a web user experience expert and editing and imputation methodologists, to agree programming specifications.
  - The workshops considered the proposals and issues from the desk review and also overarching issues.
  - The collaborative approach has brought different specialisms and perspectives together. Individuals have taken issues away for further investigation or deliberation and reported back. Work has been done as a team but rigour and challenge have been brought to the table. Lessons from international work have been used. Design issues have been discussed in a systematic way. Issues have been more likely to have been identified and addressed than if individuals had been working in isolation.
- 5. Cognitive/usability testing, taking the resulting initial web instrument out to members of the public to assess the effectiveness of the design. (Not all the questions that have been through the desk review and workshop process have been cognitively tested yet).
- 6. Subsequent review of test findings, another round of cognitive/usability testing where relevant and possible, and respecification of designs, in preparation for a quantitative test later in 2015.

The work has required close collaboration across fields of expertise, a flexible and adaptive approach, learning and changing as we go, and the need to identify, manage and document numerous issues and high volumes of detail evolving over time. Wireframes of proposed designs have been created to aid everyone's understanding of the

specifications. Audit trail documents have been kept, recording initial reviews, proposed designs, workshop discussions, test findings, review discussions and final specifications.

### **3. EMERGING FINDINGS**

In the presentation we will provide a few examples of the challenges faced, evolving designs, test findings and latest specifications, such as the following.

1. Making the instrument design more visually appealing and easier to use
  - The visual design and functionality of a web questionnaire is important in gaining and keeping a respondent's attention and motivation. During the course of our development we have changed the version of the survey software used to gain more control over the visual display of the survey. We have influenced the software functionality by requesting adaptations and identifying problems to its developers. We have changed page layouts and other design features. Cognitive/usability testing has provided valuable feedback and influenced design changes.
2. Employing the advice of the web user experience expert to improve design features
  - Collaboration with a web user experience expert aided the development of the visual design and functionality. cognitive/usability testing found that aspects of the design were not being seen or used by respondents on first sight, such as help and guidance. Following his advice we changed the font sizes and weights used for different elements of the screen and improved the presentation and usability of instructions and guidance (whether always-presented or respondent-initiated).
3. Engagement with respondents
  - Cognitive testing of a previous online pilot in 2011 indicated to us how important interviewers are in explaining to respondents the importance of the data that they are providing. In our initial web design only text was provided to explain the purpose of the LFS. In response to findings from cognitive/usability testing, along with digital publishing and design experts we created an infographic that displays high level economic activity status results from the LFS. The aim was to illustrate recent statistics about the types of data collected and explain who the users of the data are. Further cognitive testing has indicated that respondents now see more worth in what they are doing and are therefore more willing to complete the survey.
4. Using respondent feedback to clarify specific points of guidance
  - LFS questions are sometimes superficially easy to understand but actually have very specific requirements in terms of what should and should not be included in the answer or at particular response options. To be effective – noticed, read, understood and followed as appropriate - guidance needs to be easily understood and not add extra cognitive burden to respondents. Cognitive testing has provided useful feedback enabling us to improve the layout and content of guidance at specific questions.
5. Breaking-up long/complex questions
  - Results from our cognitive testing illustrated that a single question is not always the best design for respondents in terms of gathering accurate data. In our first round of testing a question was asked to determine a respondent's economic activity status. Response categories were not mutually exclusive so respondents were instructed to select the first answer from the list. The guidance needed to clarify for all economic statuses was extensive. However respondents did not always see or follow the 'select the first answer that applies' instruction; some explained that they preferred to express their



opinion of their ‘primary’ status (for example ‘student’ rather than ‘employed’ even if they had a part time job), rather than follow our prescription. Therefore we split the question into two. This resulted in better quality responses, meeting data requirements while being less cognitively burdensome and allowing respondents to give answers that were more meaningful to them

6. How to allow ‘don’t know’ and ‘prefer not to say’ responses
  - In interview modes interviewers are able to record spontaneous don’t know and refusal answers without them being presented to respondents. In web mode we need to allow these responses, for equivalence, but without increasing their prevalence due to satisficing. We utilised a method whereby these options were only presented if respondents tried to skip a question without answering it. Early cognitive/usability testing indicated that respondents were unaware they could respond in this way, so we added information about it in the general instructions at the start of the survey.
7. Streamlining the questionnaire to reduce questions and avoid confusion
  - In interviewer modes some questions are ‘ask or record’, i.e. if the interviewer has clearly established an answer from what respondent said at a previous question or in conversation they do not need to ask the question. In self-completion, this kind of question can be confusing and cause frustration and annoyance at apparent repetition. For example, there is a series of questions on a respondent’s legal marital status, whether they are cohabiting with another household member and relationships between household members (which include spouse and cohabiting partner categories). For web mode we changed the order, asking relationships first, then derived legal marital status and living as couple from relationships where this was possible to do. This results in fewer questions for some respondents and avoids confusion and frustration.

#### **4. FUTURE WORK**

It is planned that a quantitative test will be conducted in 2015. This exercise - called the Alpha pilot – is limited in extent, not a full LFS. It is intended primarily to assess 1) the level of take up of an invitation to register online to take part in the survey and collect limited information about the household and its members (known as Wave 0); and 2) take up of a subsequent invitation to those who register to take part in a cut down version of Wave 1. It will include the questions that have been through the development process we have described.

Development plans for beyond the Alpha pilot are still under consideration.

#### **REFERENCES**

- [1] S. Aubrey-Smith, K. Blanke, D. F. Gravem, M. Järvensivu, V. Meertens and P. Rünz, (eds K. Blanke and P. Rünz), Report WP11 “Web Data Collection”: Testing Web Questionnaires for the Labour Force Survey in Five Countries, ESSnet Project on “Data Collection for Social Surveys Using Multiple Modes”, (2014).
- [2] M. Couper, *Designing Effective Web Surveys*, Cambridge, (2008).
- [3] R. Tourangeau, F. G. Conrad and M. Couper, *The Science of Web Surveys*, Oxford, (2013).

# Recommended Practices for the design of business surveys questionnaires

Stefania Macchia ([macchia@istat.it](mailto:macchia@istat.it))<sup>1</sup>

**Keywords:** questionnaires, CAWI, business surveys.

## 1. INTRODUCTION

ISTAT Business Statistical Portal implements a new approach for the organization and management of data collection processes. It constitutes a single entry point for web-based data collection from enterprises according to a “business-centric” perspective. It provides new integrated functions supporting respondents in several areas: survey unit management, information updating, data collection activities management, delegation facilities in filling-in questionnaires, and electronic questionnaires access. In particular, a single software tool is used for electronic questionnaires: GX (*Generalised Italian (Data) Collection System XML*), which is a generalised in-house product (it can be used for whatever theme to be treated) based on XML, so that it can represent the main survey’s contents: survey metadata, survey variables, questionnaire structure, check plan and skipping rules [1]. This new architecture implies the acceleration of the harmonization of survey questionnaires both in terms of concepts treated and features of electronic questionnaires, with the perspective of improving quality and containing the respondent burden. This aim is pursued through the definition of Recommended Practices for questionnaires design, focusing on business surveys questionnaires, self-administrated through Web technique.

## 2. METHODS

The methodological approach for the questionnaire design in the survey process has already been treated in a number of Manuals which represent standards at international level, for instance: *i)* the I’Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System [2], deliverable of ESSNet QDet project aimed at identifying Recommended Practices by specifying the requirements of the Code of Practice [3], *ii)* the Handbook on MEthodology of MODern BUSSINESS STatistics, Handbook on Methodology [4], produced by the Memobust ESSNet project, which has a chapter on questionnaire design.

All these sources have been considered, together with other international literature [5], but two criteria have inspired in defining the Recommended Practices: first of all to consider the peculiarities of business surveys questionnaires self-administrated through CAWI (Computer Assisted Web Interviewing) technique and secondly to focus on the aspects of the questionnaire design process which are strictly connected with the data capturing technique, leaving out other aspects not pertinent in this context. As a matter of fact the premise was that some of the steps of the conceptual frame of the questionnaire, like the literature search, the specification of the survey objectives and basic concepts, the decision on the data collection mode should not be treated because in this context the aim was not to plan new surveys *ex novo*, but to restructure and harmonize surveys to be included in Business Portal. So, for instance, survey concepts are already known as well as the data capturing technique already chosen.

---

<sup>1</sup> Istat – Italian National Institute of Statistics

In particular, considering the GSBPM model (Generic Statistical Business Process Model), these Recommended Practices take into consideration all the activities pertaining to the overlapping area between sub-process 2.2 (Design variable descriptions) and sub-process 2.3 (Design collection), consistently with the assumption that *'The sub-process 2.2 may need to **run in parallel** with sub-process 2.3, as the definition of the variables to be collected, and the choice of collection instrument may be inter-dependent'* [6].

Finally, experiences already gained in Istat on business surveys, which already adopted the CAWI technique, were considered. As a matter of fact, a great number of business surveys already used Web questionnaires, even if not in an integrated architecture like the Business Portal, adopting different IT solutions (PHP, Excel, ecc.) and different data capturing modes (online/offline).

In addition, in defining the Recommended Practices, it has been considered how some aspects of the design of questionnaires and of the features to be implemented in the electronic questionnaire have an impact on the steps of the response process, represented in a hybrid response model based on a cognitive approach. This model identifies four steps [7], to which other four steps, related to organisational aspects, have been added [8] [9], which are typical of business surveys. The steps of the original model are: *i)* Comprehension, *ii)* Retrieval, *iii)* Judgement and *iv)* Communication. While the frame steps are: *i)* Encoding in memory/record formation, *ii)* Selection/Identification of Respondent(s), *iii)* Assessment of Priorities (Motivation) and *iv)* Release of the Data.

Considering the different aspects of survey questionnaire, the questions wording, their structuring, the layout and all the instructions and information at disposal of respondents related to each question surely have an impact on the comprehension and retrieval processes. On the other hand, the skipping and consistency rules implemented in electronic questionnaire constitute a support for the judgement step. Finally, the electronic questionnaire design has an impact on the communication process, as it predefines the structures according to which responses have to be given.

Regarding the organisational steps, the questionnaire design certainly has an impact on the Encoding in memory/Record formation step because as more the way data are stored in the respondent data base is considered in structuring questions, as more the effort needed to provide data is contained. On the other three steps, the features implemented in the Business Portal architecture have a direct impact.

### 3. RESULTS

A manual of Recommended Practices for business surveys questionnaires, self-administrated through Web technique, has been produced. In this manual general principles which should guide in designing questionnaires and other particular ones related to the CAWI technique are shortly described. The set of recommendations provided are organised according to the specific homogeneous themes reported hereafter.

#### *i)* Design of questionnaire sections/Web pages

One step of the questionnaire design concerns its structuring in sections in which questions attaining to specific issues are grouped. Suggestions are given on how designing the sections, taking into considerations the Web pages, which in turn, can be composed by more than one computer screen-shots.

#### *ii)* Questions design

Questions design is composed by two steps: definition of questions wording and identification of the structure to be used to provide the answer. The recommendations regard this second aspect above all, providing suggestions on quantitative questions (how to manage decimal values, scale factors, measure units, etc.), closed questions (when using radio-buttons, drop-down-boxes, check-

boxes), free texts questions to be coded (when it is preferred to use assisted coding functions) and, finally, questions organised in tables/arrays.

*iii) Use of classifications*

Classifications can be used during data collections for different purposes: to custom the questionnaire for the different respondents (e.g. classifications concerning municipalities/countries are linked to display textual descriptions corresponding to codes) or to classify textual responses. In this last case, suggestions are given on different algorithms for textual matching which can be used.

*iv) The structure for variables definitions*

Harmonising the definitions in terms of their content is surely the first problem to be faced, even if the content often depends on European Regulations. On the other hand, how this content is expressed and structured has in impact on the comprehension process as well. For this reason a modular structure, suggested as a standard for definitions of all variables, has been defined: the main module contains the real meaning, which should be expressed synthetically, with simple and not ambiguous words, while other modules contain inclusion/exclusion clauses, peculiarities of surveys or references to national laws.

*v) The management of skipping rules*

The automatic management of skipping rules represents a strength of data collection computer assisted techniques as it prevents from this type of errors and helps in guiding respondents in filling in questionnaires. Concerning business surveys questionnaires, different suggestions are given for skipping rules concerning questions belonging to the same section/web page or to different section/web pages. While for the first ones the management can be automated, for the second ones, it is often to be preferred not to do that because it is not unusual that different persons fill in each section due to the fact that the themes treated request a specific knowledge.

*vi) The checking plan*

Suggestions are given on the management of the consistency rules, taking into consideration different aspects, like: rules regarding variables of the same web-page or different ones, number of checking rules on a single web-page, relation between skipping and checking rules, etc.

*vii) The screen design*

Concerning this aspect, standards already defined for GX in the Business Portal architecture have been adopted. Some details have been provided on graphic symbols to be used to link definitions, instructions, helps, etc.

*viii) The guide and instructions for respondents*

A set of documents is always provided to respondents (e.g. brief description of the survey, normative aspect regarding confidentiality, compilation guide, etc.). Common structures, templates and access modes have been designed for these documents, as they can have an impact on the comprehension process as well.

*ix) Questions to estimate the perceived respondent burden*

The migration of surveys questionnaires in GX in the Business Statistical Portal architecture has been considered a good occasion to collect information on respondent burden directly from respondents. For this purpose a set of questions has been proposed, which distinguish possible burden due to difficulty in retrieving the requested information and in filling in the questionnaire. The responses to these questions could be analysed together with other indications derivable from paradata.

*x) The use of paradata for data collection monitoring and optimising questionnaire design.*

Paradata can be analysed for two different purposes: to monitor the data collections process and to have indications on problems in filling in the questionnaire, useful to optimise its design. The Business Portal already produces some indicators for the first purpose (e.g. the status of questionnaires, the status of consistency rules, etc.). A set of other indicators have been identified which can be used by the researcher to optimise the questionnaire design, specifically when some 'alerts' are highlighted by the answers given to questions on the respondent burden (e.g. number of entries for web-pages, connection time for the entire questionnaire or for each web-page, length of navigation routings, etc.).

#### 4. CONCLUSIONS

These Recommended Practices guarantee a uniform way of communication between Istat and survey respondents not only thanks to a common questionnaires layout, but mostly through the same features, the standard structures for definitions and the homogeneous documentation of the concepts treated. These represent a tool, together with the harmonisation of variables definitions, to have a common base which allows to reduce redundancy and to design integrated surveys on different but similar statistical domains.

#### REFERENCES

- [1] N.R. Fazio, M. Murgia and A. Nunnari, The Business Statistical Portal: a new way of organising and managing data collection process for business surveys in ISTAT, UNECE New frontiers for Statistical data Collection (2013)
- [2] Eurostat, Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System,  
[http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook\\_of\\_Practices\\_for\\_QUEST.pdf](http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook_of_Practices_for_QUEST.pdf) (2006)
- [3] Eurostat, European Statistics Code of practice <http://epp.eurostat.ec.europa.eu/>, (2011)
- [4] Eurostat, Memobust handbook - Handbook on Methodology of Modern Business Statistics, <http://www.cros-portal.eu/content/memobust> (2014)
- [5] J.Bethlehem, S.Biffignandi, Handbook of Web Surveys, eds Wiley (2011)
- [6] Eurostat, GSBPM v.5 <http://www1.unece.org/stat/platform/display/GSBPM/> (2013)
- [7] R. Tourangeau, R., Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington, DC: National Academy Press (1984), pp. 73–100
- [8] S. Sudman, D. K. Willimack, E. Nichols and T.L.Mesenbourg, Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies, Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association (2000)
- [9] D.K. Willimack, E. Nichols, Building an alternative Response Process model for Business Survey, Proceeding of the annual meeting of the American Statistical Association (2001).

# Interviewer Effects in Real and Falsified Interviews

## - Results from a Large Scale Experiment -

Peter Winker (Peter.Winker@wirtschaft.uni-giessen.de)<sup>1</sup>, Karl-Wilhelm Kruse<sup>1</sup>, Natalja Menold (natalja.menold@gesis.org)<sup>2</sup>, Uta Landrock (uta.landrock@gesis.org)<sup>2</sup>

**Keywords:** Interviewer effects; Interviewer falsifications; Indicators for falsifications

### 1. INTRODUCTION

Interviewers influence data quality in surveys unintentionally or intentionally (see [1] and [2]). Interviewers' motivation to produce accurate data might be affected by demoralisers linked to questionnaire design and administrative procedures [1]. In [3], interviewers' burden and their satisficing are discussed. In addition, psychological theories may explain interviewers' effects. One such theory provides an explanation for goal-directed human behaviour (motivation) and combines important factors which can explain the interviewers' motivation to produce correct survey data [4].

Based on these theories, one can expect an effect of work conditions – such as payment – on the quality of the data obtained. Furthermore, interviewers' ability, personal factors and skills may affect their motivation and persistency on a task resulting in an additional effect on survey data. Empirical evidence can be obtained from the research on interviewers' characteristics, which found, e.g., gender to impact data collection [5]. It should also be mentioned, however, that many authors did not find any association between interviewer characteristics and respondents' responses [6]. The effect of interviewers on the survey data becomes more pronounced if they decide to fabricate complete interviews or some parts of them [7].

We analyse influences of interviewers' characteristics and payment schemes on falsified and real data making use of a unique experimental dataset described in more detailed in Section 2. To the best of our knowledge, studies in which payment of interviewers has been the subject of variation are hardly available. Section 3 presents results of a detailed empirical analysis of interviewer effects both in real and falsified data. A summary of the findings as well as conclusions for survey practice are provided in the final Section 4.

### 2. METHODS

#### 2.1. Experimental data

A unique experimental dataset is used. It contains information on interviewers and the interviews. First, real face-to-face interviews were conducted, and afterwards the interviewers generated falsified data. The interviewers also filled in the questionnaire of the survey themselves as well as a further short questionnaire regarding their attitudes and strategies when generating the falsified interviews. Two settings for payment have been used – payment per hour and payment per interview. This allows conclusions regarding the impact of payment and interviewers' characteristics on data quality.

---

<sup>1</sup> Justus-Liebig-University Giessen, Center for International Development and Environmental Research (ZEU), Giessen, Germany

<sup>2</sup> GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Details of the experiment are reported in [8,9]. 78 interviewers (students) collected first about 700 real interviews from students of the same university. Audio recording was used to ensure that only real interviews are considered when real data are analysed. In a second step, the same interviewers were asked to fabricate survey data. To this end, each interviewer received personal descriptions of real survey participants interviewed by his/her colleagues. These descriptions contained some basic characteristics, e.g., gender, age, subject of study, number of semesters enrolled, marital status, residence and country of origin. The students, now acting as “falsifiers”, were instructed to imagine conducting a face-to-face interview with a real person with those characteristics and then to fill in the survey questionnaire.

## 2.2. Payment scheme and interviewer characteristics

To disentangle the effect of specific characteristics of interviewers or interviews, we focus on specific sub-groups. In particular, we concentrate on the effects of the payment scheme, interviewers’ gender, and a split according to interviewers’ optimism about future economic development. Within each of these subgroups, we further differentiated between actual interviews and those fabricated by the interviewers.

The first subgroup considered is defined by the payment scheme (per interview or per hour). Half of the real interviews and half of the falsifications were conducted with a fixed payment per interview, while the other half of both real and falsified interviews received a payment per hour. The groups differed only in the instruction on payment (per completed interview versus per hour) and not in any other aspect. A second split of the interviewers is made with regard to gender as several results in the literature point at a gender effect in real face-to-face interviews [5]. In addition, we are interested to see whether gender of the person producing falsified data also has an impact on the properties of these falsifications. Finally, we consider subgroups defined by the degree of optimism. To this end, a question about the assessment of the economic development of Germany with a 5-point-scale was used to split both interviewers and respondents into optimists (top 2 categories) and pessimists (bottom 3 categories).

## 2.3. Meta-Indicators

As our interest is in general interviewer effects, we focus on a set of meta-indicators of the data collected by each interviewer. Such meta-indicators have been used in previous research to identify potential falsifications [7,8,9]. We will only present a few of them here due to space constraints. In addition, we also have at our disposal data on the duration of the interviews.

The meta-indicators considered are: **1. Acquiescent responding style (ARS)**: This indicator measures how often the respondents agree to the interviewer without really thinking about what they were asked; **2. Participation**: In contrast to most other meta-indicators, this indicator requires a question about the past political activities of the respondent. She/he is asked to check all options that apply; **3. Non-differentiation (ND)**: For this indicator, first the standard deviation (sd) of each appropriate multi-item question in the questionnaire was calculated. ND is obtained as the average of  $(1-sd/10)$  over all multi-items considered; and **4. Semi-open**: The indicator value is given by the relative frequency of choosing the „others, please specify“-option across all semi-open-ended questions in the questionnaire.

### 3. RESULTS

We report some of the most striking differences found between subgroups of interviewers and between real and falsified interviewers in the following. Table 1 summarizes the findings regarding **interview duration**. It reports the descriptive statistics for interview duration both for actual interviews (“Real interviews”) and for interviews fabricated in the laboratory (“Falsified interviews”). The information is provided both for the full groups and for subgroups according to payment scheme.

**Table 1.** Duration of real and falsified interviews (in minutes)

	N	min	max	mean	std.dev.
Real interviews	696	15	134	33.99	9.66
- payment per hour	328	18	134	34.80	
- payment per interview	368	15	87	33.28	
Falsified interviews	690	2	132	15.59	8.16
- payment per hour	342	1	79	16.54	
- payment per interview	348	1	132	14.66	

The duration of real interviews in mean is slightly longer than half an hour (33.99 minutes) with a substantial variation (std.dev. 9.66), while falsifications are done much faster (about 50%). It is found that both real and falsified interviews take more time when payment is per hour. In both cases, the difference is statistically significant (5% level). For real interviews this might translate into improved data quality if payment is per hour.

For the meta-indicator of **acquiescent responding style** (ARS), we find interesting differences both in real interviews and between real and falsified interviews. For the real interviews, we find a significantly (at 1%-level) higher value if the interviewer is male. Taking also into account the respondent’s gender, we find relevant effects of interviewer respondent pairing as in [5]. For the falsified interviews, we find that both women and men as falsifiers underestimate the extent of acquiescence.

According to [8], falsifiers report less political activity (**Participation**) of respondents than real respondents. We also find a significant difference for the real interviews according to the payment scheme. If the interviewer is paid per hour, the frequency of reported political activity increases as compared to a situation when payment is per interview (difference significant at the 1%-level). This finding provides strong evidence that the payment scheme might affect the actually collected data. We also spot a significant (1% level) impact of the attitude of the falsifiers.

For **non-differentiation** we find a significant (5%-level) impact of the interviewer’s gender on the outcome in real interviews. A female interviewer seems to reduce the non-differentiation. The effect does not depend in a significant way on the gender of the respondent. The gender of the interviewer for falsified interviews does not affect the degree of non-differentiation. When focusing on falsified interviews, we find a highly significant impact of the payment scheme. Payment per interview seems to decrease non-differentiation as compared to payment per hour.

The frequency of choosing the “others, please specify” option in **semi-open** questions is significantly (1%-level) lower for falsified interviews independently of the payment scheme. This effect is weaker in male interviewers than in women. We also find a significant (5%) difference for real interviews when considering the subgroups of optimistic and pessimistic interviewers. Pessimistic interviewers are more likely to gather “other” answers indicating effects of interviewers’ characteristics to be quite subtle.



#### 4. CONCLUSIONS

Based on a unique experimental data set, properties of data collected by interviewers or fabricated by them are studied. Significant differences are found both with regard to real and falsified data depending on the payment scheme and characteristics of the interviewer. As a first conclusion for survey practice, it is recommended to collect as much information about interviewer characteristics as possible. This might be used in later analysis to control for interviewer effects. It can also be used to improve data driven methods for the identification of falsifications. A second conclusion concerns payment schemes. As expected, the kind of remuneration for interviewers might have a significant impact on data quality. Apart from anecdotal evidence that payment per conducted interview might result in changing the respondent in case the contact person is not available, it might have a direct influence on the data collected. Finally, the observed heterogeneity of meta-indicator values across subgroups of interviewers both for real and, in particular, for falsified interviews suggests considering cluster procedures allowing for more than one cluster of falsifiers. This is on the agenda of our future research as well as a more detailed analysis of the reasons for the observed differences and a theory based selection of potentially relevant characteristics.

#### REFERENCES

- [1] L.P. Crespi, The cheater problem in polling, *Public Opinion Quarterly* 9/4 (1945), 431-445.
- [2] B.T. West, F. Kreuter and U. Jaenichen, "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics* 29/2 (2013), 277–297.
- [3] L. Japiec, Quality issues in interview surveys, Some contributions. *Bulletin of Sociological Methodology* 90/1 (2006), 26-42.
- [4] J.W. Brehm and E.A. Self, The intensity of motivation, *Annual Review of Psychology* 40 (1989), 109–131.
- [5] L.M. Matikka and H.T. Vesala, Acquiescence in quality-of-life interviews with adults who have mental retardation, *Mental Retardation* 35/2 (1997), 75-82.
- [6] J.J. Hox, E.D. de Leeuw and I.G.G. Kreft, The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In: *Measurement Errors in Surveys* (P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, Eds.) Wiley, Hoboken, NJ (1991), 439- 461.
- [7] S. Bredl, P. Winker and K. Kötschau, A statistical approach to detect interviewer falsification of survey data. *Survey Methodology* 38/1 (2012), 1–10.
- [8] N. Menold, P. Winker, N. Storfinger and C.J. Kemper, A method for ex-post identification of falsifications in survey data. In: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention* (P. Winker, N. Menold and R. Porst, Eds.) Peter Lang. Frankfurt (2013), 25-47.
- [9] S. De Haas, S. and P. Winker, Identification of partial falsifications in survey data, *Statistical Journal of the IAOS* 30 (2014), 271-281.

# User-friendly framework for metadata and microdata documentation based on international standards and PCBS Experience

Haitham Zeidan ([Haitham@pcbs.gov.ps](mailto:Haitham@pcbs.gov.ps))<sup>1</sup>, Geoffrey Greenwell ([geoffrey.greenwell@oecd.org](mailto:geoffrey.greenwell@oecd.org))<sup>2</sup>

**Keywords:** Metadata, Microdata, DDI, DCMI, SDMX, RDF, XML, Semantic web, NADA, PCBS, OECD.

## ABSTRACT

This paper discussed and investigated the experience of Palestinian Central Bureau of Statistics (PCBS) [1] in designing documentation model and user-friendly framework for metadata and microdata documentation. PCBS uses two metadata specifications: the Data Documentation Initiative (DDI) [2] and the Dublin Core Metadata Initiative (DCMI) [3]. Both are defined in the Extensible Mark-up Language (XML) and the Resource Description Framework (RDF). This paper focused also on the DDI and DCMI as well as its relationship to other relevant metadata standards (e.g., The Statistical Data and Metadata Exchange (SDMX)) [4] and the semantic web technologies, we addressed the features of these standards as Richer content, Coverage, On-line analytical capability, Search capability and Interoperability since these standards are defined in the Extensible Mark-up Language (XML).

## 1. INTRODUCTION

Good documentation has a number of features. It should accurately describe the data. The information should be clear so that the data are not incorrectly used. It should also be comprehensive, so that the statistical agency is not dependent on the institutional memory of staff. A basic principle is that all information that can foster the effective and accurate use of datasets by secondary users should be preserved and disseminated.

Unfortunately, documentation is often the last step of the survey process, and it is then often too late to capture all metadata produced during the life cycle of the survey activities. This results in the loss of useful information generated at early stages, such as the comments received from various stakeholders at the stage of questionnaire design, problems encountered during pilot-testing of the questionnaire, etc. Treating documentation as an ongoing part of survey activity will reduce the documentation costs and increase its quality. Using the international metadata standards, such as the Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI) specifications, can reduce the burden considerably, because they provide a rigorous framework for organizing the process and will help to address the technical issues related to documentation, preservation and dissemination process of the surveys, in addition to improve management and use of microdata.

## 2. OBJECTIVES

The objectives of this study is to display the user-friendly framework for metadata and microdata documentation that introduced in Palestinian central bureau of statistics (PCBS) for better documenting, preserving, anonymizing and disseminating of existing

---

<sup>1</sup> Palestinian Central Bureau of Statistics (PCBS)

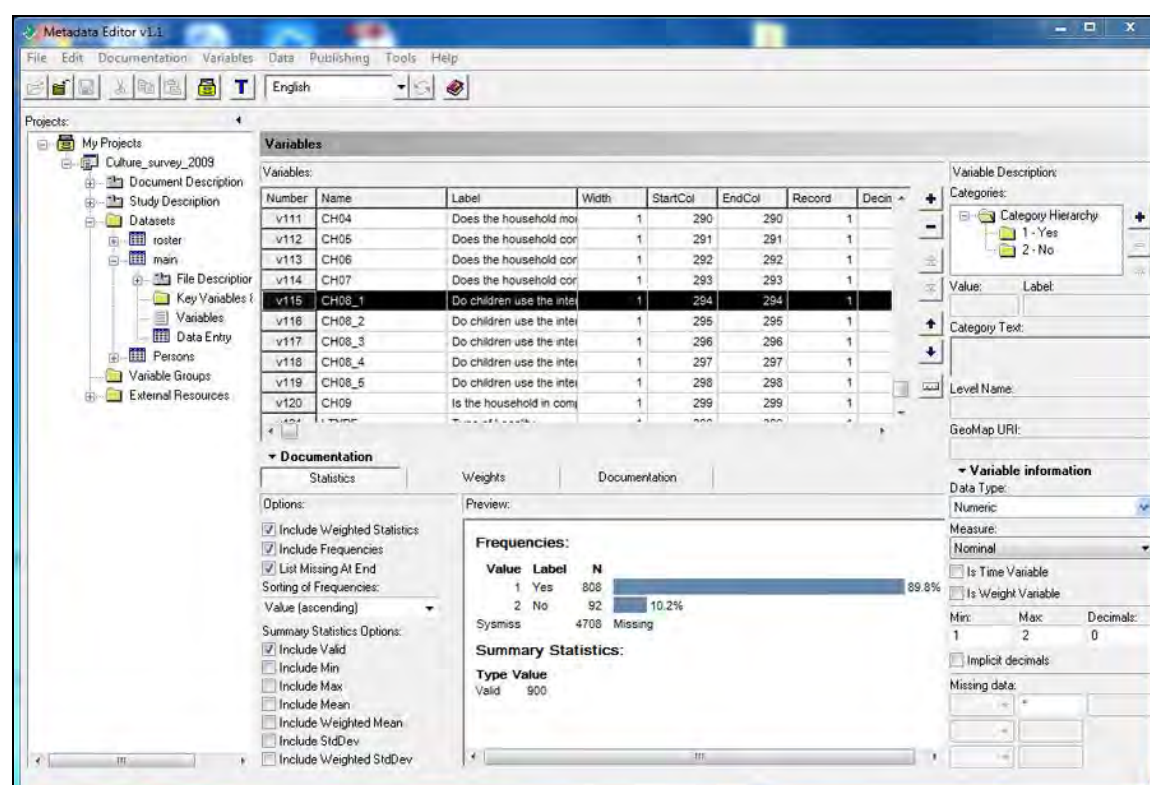
<sup>2</sup> The Organization for Economic Cooperation and Development (OECD)

microdata, this framework produced based on international standards: Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI).

### 3. USER-FRIENDLY FRAMEWORK FOR METADATA AND MICRODATA DOCUMENTATION

Our framework produced based on international standards: Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI). We use in PCBS The IHSN Metadata Editor, also known as the Nesstar Publisher [5] as shown in figure (1), which is a rich editor for the preparation of metadata and data for publishing in an online catalog, such as the IHSN-developed National Data Archive (NADA) [6]. The metadata produced by the Editor is compliant with the Data Documentation Initiative (DDI) and the Dublin Core XML metadata standards. The application is developed by Nesstar at the Norwegian Social Science Data Archive (NSD) and is distributed as freeware. The features of metadata editor are:

- Easy editing/creation and export of DDI documented datasets with XML experience needed.
- Tools to validate metadata and variables.
- The ability to include automatically generated frequency and summary statistics for each variable.
- Tools to compute/recode/label new, or existing, variables to be added to a dataset before publishing.
- The ability to import and export data to the most common statistical formats, including delimited files.
- Multilingual - Arabic, Chinese, English, French, Portuguese, Russian and Spanish.



## **Figure 1. The Features of Metadata Editor**

### **3.1. Data Documentation Initiative (DDI)**

The DDI developed standards that provide a structured framework for organising the content, presentation, transfer and preservation of metadata in the social and behavioural sciences. It enables documenting even the most complex microdata files in a way simultaneously flexible and rigorous.

The DDI seeks to establish an international XML based standard for microdata documentation. Its aim is to provide a straightforward means of recording and communicating to others all the salient characteristics of micro-datasets. The DDI specification is a major transformation of the once-familiar electronic ‘codebook’: it retains the same set of capabilities but greatly increases the scope and rigour of the information contained therein.

#### **3.1.1 DDI Features**

**Interoperability:** DDI-compliant documentation can be exchanged and transported seamlessly, and applications can be generically written, because the documents are homogeneous.

**Richer content:** The DDI provides data analysts with broader knowledge about data content, because the DDI initiative provides a comprehensive set of elements that can describe micro-datasets as completely and as thoroughly as possible.

**Multipurpose documentation:** A DDI codebook can be restructured to suit different applications, because it contains all the information necessary to produce different types of output.

**On-line analytical capability:** DDI documents can be easily imported into on-line analysis systems, rendering datasets more readily usable by a wider audience. This is made possible because the DDI mark-up extends down to the variable level and provides a standard uniform structure and content for variables.

**Search capability:** Field-specific searches across documents and studies are made possible, because each of the elements in a DDI-compliant codebook is tagged in a specific way.

#### **3.1.2 DDI Coverage**

The DDI specification has been designed to fully encompass the kinds of data generated by surveys, censuses, administrative records, experiments, direct observation, and other systematic methodologies for generating empirical measurements. In other words, the unit of analysis could be individual persons, households, families, business establishments, transactions, countries, or other subjects of scientific interest. Similarly, observations may consist of measures taken at a single point in time in a single setting, such as a sample of people in one country during one week, or they may consist of repeated observations in multiple settings, including longitudinal and repeated cross-sectional data from many countries, as well as time series of aggregate data. The DDI specification also provides for full descriptions of the methodology of the study (mode of data collection, sampling methods if applicable, universe, geographical areas of study, responsible organization and persons, and so on).

### 3.1.3 DDI Structure

The DDI specification permits all aspects of a survey to be described in detail: the methodology, responsibilities, files and variables. It provides a structured and comprehensive list of hundreds of elements and attributes that may be used to document a dataset, although it is unlikely that any one study would use all of them. However, some elements, such as “Title,” are mandatory (and must be unique). Other elements are optional and can be repeated, for example “Authoring Entity/Primary Investigator”, since it includes information on the person(s) and/or organization(s) responsible for the survey.

The DDI elements are organized in five sections:

Section 1.0: Document Description: A study (survey, census or other) is not always documented and disseminated by the same agency as the one that produced the data. It is therefore important to provide information (metadata) not only on the study itself, but also on the documentation process. The Document Description consists of overview information describing the DDI-compliant XML document, or, in other words, “metadata about the metadata”.

Section 2.0: Study (Survey) Description: The Study Description consists of overview information about the study. This section includes information about how the study should be cited, who collected, compiled and distributes the data, a summary (abstract) of the content of the data, information on data collection methods and processing, and so on.

Section 3.0: Data File Description: This section is used to describe each data file (Microdata) in terms of content, record and variable counts, version, producer, and so on.

Section 4.0: Variable Description: This section presents detailed information on each variable, including literal question text, universe, variable and value labels, derivation and imputation methods, and so on.

Section 5.0: Other Material: This section allows for the description of other materials related to the study or survey. These can include resources such as documents (questionnaires, coding information, technical and analytical reports, interviewer's manuals, and so on), data processing and analysis programs, photos, and maps. However, the Dublin Core Metadata Initiative (described below) is better suited for the framework requirements.

### 3.2. The Dublin Core Metadata Initiative (DCMI)

The DCMI Metadata Element Set (ISO standard 15836), also known as the Dublin Core metadata standard, is a simple set of elements for describing digital resources. This standard is particularly useful to describe resources related to microdata such as questionnaires, reports, manuals, data processing scripts and programs, etc. It was initiated in 1995 by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) at a workshop in Dublin, Ohio. Over the years it has become the most widely used standard for describing digital resources on the Web and was approved as an ISO standard in 2003. The standard is maintained and further developed by the Dublin Core Metadata Initiative - an international organization dedicated to the promotion of interoperable metadata standards.

### 3.2.1 DCMI Elements

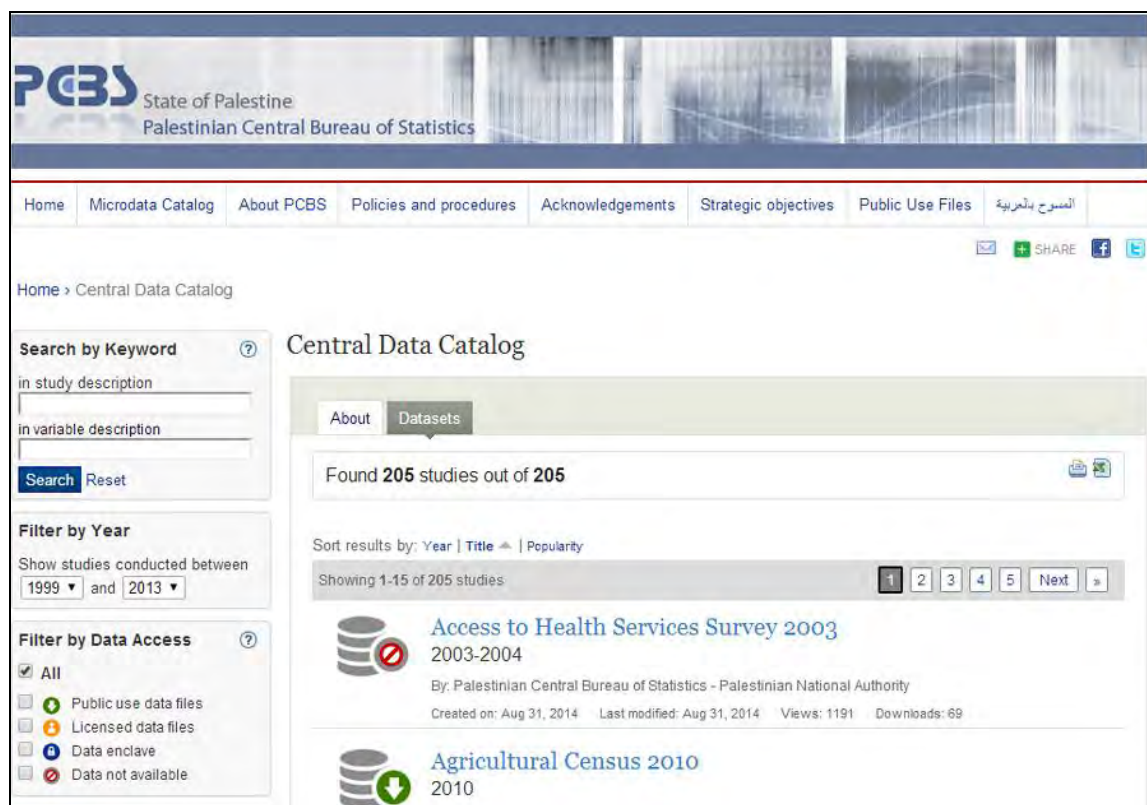
The Dublin Core metadata standard is based on the same principles as the DDI specification. It consists of a set of elements (or “tags”), organized to form an XML file. The Dublin Core standard includes two levels: Simple and Qualified. In the framework, only the Simple Dublin Core elements are used. They include the following fifteen elements as shown in Table (1) below:

**Table 1. The Dublin Core Metadata Initiative (DCMI) Elements**

<b>Element</b>	<b>Details</b>
Title	The name by which the resource is formally known.
Subject	The topic of the resource.
Description	An abstract, a table of contents, or a free-text account of the content.
Type	The nature or genre of the content of the resource (e.g., a survey questionnaire, a data processing syntax program, a map).
Source	A reference to a resource (e.g., a PDF filename, or a website URL).
Relation	A reference to a related resource (this element will rarely be used).
Coverage	The extent or scope of the content of the resource. Coverage will typically include spatial location (e.g., a country), or a temporal period (a date or date range).
Creator	The person(s), organization(s), or service(s) responsible for making the content of the resource.
Publisher	The person(s), organization(s), or service(s) responsible for making the resource available.
Contributor	The person(s), organization(s), or service(s) having contributed to the content of the resource.
Rights	A rights management statement for the resource.
Date	A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource.
Format	For use in determining the software, hardware or other equipment needed to display or operate the resource (e.g., “STATA Version 8”; or “MS-Excel 2000”).
Identifier	An unambiguous reference to the resource within a given context. Examples of formal identification systems include the Uniform Resource Locator (URL), and the International Standard Book Number (ISBN).
Language	A language of the intellectual content of the resource.

### 3.3. Dissemination Surveys Using National Data Archive (NADA)

After documentation process we in PCBS disseminate the surveys on National Data Archive (NADA) portal as shown in figure (2), NADA is a web-based cataloging application that allows for the creation of portals that allows users to browse, search, compare, apply for access, and download relevant census or survey information. It was originally developed to support the establishment of national survey data archives. The application is used by a diverse and growing number of national, regional, and international organizations. NADA, uses the Data Documentation Initiative (DDI), XML-based international metadata standard.



**Figure 2. National Data Archive (NADA) portal used to disseminate documented surveys.**

#### **4. DDI AND SDMX STANDARDS**

Recently, two technical standards for statistical and research data and metadata have been receiving much attention. Particularly for those working with both micro-data and time-series aggregates, there can be some confusion as to the relationship between these standards, and questions about which may be more appropriate for use in a particular application or institution. We describes the basic scope of each standard.

The Statistical Data and Metadata Exchange (SDMX) technical specifications come out of the world of official statistics and aim to foster standards for the exchange of statistical information. They have been created by the Statistical Data and Metadata Exchange Initiative. The initiative is a cooperative effort between seven international organizations: the Bank for International Settlement (BIS), the International Monetary Fund (IMF), the European Central Bank (ECB), Eurostat, the World Bank (WB), the Organization for Economic Cooperation and Development (OECD), and the United Nations Statistical Division (UNSD). The output of this initiative is not just the technical standards, but also addresses the harmonization of terms, classifications, and concepts which are broadly used in the realm of aggregate statistics.

The Data Documentation Initiative (DDI) is a specification for capturing metadata about social science data. It is maintained by the Data Documentation Initiative Alliance, a membership-driven consortium including universities, data archives, and national and international organizations. The specification was originally created to capture the information found in survey codebooks, which remains the focus of the first two versions. The DDI 3.0 version covers the whole data lifecycle, from the survey instrument design to archiving, dissemination and repurposing, allowing for a description of re-codes, processing, and comparison of studies by design or after-the-fact [7].



SDMX and the latest version of the DDI have been intentionally designed to align themselves with each other as well as with other metadata standards. Because much of the micro-data described by DDI instances is aggregated into the higher level data sets found at the time-series level, it is not surprising that the two have been designed to work well together. Although there is some overlap in their descriptive capacity, they can best be characterized as complementary, rather than competing [8].

#### **4.1. DDI/SDMX Overlap**

SDMX provides XML formats for describing data and independent metadata structures, which can be user-configured to hold any concepts desired. They also provide XML formats based on these configurations. The concept of exchanging a data set or a metadata set is the primary focus in SDMX, which is optimized for the exchange of aggregate data. The typical case is the exchange of time series data.

DDI also provides the ability to describe a rich set of metadata in an XML format, with an emphasis on micro-data, but also allowing for tabular formats and multidimensional cubes. In the 3.0 version, DDI supports all phases of the lifecycle from a description of concepts and the survey instrument used to collect data to the end product held in a data archive and used for analysis. DDI 3.0 also provides an XML format for micro-data and tabular/multi-dimensional data, but very often the data is held in text or statistical software specific binary files. The user-configurable aspects of DDI ("variables") are mixed with specific metadata fields.

These two standards are well aligned means that they can be combined in powerful ways, and that users of the two standards can move data from one standard format to the other fairly easily.

### **5. CONCLUSION AND FUTURE WORK**

This research aimed to introduce and to display the user-friendly framework for metadata and microdata documentation in PCBS based on international standards DDI and DCMI, the features of these standards and relations with other standards like SDMX. We introduced also dissemination surveys using national data archive (NADA).

Future work includes extending our framework and using it in other ministries and agencies to build centralized nada portal for all documented surveys, this will enhance and support the national statistical system (NSS) and the national strategy and will improve the documentation and dissemination policy.

#### **REFERENCES**

- [1] Palestinian Central Bureau of Statistics (PCBS): [http://pcbs.gov.ps/site/lang\\_en/1/default.aspx](http://pcbs.gov.ps/site/lang_en/1/default.aspx)
- [2] Data Documentation Initiative (DDI): Available at: <http://www.ddialliance.org/>
- [3] Dublin Core Metadata Initiative (DCMI): Available at: <http://dublincore.org/>
- [4] The Statistical Data and Metadata Exchange (SDMX): Available at: <http://sdmx.org/>
- [5] The IHSN Metadata Editor, also known as the Nesstar Publisher: Available at: <http://www.ihsn.org/home/software/ddi-metadata-editor>.



- [6] IHSN-developed National Data Archive (NADA): Available at:  
<http://www.ihsn.org/home/software/nada>.
- [7] M. Vardigan, P. Heus, W. Thomas, Data Documentation Initiative: Toward a Standard for the Social Sciences, *The International Journal of Digital Curation*. ISSN: 1746-8256 (2008), Vol. 3, No. 1, pp. 107-113.
- [8] A. Gregory, P. Heus, DDI and SDMX: Complementary, Not Competing, Standards, Open Data Foundation (2007).

# Sampling design data file

Seppo Laaksonen ([Seppo.Laaksonen@Helsinki.Fi](mailto:Seppo.Laaksonen@Helsinki.Fi))<sup>1</sup>

**Keywords:** Auxiliary variables, calibration, data quality, fieldwork, inclusion probability, non-response.

**Abstract:** The paper first determines the term ‘sampling design file’ that is not commonly used in survey sampling literature. The methodology behind this term is, of course, used to some extent, but only implicitly. Its explicit determination facilitates many things in survey practice and also gives a clear target for one big part of a survey, that is, sampling, fieldwork and finally for estimation. The sampling design file consists of all the gross sample units and its variables include those that give opportunity to create sampling weights, to analyse the survey quality, and to estimate. The file is possible to complete after the fieldwork. Its most important characteristics, including sampling design variables and weights, will be finally merged together with the real survey variables at respondent level, and then the survey analysis is ready to begin.

## 1. INTRODUCTION

The term ‘sampling file’ or more broadly ‘sampling design data file’ is rarely used in standard survey literature. One of this first users is the sampling expert panel of the European Social Survey (ESS) that was established in 2001 (see more information about this survey that initially started in 2002, [europeansocialsurvey.org](http://europeansocialsurvey.org)). The document of the panel says: “The Sampling design data file (SDDF) is routinely generated by an ESS country’s National Coordinator after fieldwork has finished. It includes information on the implemented sample design such as inclusion probabilities and clustering. As such, it serves the sampling team with the data required for computation of design weights, design effects and as a general basis for benchmarking the quality of sampling. The ESS analyst may use it for several purposes such as incorporating cluster information in her/his analyses.”

A SDDF is required for all types of surveys, thus for surveys from households, individuals, businesses and corporations. Here we concentrate on surveys of individuals who are members of households.

## 2. BASIC TARGETS OF THE SAMPLING FILE

The statistical units of the sampling file should ideally cover all the gross sample units of the survey. Such units are selected addresses in the case of address-based samples (but there are individuals behind these addresses or dwelling units), and selected individuals in the case of individual-based samples. In the end, the file of these statistical units thus covers the respondents, the non-respondents and the in-eligibles. It might be difficult to completely numerate in-eligibles for the file, since any contact for some individuals/addresses cannot be made and hence the file may be inaccurate, but all efforts to complete the file with appropriate information should be done. It follows that such a unit may thus be either an in-eligible or a non-respondent. Correspondingly, some bias in estimates necessarily follows.

---

<sup>1</sup> Affiliation

The first-order sampling file is good to create while the gross sample has been drawn. In this case, the file includes:

- Non-confidential and confidential identifier
- Sampling frame variables and respective statistics
- Stratification variables, explicit strata in particular
- Implicit strata if they include useful information; implicit stratification specifies the order of the units selected by equidistance or other systematic selection but basically this design corresponds to a simple random selection.
- Inclusion probabilities of each stage within explicit strata.

In the case of multi-stage sampling, all inclusion probabilities are not maybe available before the end of the fieldwork. This is typical in a three-stage sampling if the primary sampling units (PSU) are small-areas and the secondary sampling units (SSU), respectively, are addresses or households, but the third stage units are individuals. This missingness for the third stage units is due to the problem to contact a dwelling unit or an address in order to know how many target population members there exists. Even in register countries such information is hard correctly to get, since the register is not up-to-date for a survey period.

It is possible and also useful to calculate the gross sample design weights immediately when the first-order file is available. This gives opportunity to check basic figures and the quality of the sampled file of this phase. For example, when summing up these design weights we should obtain the correct target population statistics that represent the final target population if no missingness occur. At contrast, if the third-stage units, for instance, are missing, the target population of the households or addresses only can be computed.

The above variables derived from a sampling frame are minimal requirements but not sufficient. It is rational at the same occasion to download other useful information for the sampling file from the same frame that we here call the second-order sampling file. For example, in register countries, the sampling frame has been created from the population register, that is up-to-date reasonably. The sampling design only requires aggregate population statistics by large region, age group and gender, for example. But the same information can be matched at micro level into gross sample units too. In addition, the same data source consists of many other variables that are beneficial to download to the second-order sampling file at the same time since it is basically free of charge. It is not common even in Finland to distend over the minimum although it is possible easily to increase the file with the following auxiliary variables, among others: marital status, year of marriage, multi-marriage, number of children, house size, type of house, citizenship, mother tongue, coordinates of the house and municipality at birth.

The second-order sampling file can further be completed from the other sources at the same time as the first-order file has been created. This usually may require some additional administration and paper work but it is best to do as soon as possible since the data sources cannot be long up-to-date, or even some data are destroyed. Typical other sources are: formal education, tax register information on income and wealth, jobseekers' register. Section 4 presents a Finnish example on this issue in more details.

The third-order sampling file can be created as soon as the fieldwork has been completed. In this case, the most important new variable is the outcome of the fieldwork that indicates who is a unit respondent, and who is a non-respondent and an in-eligible, respectively. As said above, the two last categories are often hard to determine definitely correctly. This seems to be a worsening problem in Europe due to more or less permanent absence of the official address (home). A reason for this is working outside the country over several months, or using a second home in another country, respectively.

A drawback, in many countries as said already above, is that all inclusion probabilities cannot be known after the fieldwork. In the ESS, this is concerned the selection of one individual within the selected household or address. As a consequence, it is not possible to calculate a complete inclusion probability for the individuals of the gross sample, but only for the second stage address/household. The sampling weight for the respondent can be, nevertheless, calculated, assuming, for example, that the response mechanism for the third stage is ignorable within strata.

After the fieldwork, the sampling file can be further reinforced with other data on fieldwork. Opportunities for that are depending also on the survey mode used. Face-to-face interviewers can collect information about the quality of the location where a potential respondent lives. For example, an interviewer can classify the quality of the living area or the type of house. This indicator is of course useful only if a valid measurement is available and the same information is available both for the respondents and for the non-respondents. Moreover, the interviewer information (e.g. their basic characteristics, attitudes toward this survey) can be added to the sampling file too.

### **3. WHAT TO DO WITH THE SAMPLING FILE?**

The sampling file is necessary in order to calculate the sampling weights for the respondents. This thus requires that the inclusion probabilities are available in the file. Naturally, the identifiers of the respondents should be available in the sampling file in order to match the sampling weights and other sampling design variables into the survey data file of the respondents.

The narrowest correct sampling file is such in which case the sampling design is simple random sampling. In this case, the file consists only of an identifier and one constant inclusion probability, and the survey outcome variable that identifies the respondents, the non-respondents and the in-eligibles. These data allows only calculate the single sampling weights for each respondent. No real non-response analysis can be done due to completely missing auxiliary data.

If a two- or three-stage design has been used, there are more variables, including PSU's as clusters, and SSU's, respectively. Even though there are no any strata or auxiliary variables, it is possible to review non-response by PSU and SSU, respectively. This gives opportunity to adjust the weights as well to some extent since non-response may vary by SSU conditional to PSU. Hopefully, all PSU's are still in the file. Otherwise, the fieldwork has been failed.

The sampling file is primarily needed to create the weights for the respondents although these are good to first create for the gross sample. Secondarily, the file is for analysing the success of the fieldwork. It is possible that a particular survey may use more than one survey mode, like in the case of a mixed-mode design. The sampling file naturally must

hence include also the mode used in data collection for all individuals. If two or more modes are used for one individual, this should be coded at variable level too.

A good sampling file is naturally very useful to analyse survey quality. Auxiliary variables particularly are needed for this purpose. Also, we would be happy if some variables of the fieldwork file would be merged with the sampling file.

#### **4. AUXILIARY VARIABLES IN THE SAMPLING FILE**

We have above given examples of auxiliary variables of a good sampling file. Now, we concretise this issue. It is good to recognize that all such variables are given for individual gross sample units whatever they are. In the case of multi-stage sampling, such variables can be more problematic since they are first concerned clusters of the target units. There can thus only be such variables that are concerned clusters. If the clusters are small-areas, regional information is available but it is more difficult to know, for example, about the education of all cluster persons. But this is not necessary since most important is to gather information about education of the respondents and the non-respondents within this cluster.

Auxiliary variables can thus be either macro or micro. Both of these variables are useful and even for the same purpose; they can be derived from the same basis. For example, age of an individual can be used in non-response analysis in several forms, like age as such or as age groups. But the same variable is useful as target population statistics and thus being a macro auxiliary variable indicating how many target population members are in each age group. This is an example of the benchmarking information, and they can be used in calibration methods that require macro auxiliary data, that is, known population margins (e.g. Deville and Särndal 1992). There can be several population margins in calibration at the same time. And if such information is available in the sampling file, it is easy to compute the calibrated weights, respectively, using the French software Calmar 2, among others (see Le Guenne & Sautory 2005).

Macro auxiliary variables can thus be margins of known population figures giving opportunity to use these in calibration. They can also be relative frequencies of small areas like PSU's, concerning for instance register unemployment rates, rates of highly educated people, or crime and poverty rates. Such variables could be used for analysing reasons of nonresponse.

The richness of the auxiliary variables in the sampling file facilitates in analysing the success of the fieldwork. For example, unit non-response can be assessed against these variables and the multivariate response propensity model estimated, consequently. This model may respectively be a good starting point for adjusting the sampling weights to take into account the variation in non-response (e.g. Laaksonen 2007, Laaksonen and Heiskanen 2013)).

The sampling file should be explicitly available, that is, for all gross sample units, and all inclusion probabilities should be in the file. Sometimes, these probabilities are only implicitly available. For simple random sampling it is most common since the inclusion probabilities are unique that only requires one population statistics figure and the gross sample size. Hence it is impossible to check based on this probability that everything has been correctly done. Thus whether SRS is really good or not?

Another annoying situation is two-stage sampling when the equal absolute sample sizes are used in the second stage. This leads to the final inclusion probabilities in which the

PSU sizes of the first stage clusters will disappear. It means that this size is not necessarily needed in the formula of the inclusion probability. Unfortunately, there exists sampling files where there is only one 'final' probability of this kind. One example is in Burnham et al (2006) that Laaksonen (2008) criticised due to missing inclusion probabilities or even so that all concrete information about first stage sampling is missing. So, it is possible that everything has not been done completely correctly. We are not saying that they bluffed but it is in this case and always possible if the sampling design information is almost lacking. The same bluffing possibility is in all designs, even in simple random sampling, since the sampling data file is so restricted that very little only can be checked. Good auxiliary data (macro and micro) also lever confidence in the survey data and hence it should be recommended to collect.

## **5. A FINNISH EXAMPLE**

In the end, an example from the Finnish security survey (FSS) 2010 is presented (Aromaa 2010, Laaksonen and Heiskanen 2014). The characteristics of its sampling data file are given below.

The number of statistical units of the FSS is 7933. They are thus gross sample units.

Table 1 illustrates the variables of the sampling file.

This list is rather long, and good in many meanings. It gave opportunity to analyse non-response by various auxiliary variables. Based on the data, we also created the so-called adjusted sampling weights. This first exploits the response propensity modelling and finally the stratification based on such calibration that the known population statistics match with our gross sample design weights by strata.

Naturally, we used the data also for survey quality including the analysis of problems in the fieldwork. This was possible for two reasons: (i) based on para data, we were able to follow the interviewing time that was shortening during the fieldwork; the response time vary by mode also so that web took least time and face-to-face most, (ii) we made a special survey for the interviewers and found that the point (i) was in telephone interviewing due to the hecticness of their job in the end of the fieldwork. Naturally, the results were not ideal.

Our sampling file thus is rich but it is not common everywhere. The file content also depends on the survey practice. Our European Social Survey team has met various interesting things that should be included in the sampling file. One is the so-called reserve sample that is initially created to guarantee that enough respondents will finally be found. It is clear that this reserve should be probability based, but if the reserve part is not included in the sampling file, it will be hard to follow the fieldwork well and even to calculate correct response rates. This reserve sample option is now in our template. It is interesting that a certain country found this option there and wanted to take a reserve sample even though this was not in their sampling file. This is not correct.

## **6. END NOTES**

I really hope that survey organisers will pay attention to create an as good sampling data file as possible and such that helps in getting improved estimates from the survey. Unfortunately, this concept is not in standard literature currently. Hopefully it will be so in a future.

**Table 1. The key variables of the sampling data file for the Finnish Security Survey.** Symbols in the column 'Source': SO = Created by survey organisation, M = Computed by methodologist, PR = Population Register, FER = Formal Education Register, ER = Employment Register, TR = Tax Register. The alternatives for use: R = Merging with respondent data, S = Sampling, U = Unit non-response, W = Weighting, E = estimation

Variable	Source	Use for
Identifier	SO	R
Survey mode (face-to-face, telephone or web)	SO	W E
32 Explicit strata by region, gender and age group, code	M	S U W E
PSU's, 100 out of 600 small areas, anonymous code	M	S U E
PSU size, Stratum size (incl. size of the target population)	M	S E
1st stage inclusion probabilities for PSU's and 2 <sup>nd</sup> stage probabilities for individuals	M	S W E
Age in years	PR	S U W
Age group respectively, both micro codes macro statistics	PR	S U W
Gender, code and macro	PR	S U W
Regional variables including municipality, postal code, co-ordinates of home, code and for some also macro	PR	S U W
Marital status with different options, year of marriage, number of marriages, code	PR	U W
Native language and citizenship	PR	U W
Occupational or socio-economic status (fairly rough only available)	PR	U
Household composition including number of children at different age groups	PR	U W
House variables such as size, number of rooms and type of kitchen	PR	U W
Level and field of education	FER	U W
Unemployed or not, number of months unemployed	ER	U W
Taxable income	TR	U W
Fieldwork outcome (respondent, non-respondent, in-eligible)	SO	M U W E
Reason for non-response (well for face-to-face, badly for web)	SO	U
Para data, e.g. interviewing time, responding time	SO	E

## Acknowledgements

The author is grateful to the sampling expert team of the European Social Survey for useful discussion over several years. The team consisted when I was writing this note, in addition to me, of Matthias Ganninger, Sabine Häder and Siegfried Gabler from Mannheim, and Peter Lynn from Essex.

## REFERENCES

- [1] Burnham, G., Lafta, R., Doocy, S. and Roberts, L. (2006) Mortality After the 2003 Invasion of Iraq: a Cross-sectional Cluster Sample Survey. *The Lancet* 368, 1421–1428.
- [2] Deville, J-C. & Särndal, C-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 376-382.
- [3] Laaksonen, S. (2007) Weighting for Two-Phase Surveyed Data. *Survey Methodology*, December Vol. 33, No. 2, pp. 121-130, Statistics Canada.
- [4] Laaksonen, S. (2008) Retrospective Two-Stage Cluster Sampling for Mortality in Iraq. *International Journal of Market Research* 50, 3, 403-417.

# Quality in the web data collection: Standardising online questionnaires, integration with administrative sources and development of bias control mechanisms

## Basque Statistics Office (EUSTAT)

CRISTINA PRADO, e-mail: [cristina\\_prado@eustat.es](mailto:cristina_prado@eustat.es)

PATXI PIZARRO, e-mail : [patxi-pizarro@eustat.es](mailto:patxi-pizarro@eustat.es)

CARMEN GUINEA, e-mail: [mcarmen\\_guinea@eustat.es](mailto:mcarmen_guinea@eustat.es)

**Keywords:** Quality, Security, Integration, Costs

### 1. INTRODUCTION

Three years ago, in the 2011 NTTS seminar, we presented the project called "Standardising the Web Channel" which was being implemented at this moment and whose primary aim was to create a software platform that allowed the electronic questionnaires used up to now to be redesigned, so that they take into account new design guidelines and computing systems and thereby improve the quality of the data gathered via the Internet and increase direct collection from the individual surveyed as opposed to other forms of collection.

This project is fully operational in the production systems of our statistical institute: all questionnaires relating to both economic and social statistic operations have been completed and the design standards and technical architecture established in the web channel standardisation project are being followed.

With the use of electronic questionnaires, respondents have to resolve problems and pitfalls that may arise on their own while completing the questionnaires, and with no personalised help available, in contrast to other forms of data collection. For this reason the quality of the data collected could have a major impact upon the statistical results, directly effecting the precision of estimations.

Our presentation aims to explain the solution adopted in the design of the electronic questionnaires and the improvements that this project has brought to the data collection processes, both in business and social statistics.

### 2. IMPROVEMENT OF THE WEB CHANNEL IN GATHERING DATA

In this section we will explain how the electronic questionnaires were designed from the following points of view:









## 2.1. Data quality

Until fairly recently, data collection, whether on paper or over the telephone, was directed by interviewers whose job was not only to obtain answers but to ensure that these answers were as truthful as possible. With information collected via electronic questionnaires, the knowledge of the professionals who prepared the questionnaires had to be transferred to the online questionnaires so that the data quality was not affected, or could even be improved. In our project, the questionnaire quality mechanisms have centred on the following aspects:

### Design of the questionnaire



In order to prepare the style guide where all the questionnaire design standards are defined, we studied the recommendations made by international experts such as Mick Couper, and established the following guidelines, among others:

-  The use of a section-by-section scheme that is identical on every page of the questionnaire, with each section having a specific function so as to facilitate the supply of information.
-  The use of a "Pagination" design, avoiding the need to scroll up, down or sideways. Every page has more than one question, but there is any need to scroll down.
-  The use of basic elements (radio buttons, check boxes, dropdown lists, tables and images...etc) adapted to the nature of the questions, selecting whatever element is best suited to the question and always aiming to avoid biased answers.
-  Using formats for non-verbal answers that make it easier to enter the data.
-  A specific design of data tables that facilitate the understanding of this type of data.
-  Help systems defined at different levels that allow the respondents to resolve problems and/or understand questions.

### Information verification systems

In order to avoid erroneous data as far as possible, we have designed a validation and error control system that aims to minimise possible errors in answers. Three different types have been defined:

- Page validations: validations of controls or between controls on the same page.
- Cohesion validations: Validations between controls on different pages.
- Length validations: warning validations on specific controls. These are used to verify answers of questionable veracity.

The page and cohesion validations are organised into "hard" and "soft" in a manner that is user-friendly for the respondent, and aim to avoid unanswered questions. Special

attention has also been given to the way messages are presented, so as to make the questionnaire more comprehensible.

### **Browser**

We have defined a browser logic that makes it possible to change pages using sequential progress buttons or using the "browser map" through non-sequential jumps. The browser includes storage for questionnaire control information such as the period, completion date, etc. In addition to the error warning system, it includes styles and functionalities that make it possible to identify the pages that have been visited or and those that have not.

Finally, browsing between questionnaire pages includes a graph control system that activates and deactivates questions or blocks of questions and even whole pages based on answers given to key questions. This system allows a single questionnaire to be filled in by different members of a family, activating the pages and/or sections that each person is required to answer according to their demographic characteristics or their specific answers to previous questions. The intention is that the respondents focus on the questions they are required to answer and that they are not "distracted" by the software tool.

## **2.2. Data security**

Regarding data security, a pre-questionnaire has been designed as a single starting point for all electronic questionnaires on all statistical operations. It includes various security measures for both data protection and access control, and records other information of interest that allows the "workload" of the respondents in completing the questionnaire to be identified and analysed.

## **2.3. Integration with administrative sources**

In those statistical operations for which administrative data is available, these data have been integrated into the electronic questionnaires so that they can be used for one of the following functions:

**Suggested response:** When the administrative data refers to data in the questionnaire itself, they are presented in the questions as an optional response that can be selected or modified. The system has been developed to make respondents to confirm the truth of the suggested datum and to prevent it from being used inappropriately.

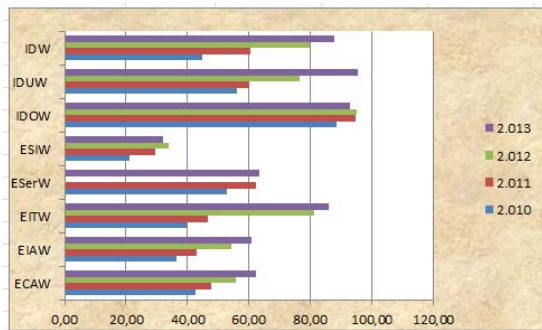
**Response verification:** When the administrative data refers to information on the respondent him- or herself, they are used to verify the answers given to certain questions as internal mechanisms for bias control.

## **2.4. Costs**

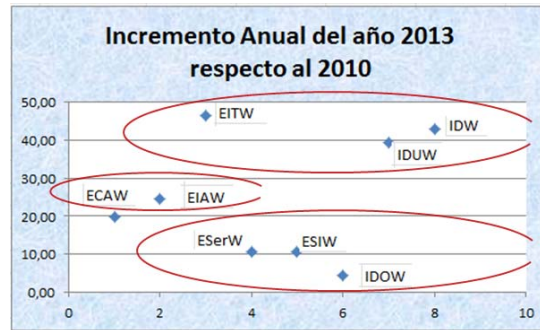
The system developed in turn allows for a reduction in costs both in the production of new electronic questionnaires and in the maintenance of existing ones, thanks to the creation of a software repository that is established as the single framework for building the web questionnaires.

### 3. RESULTS

The business questionnaires are operative with the new design defined in the web standardisation project since 2012, and therefore the percentage of results obtained through this channel has been measured, with the following results:



**Figure 1. Percentage of responses to economic questionnaires via the Web channel**



**Figure 2. Increase in the Web canal responses in 2013 with respect to 2010**

According to data, the percentage of data collection via the Web channel has increased over the years, with a very high percentage of responses exceeding 60% of the total.

The social questionnaires are operative since last year, and now we are analysing the results and comparing with collected data in previous years.

### 4. CONCLUSIONS

The Project for standardising electronic questionnaires has allowed us to take a qualitative step forward in the continuous improvement of data collection processes, adapting these processes to the use of Internet technology and making use of the facilities provided by new technologies to improve data quality, reduce response bias and non-responses and increase the precision of statistical data. In the full paper, we are going to use the Force Labour survey's Web questionnaire to explain the implemented solution in Web data collection mode.

### REFERENCES

- [1] Web Survey Methodology: Interface Design, Sampling and Statistical Inference. Mick P. Couper. Research professor at the Survey Research Centre.
- [2] Best practices for design LUKE WROBLEWSKI
- [3] ESSnet Project: Intermediary report, available at: <http://www.cros-portal.eu/content/intermediary-report-public>
- [4] Ventajas e inconvenientes de la encuesta por Internet [Advantages and Disadvantages of Internet Surveys]. Vidal Díaz de Rada. University of Navarra.

# EMOS - European Master in Official Statistics

Zivile Aleksonyte-Cormier (zivile.aleksonyte@ec.europa.eu),  
Markus Zwick (markus.zwick@ec.europa.eu)<sup>1</sup>

**Keywords:** Official statistics, statistical education, statistical literacy, lifelong learning

## 1. INTRODUCTION

The European Master in Official Statistics (EMOS) is a project aimed at developing a programme for training and education in Official Statistics within existing Master programmes at European universities. Furthermore, EMOS is a European network of universities and national statistical authorities working together to share and further develop issues of Official Statistics.

After two years of discussions and two further years of preparatory work, the first EMOS call for interest was published in August 2014. 22 universities from 13 European countries expressed their interest to participate in the EMOS network. When approved by the European Statistical System Committee (ESSC) in its meeting of 20-21 May 2015, the first Master programmes with the EMOS label will start in autumn 2015.

The article describes the main steps towards the implementation of EMOS, in particular the EMOS curriculum and the governance model, next steps and long-term prospects.

## 2. METHODS

Based on the recommendations of the EMOS feasibility study in 2013, Eurostat and the Group of Experts developed the models for the EMOS curriculum and EMOS governance [1].

The concept for the implementation of EMOS was presented to the General Directors of NSIs in their ESSC meeting on 14 May 2014. Following the favourable opinion, the concept was presented at the EMOS workshop in Helsinki to the EMOS community in June 2014 [2].

The recommendation of the feasibility study was to implement EMOS as a label for existing university Master programmes. The consensus by the Group of Experts was that the requirements should be flexible enough that all interested universities could participate under the condition that the candidate Master programme meets the established requirements.

---

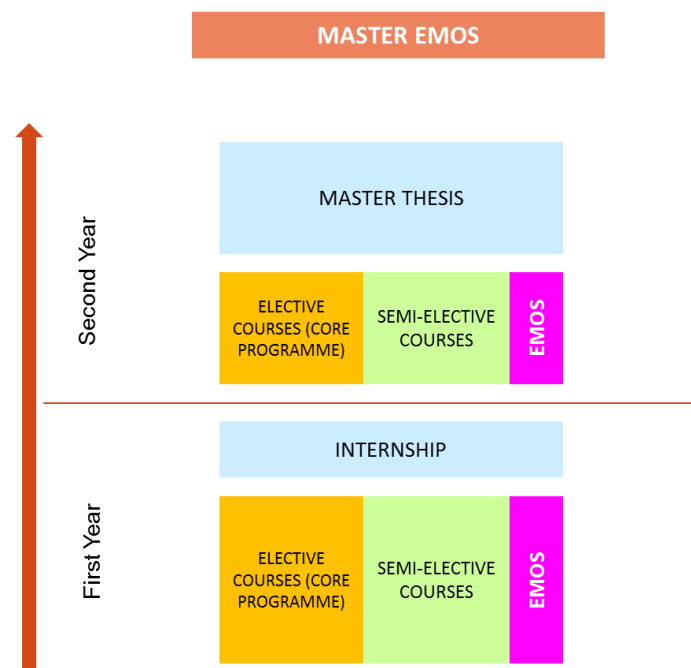
<sup>1</sup> European Commission, Eurostat

### 3. RESULTS

#### 3.1. EMOS Curriculum

Thus an EMOS-labelled Master will be made up of four main parts:

- EMOS module (approx. 10% of ECTS credits);
- Semi-elective courses (approx. 30% of ECTS credits);
- Elective courses (approx. 25% of ECTS credits);
- Internship and Master thesis (approx. 35% of ECTS credits).

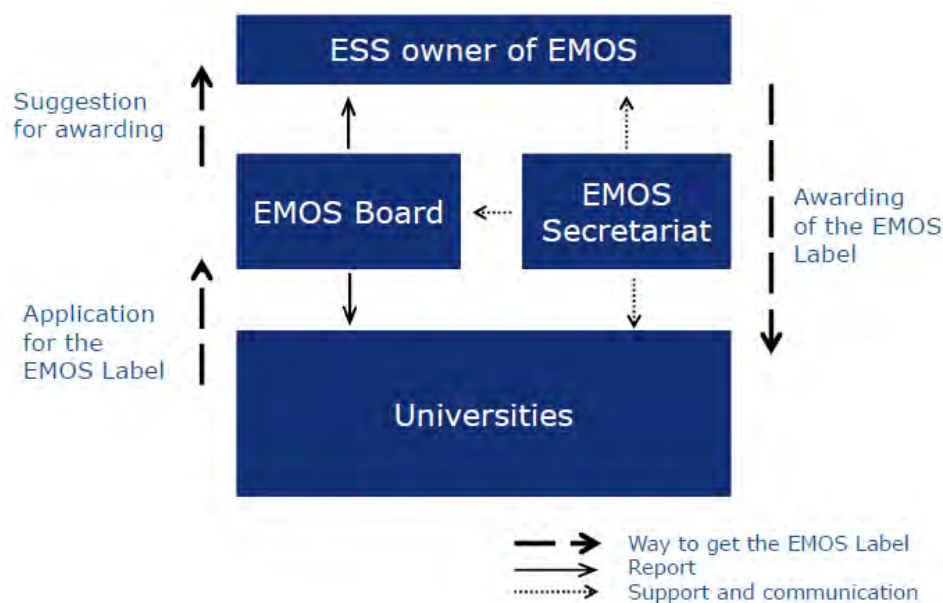


**Figure 1. EMOS curriculum**

#### 3.2. EMOS governance

The EMOS governance model includes several bodies involved in the development of EMOS, including the labelling process. At the top of the model is the ESSC who is the owner of the EMOS label. The ESSC is responsible for awarding university Master programmes with the EMOS label, also for quality and development of EMOS.

The EMOS Board is an essential part of the EMOS governance framework. The Board will assist the ESSC with regard to the development of EMOS, in particular the award of the EMOS label to Master programmes in compliance with the EMOS requirements. It will also contribute to quality monitoring of the EMOS Master programmes in order to ensure that the required standards are achieved and maintained through evaluation of new applications and extensions of the EMOS label.



**Figure 2. The EMOS governance model**

#### 4. NEXT STEPS

In January Eurostat published the EMOS call for applications addressed to the 15 pre-selected universities. The universities have been requested to submit full applications to prove the implementation of EMOS in their Master programmes from autumn 2015.

In March 2015, the EMOS Board will assess the applications and will propose for the European Statistical System Committee (ESSC) to award the selected Master programmes with the EMOS label in its meeting of 20-21 May 2015.

Following the award of the first master programmes with the EMOS label, Eurostat will publish the second EMOS call for applications. The second round will benefit from the experience of the ongoing selection, yet an evaluation of the project will be carried out in 2017 to assess the development of the EMOS network. Following the evaluation, the ESSC will have to decide on further steps of EMOS.

#### 5. CONCLUSIONS

Official Statistics is seeing continuous developments, which means that future and current staff members of statistical offices need to continuously acquire new knowledge and update their current skills. An intensive cooperation with universities is a key in meeting these challenges; therefore, one of the next steps is to see how EMOS could be used for training staff in NSIs.

The European Master in Official Statistics is getting real as a label and growing as a community; however, there is a lot of work and further development ahead.

## REFERENCES

- [1] <http://www.cros-portal.eu/content/emos>
- [2] <https://wiki.helsinki.fi/display/EMOS2014/Home+EMOS>

# Going beyond GDP: a challenge also for training statisticians. A new Master proposal and experience

Filomena Maggino<sup>1</sup> (filomena.maggino@unifi.it) Maria Pia Sorvillo<sup>2</sup> (sorvillo@istat.it)

**Keywords:** Well-being measurements, post-graduate master, cooperation university-NSI

## 1. INTRODUCTION

Italy has recently deployed some important initiatives related to the development of a set of measurement to describe the well-being of population, beyond the merely economic ones summarized by GDP.

These experiences were mainly based on a strong collaboration between Academia, the National Institute of Statistics (ISTAT) and other relevant organizations, such as the National Council for Economics and Labour. The first result attained was the definition and the elaboration of a set of indicators, covering the dimensions of well being from its different perspectives. The project is known with the acronym BES (Benessere Equo e Sostenibile, Fair and Sustainable Wellbeing; cfr. 1).

The intense methodological and conceptual work that brought to these results has also given the opportunity to collect a number of different expertise (on statistics, sociology, economics, IT, communication) activating them around a common theme and a common effort.

Furthermore, the release of first results has boosted the interest in the subject of “beyond GDP measures” in a number of organizations, especially at the local level, and shown the opportunity to set up a new academic educational proposals giving the possibility to young statisticians to be trained in the new perspective of “going beyond GDP”.

## 2. METHODS

### 2.1 The QoLexity Master

On this basis, the University of Florence (Department of Statistics, Informatics, Applications "G. Parenti" - DiSIA) developed a new Master named "QoLexity Measuring, Monitoring and Analysis of Quality of Life and its Complexity", in strict cooperation with ISTAT and with the Eurostat's sponsorship (Eurostat offered some stage position for students).

“QoLexity” is a neologism, coined on the occasion of first international workshop of the Italian Association for Quality-of-Life Studies (AIQUAV) which was held in September 2011 in Florence (cfr. 2). It refers to a complex approach to Quality of Life (defined QoLexity), covering the issues of defining, measuring, monitoring and analysing the quality of life in quantitative terms, and involving different academic disciplines (philosophy, sociology, psychology, statistics, economics, politics sciences). Hence, the program was built adopting a logic leading from concept definition to statistical indicators, a guiding concept which is "complexity", and an approach that leads from research questions to data to analysis and communication of results.

---

<sup>1</sup> University of Florence

<sup>2</sup> Istat



The master aims at developing the well-known “knowledge triangle”, putting into practice the linkage between innovation (new approaches and new skills in statistics), education (taking into account the need to provide methodologies to define, measure analyse and communicate new issue) and research (also through practical experiences and traineeships).

## **2.2 Contents and didactic organization**

A “second level master degree”<sup>3</sup> was identified as the right formula to address the educational need at the academic level. Access to the second level master is reserved to students in possession of any second level degree (master degree) who has in their curriculum studiorum at least one exam in Statistics.

The master develops four main topics, addressing different questions related to: conceptual definitions, data sources and collection, analytical tools and strategies, findings’ communication and dissemination.

The master provides 60 Credits<sup>4</sup> with almost one third given to internship and thesis (respectively 13 and 5) to stress the importance of practical activities and applications in the economy of the Master.

Teachers belong mainly to the University of Florence and the Italian National Institute of Statistics, but also to other Italian universities and international organizations with a relevant expertise in the specific field, such as the European Commission, GESIS (Leibniz Institute for the Social Sciences) and OECD.

## **3. RESULTS AND FUTURE PLANS**

The first edition of the master started in January 2014 and the presentation of the thesis is planned in January 2015. It was hold partly in Florence (at the University’s premises) and partly in Rome (at Istat), taking the students close to where official statistics are actually produced.

Five students were trained for the entire program, while other two took the opportunity to follow only some modules. They are already working in public administrations and banks, and will be able to apply immediately their new skills in a working environment, either managing their task with more awareness of new methods and techniques or developing new strands of activity in the “Beyond GDP” framework.

Students expressed an high level of participation and the evaluation of the program was quite good, both as a whole and for each module. The mix of theoretical lessons and practical applications, especially those related to the production and use of official statistics directly linked to the BES dimensions, was considered particularly effective.

The project proved to be a practical example of how academia and official statistics can effectively collaborate to create an innovative training offer that make the most of respective strong points, complementing each other’s expertise in a fruitful way.

These are the reasons why a new edition of the master is being planned, confirming the collaboration between the University of Florence and ISTAT, who are jointly evaluating the possibility to enlarge the cooperation to other subjects. In fact, during the first edition an interest

---

<sup>3</sup> According to the Italian University system, the access to the second level master is possible to anyone in possession of a qualification corresponding to bachelor’s degree (3-year degree) + master degree (2-year degree).

<sup>4</sup> In the Italian University System, each credit corresponds to 6 to 12 hours of work (lessons, individual study, laboratory, and so on).

about the initiative was shown by other universities, that proposed integrations to the curriculum and offering their professors' expertise to contribute to the training offer.

#### **4. CONCLUSIONS**

The experience of addressing training issues in the field of "Beyond GDP" measurement proved to be fruitful. The pilot edition of the master was also useful to test the collaboration between an academic establishment and a national statistical institute with respect to an high-level training initiative.

One issue should be particularly addressed in view of the coming editions, i.e. the need to improve communication towards the potential participants. That was actually a weak point in the first edition, as it was not possible to realize enough specific actions to spread information and increase awareness about the new master. New ways to disseminate information, in addition to those already used such as the institutional web sites of the two organizations, should be identified and tested in order to increase the participation in the following editions of the Master.

#### **REFERENCES**

[1] <http://www.misuredelbenessere.it/index.php?id=51>

[2] [http://www.aiquav.it/newsletter/NewsletterAiquav\\_Dicembre\\_2011.pdf](http://www.aiquav.it/newsletter/NewsletterAiquav_Dicembre_2011.pdf)

# **TITLE: A new job for statisticians: the data scientist. Which skills, how to build them**

Ludovico Antonio Ottaiano ([ottaiano@istat.it](mailto:ottaiano@istat.it))

**Keywords:** data, data scientist, competence profile, skills development

## **1. INTRODUCTION**

The recent explosion of digital data made available a large amount of data to be explored, analyzed, connected in order to build knowledge and create value.

In the last few years, there has been an explosion in the amount of data available. People spend more and more time online, leaving behind tracks of the data they use. The web is full of *data-driven apps*: e-commerce applications, for instance; and any web front end has a database behind it, collecting data from users. And we leave a data trail behind us whenever we surf the web, chat with our friends on Facebook, or buy something in a shop.

Mobile applications leave an even richer data exhaust: many of them are geolocated, and make available a great deal of data, all of which can be mined. Point-of-sale devices and frequent-shopper's cards make it possible to capture data also from retail transactions, not just from the online ones.

To take advantage of that, organisations more and more need people able not only to simply collect and report on data, but also able to look at those data from many perspectives, determine what they mean and suggest how to apply them in the business strategy .

“Data science” is the label used for this kind of competences and activities. And the “data scientist” is the new job profile related to a more powerful use of this huge amount of data available.

But what are the skills a data scientist is required to have? How can an NSI build such a competence profile?

## **2. METHODS**

Actually, using data isn't really what we mean by “data science”. As Mike Loukides explains in his *What is Data Science?* “a data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product. Data science enables the creation of data products”.

Significant worldwide companies (Google, Facebook, LinkedIn, Amazon) had their gains by using data creatively, turning it into something of value. By implementing its PageRank algorithm, for example, Google was among the first to use data outside of the page itself, in particular, the number of links pointing to a page. Tracking links made Google searches much more useful, and PageRank has been a key factor to the company's success. The PYMK (People You May Know) algorithm – that LinkedIn and Facebook use to suggest, starting from patterns of friendship relationships, people users

should know - is another example of a creative and profitable use of data collected from users. Services like I Tunes analyses music users listen to and look for, and suggest them lists of songs they may want to buy.

All these are examples of “data products” that share the same feature: they are based on the fact that data collected from users (search terms, contacts, music) provides companies with added value for their business.

Now, in such a huge sea of data, the question companies have to face is how to use data (not just their own data, but all the data available) effectively. In fact, how can they make that data useful?

Using data effectively requires something different from traditional statistics. What differentiates data science from statistics is that data science is characterized by what Loukides calls a “holistic approach”. Such an approach involves gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.

This is a job for data scientists.

The title of “data scientist” was coined in 2008 by D.J. Patil, and Jeff Hammerbacher, who then led their teams of experts of data analysis at LinkedIn and Facebook: according to them, it was the title that seemed to best express people who used both data and science “to create something new”.

A data scientist can be seen as an evolution from the business or data analyst role.

Whereas a traditional data analyst look at data from a single source, a data scientist will explore and examine data from multiple disparate sources, looking at it from many angles. The goal is different: discover previously hidden insight, to pick the right problems that have the most value to the organization.

Data scientists “make discoveries while swimming in data”, as Patil says. While doing that, they give structure to large quantities of formless data and make analysis possible. As they make discoveries, they communicate what they’ve learned and suggest its implications for new business directions.

The data scientist role has been described as “part analyst, part artist.” Anjul Bhambhri, vice president of big data products at IBM, says that “a data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization.”

In fact, often they are creative in displaying information visually and making the patterns they find clear.

Visualization is frequently the first step in analysis, and take the data scientist through each step of his/her work. To make clear what the numbers mean, to explain which stories they are really telling, data scientists have to adopt a visualization approach and use visualization techniques and tools.

Summing up what we have said so far, we can now list a first set of activities a data scientist is required to carry out:

- looking for rich data sources
- working with large volumes of data
- cleaning the data
- crossing multiple datasets
- analyzing connections among data

- visualizing the data analysis.

What kind of person does all this? What abilities make a data scientist successful? What skills does he/she need?

Patil makes a list of what makes a good data scientist:

- *Technical expertise* in some scientific discipline.
- *Curiosity*, that's the willing to go beyond the way data looks like, and find out what is hidden behind it.
- *Storytelling*: the ability to make a story based on the data available and to communicate it effectively.
- *Cleverness*: that is the ability to look at a problem in different ways and from many perspectives.

We are talking about interdisciplinary profiles, able to tackle issues from different points of view and to look at the problem they have to face as a whole, as part of a bigger one. Profiles that are required to be statisticians, mathematicians, programmers, even "artists".

Are such profiles available on the job market? Who shall we look for?

The traditional backgrounds are not helpful anymore: a data management expert is not necessary so able to analyze data as he is in organizing data; vice versa, a quantitative analyst can be valuable in analyzing data, but not so able in managing and shaping a mass of unstructured data into a form that can be analysed.

So, if to find such competence profiles available on the job market is not so easy, could it be a better solution for an organization to create and grow, rather than hire, its own data scientists? Not necessarily a profile with all a data scientist's skills required, but an integrated team in which such skills are distributed among the members.

Building a group of data scientists was the road followed by D.J. Patil at LinkedIn. That team - the team which turn out in developing products like PYMK - wasn't made only by statisticians, mathematicians

and other "data people." It was a fully integrated group that included people working in design, web development, engineering, product marketing, all able in working with data. The aim was not to re-produce those silos that traditionally separated data people from engineering, from design, from marketing, and to make the data scientists' group a full product team responsible for designing, implementing, and maintaining data products.

At LinkedIn, the strategy of building a cross-disciplinary group of experts of data science resulted successful.

But how can it work for a NSI? Can a NSI build integrated team of data scientists' competences also through training activities?

### 3. RESULTS

In Italy a reflection on data scientist's skills and training intervention started within the framework of the Italian Digital Agenda. Moving from the statement of the Scheveningen Memorandum on "Big Data and Official Statistics", adopted by the ESSC on 27 September 2013, the Italian National Institute of Statistics (Istat) gave its

contribution both to the definition of the skill profile on data scientist and to the design of an introductory course on data science.

The skill profile was defined within the framework of the Web professional profiles list published by IWA (International Webmasters Association) Italy according to the CEN guidelines in the field of Generation 3 (G3) European ICT Profiles.

Based on that, the data scientist profile was described in terms of mission, tasks, E-Competence Framework skills. Such a framework was the point from which the reflection at Istat on how to create and develop data scientist's skill started.

A project team was established, with the aim of integrating competences and experts coming from different areas: statistics, IT, organization, training.

It was decided to design a training activity made of different steps: a first step being a seminar addressed to a wide audience, at an introductory level, with the aim of verifying within Istat the interest towards such topics and collect needs of deepening the subject, to be met at a later stage; in this second step, addressed to both statisticians and IT experts, traditional lectures will be integrated by "laboratory" activities, with the aim of promoting interaction among the different competences involved.

As for the contents, some macro-areas were identified: statistical methods, IT techniques and visualizations tools. A set of topics for each of them was defined as well, to be dealt with at a different level of detail according to the target: from the fundamental of database design and management to data mining; from big data platforms and applications (Hadoop, MapReduce, Cassandra, Hive) to programming languages (Python, Perl, Php), to data modeling and machine learning.

#### **4. CONCLUSIONS**

The recent explosion of digital data made available a large amount of data to be explored, analyzed, connected in order to build knowledge and create value. To take advantage of that, organisations more and more need "data scientists", people able not only to simply collect and report on data, but also able to look at those data from many perspectives, determine what they mean and suggest how to apply them in the business strategy.

Organisations can either hire such profiles or grow their own data scientists, investing in building and developing integrated team of data scientists' competences. It's an investment worth making, since the ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — is going to be a hugely important skill in the next decades.

#### **REFERENCES**

- [1] M. Loukides, What is Data Science, O'Reilly Media, Inc.
- [2] D.J. Patil, Building Data Science Teams, O'Reilly Media, Inc.
- [3] T. H. Davenport, D.J. Patil, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review

# Geo-spatial data and statistics to support the knowledge base for monitoring natural capital

Jan-Erik Petersen ([Jan-Erik.Petersen@eea.europa.eu](mailto:Jan-Erik.Petersen@eea.europa.eu))<sup>1</sup>, Anton Steurer  
([Anton.Steurer@ec.europa.eu](mailto:Anton.Steurer@ec.europa.eu))<sup>2</sup>

**Keywords:** geo-spatial data, statistics, knowledge base, ecosystem accounts, natural capital

## 1. INTRODUCTION

### 1.1. Policy context

The context of this paper is the European Union's 7<sup>th</sup> Environmental Action Programme (7EAP) to 2020 - 'Living well, within the limits of our planet'<sup>3</sup>. The first priority of the 7EAP is to protect the Union's natural capital, which includes "biodiversity, including ecosystems that provide essential goods and services, from fertile soil and multi-functional forests to productive land and seas, from good quality fresh water and clean air to pollination and climate regulation and protection against natural disasters." The programme includes under the term also marine, coastal and fresh waters, land and forests as well as air. This objective therefore focuses on ecosystems and ecosystem services.

The 7EAP also requires as priority action 5 to enhance the environmental knowledge base for monitoring the success of environmental policies and trends in natural capital (*inter alia*). While the statistical system and environmental monitoring programmes have expanded substantially and now cover many environmental variables, many of these systems have developed in a fragmented way, with different institutions producing data sets, with gaps and overlaps in data and the data partly responding to historical needs no longer adequate to modern policymaking. There is a dual challenge of modernising the existing fragmented information system and provide key information on ecosystems where such information is currently scarce.

To help develop the knowledge base for the 7EAP a coordination group has been established at EU level – the Environment Knowledge Community (EKC), comprising DG ENV, Eurostat, JRC, DG RTD, European Environment Agency and DG CLIMA. A key component of the group's work will be a small number of 'knowledge innovation projects', one of which focuses on monitoring of, and accounting for, natural capital.

### 1.2. Analytical challenge

An EU level workshop organised by Eurostat and EEA in June 2014 on the knowledge base for the 7EAP made the following recommendations to develop sound physical data on nature and ecosystems:

---

<sup>1</sup> European Environment Agency

<sup>2</sup> Eurostat

<sup>3</sup> <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013D1386>

- To fully map available data sets between all producers, develop a joint producer & user approach and enhance cooperation;
- to promote further harmonisation and standardisation of data sets and environmental reporting (e.g. via INSPIRE) and develop an approach that makes data sets developed at national and EU level more complementary, for example in the area of land use/land cover;
- for data sampling start with a joint analytical frame and develop the sampling strategy from that; consider factors for combining different data sources in the final analysis; agree on joint data categories (e.g. on land use/cover); ensure the right spatial level for different types of natural capital.

The purpose of this paper is to propose a working method for developing an effective approach to geo-statistical accounting for ecosystems as part of an integrated and coherent Union-wide system of geo-statistical information.

## **2. METHODOLOGICAL APPROACH**

### **2.1. Identifying the key challenges**

Representative information about natural systems requires a combination of various data sets, including statistics, environmental monitoring data, LUCAS-type<sup>4</sup> ground observation surveys and the analysis of satellite images. The work on ecosystem accounting involves layers of geo-referenced data on land use and land cover along with layers of physical data about human activities, biomass production, water use and availability, carbon sequestration and release, etc. Also administrative sources (e.g. cadastre data) and sources that provide data only for part of the territory (e.g. information related to Natura 2000 areas) can be useful.

To construct these aggregated layers it is important to establish a solid basis of basic data for different components of natural capital that are all spatially referenced. As long as these basic data layers do not exist, or cannot be linked, all accounting efforts will remain "experimental" and partly theoretical. The development of these different data layers need to be informed by current ecosystem accounting methodology [1] to make sure they match analytical requirements. At the same time a sampling and data integration design should be used that is cost-effective and maximises analytical possibilities. This sampling frame could be devised on the basis of experience in other statistical domains.

A comparison with demographic statistics on population, ageing, shrinking societies, migration, households etc. serves as an example. In demographic statistics the different surveys are well organised with a view to integration in such a way that some core

---

<sup>4</sup> Land use/cover area frame statistical survey. LUCAS is a European field survey program which involves ground visits to gather harmonised data on land cover/land use and agro-environmental issues across the EU. LUCAS surveys were undertaken in 2009 and 2012. A new LUCAS survey is under way with field data collection between March and September 2015. The next survey is planned for 2018. For the statistical sample of the LUCAS survey, a regular 2 km grid with over 1,100,000 points is overlaid on the EU territory. In 2012 the final sample consisted in approximately 270,000 points which were visited in-situ by 700 field surveyors, which are trained to follow a harmonised methodology. At each visited point, the surveyors fill in a field form and document the visit by taking a number of photographs of the point, resulting in more than 2,000,000 photos per campaign. Moreover a transect (a 250 meters walk to the East of each point) is recorded; the surveyors walking the transect are requested to register all the land cover transitions they can observe according to a list of codes including both areal land cover classes and linear elements. A topsoil sample of approximately 500g is collected in 10% of the visited points.



statistical variables (population, sex, age, household composition etc.) are collected with a census type of methodology, which then serves also as a reference frame for diverse samples of well-defined sample sizes, which are dedicated to more specific and more complex issues (composition of household, employment, income etc.). With this integrated system, the highest possible effectiveness (relevance, representativity, reliability, actuality etc.) can be achieved with a very efficient production method (respecting the available budget, in advance planning of the survey modules, transparency, democratic control etc.).

## **2.2. Developing a logical sequence of analytical steps**

The following steps are proposed for identifying an approach for establishing a knowledge base for the monitoring of, and accounting for, natural capital:

- 1) Identify what are the essential environmental and other parameters to monitor for analysing natural capital trends.
- 2) Develop a comprehensive and efficient sampling frame for these variables.
- 3) Review what are the currently available statistical, geo-spatial, earth observation and other data sources relevant for monitoring the parameters identified under 1) and to what degree they match the sampling frame under 2).
- 4) Where feasible identify possible combinations of different data sets for developing input parameters for natural capital accounting to minimise investment needs.
- 5) Establish what is the right mix of data between statistical, geo-spatial, earth observation and other data sources.
- 6) Design an integrated natural capital sampling frame and geo-statistical system taking account of knowledge gained under steps 2 – 5.
- 7) Based on steps 3) and 6) describe the key gaps to be filled and the mechanisms to be used for that (adaptation of current surveys or establishing new instruments).
- 8) Agree with key partners how to move from the current situation to a more integrated system
- 9) Implement and build up an integrated and efficient data foundation. In parallel, conduct experimental accounting work using available data to provide first insights and to demonstrate the uses of the integrated data set.

## **3. FIRST CONCLUSIONS**

- a) Monitoring of, and accounting for, natural capital is now a priority objective of EU environment policy.
- b) The current data foundation derived from statistical, environmental, earth observation and other sources is not sufficient and needs to be further developed

by integrating and making better use of existing data and by new and improved data collection to fill gaps.

- c) As a characteristic of ecosystems is their complex and often fine-grained geographic distribution any new or revised data gathering exercise needs to be built on geo-referencing of the data to be collected that is commensurate with the scale of the observation units.
- d) Geographic information systems and a common geo-spatial reference frame will therefore play an important role in developing monitoring and accounting systems for natural capital.
- e) Experience from other areas of statistics, such as demography, can be fruitfully harvested for developing a cost-effective and analytically powerful sampling frame for building a good knowledge base on natural capital.
- f) The EU level process has come into place to tackle the challenges in developing a reliable and comprehensive system for the monitoring of, and accounting for, natural capital. However, the real work in identifying priority tasks and the right common sampling frame is only starting.
- g) It is essential to start establishing sufficient analytical capacity to be able to make full use of the information once the data foundation is in place.

## REFERENCES

- [1] United Nations, European Commission, Food and Agriculture Organization of the United Nations, Organisation for Economic Co-operation and Development and World Bank Group, (2014). System of Environmental-Economic Accounting 2012 - Experimental Ecosystem Accounting. Last viewed on 25/01/2015 at:  
[http://unstats.un.org/unsd/envaccounting/seeaRev/eea\\_final\\_en.pdf](http://unstats.un.org/unsd/envaccounting/seeaRev/eea_final_en.pdf)

# The European bird monitoring programmes as examples of citizen science relevant to policy and research.

Petr Voříšek<sup>1</sup> (EuroMonitoring@birdlife.cz), Ruud Foppen<sup>2</sup> (Ruud.Foppen@sovon.nl), Richard Gregory<sup>3</sup> (richard.gregory@rspb.org.uk).

**Keywords:** citizen science, indicator, biodiversity

## 1. INTRODUCTION

It has been widely recognised that biodiversity is important for human quality of life and ecosystem services. Information on biodiversity trends is therefore needed to guide policy decisions for achieving a good management of species and habitats. However, obtaining good quality information on biodiversity poses a challenge; the information is patchy and incomplete. Birds are exceptional in this context because they are very popular among the public and thousands of birdwatchers observe and report on birds as their hobby. Moreover, birds are considered to be good indicators of biodiversity. Indeed, bird indicators based on count information have been used globally, continentally and nationally. Information on (changes in) bird numbers and their distribution are used in relation to many (inter)national policy issues.

## 2. BIRD CENSUS – PAN-EUROPEAN COMMON BIRD MONITORING

In Europe bird monitoring techniques have developed towards highly standardized field methods, though simple enough to be performed by volunteer ornithologists across large geographical units. Bird monitoring schemes are organized nationally or regionally, mostly by non-governmental organizations, often in cooperation with universities and research institutes. Volunteer fieldworkers count birds in the breeding season (although winter counts also exist) every year providing long-term data spanning several decades in some countries. Sample plots were traditionally selected by volunteers themselves, but monitoring schemes established more recently (from 1990s) have used randomised selection of sample plots. Nowadays, European breeding bird monitoring schemes coordinated through the Pan-European Common Bird Monitoring Scheme (PECBMS)<sup>A</sup> provide relevant data on bird numbers from 27 European countries and more than 160 bird species. The number of fieldworkers taking part in the field surveys is estimated at well over 10 000.

Yearly population indices and trends are the most important outputs of national monitoring schemes. The index gives bird numbers to a base year, when the index value is set at 100. Usually, but not necessary, the first year of a time series is chosen as the base year. Trend values express the linear overall population change over a period of years.

National species indices are produced by the coordinators of the monitoring schemes. They compute the individual national species indices in a prescribed way using TRIM software (Trends and Indices for Monitoring data, Pannekoek & Van Strien, 2001). TRIM is a widely used freeware program (available via <http://www.ebcc.info/trim.html>). The national indices are collected by the PECBMS coordinator.

<sup>1</sup>Pan-European Common Bird Monitoring Scheme & European Breeding Bird Atlas, Czech Society for Ornithology, Czech Republic

<sup>2</sup>European Bird Census Council, Dutch Centre for Field Ornithology SOVON, The Netherlands

<sup>3</sup>Pan-European Common Bird Monitoring Scheme, Royal Society for Protection of Birds, UK

After extensive data quality checks, the supranational (regional and then European) indices of species' population change are produced. A method developed for this purpose (Van Strien et al, 2001) takes into account the differences in population sizes per country, as well as the differences in field methods and in the numbers of sites and years covered by the national schemes.

Finally, supranational species indices are combined in multispecies indicators. These are produced for groups of species according to their main habitat types (forest, agriculture etc). The rationale behind the construction of composite indicators is that each species is seen as a replicate that may respond in the same way to environmental drivers as the other species and repeats the same signal. To produce multispecies indicators, we average indices (by taking geometric means), rather than abundances in order to give each species an equal weight in the resulting indicators. The composite geometric mean represents the average behaviour of the constituent species.

PECBMS produces European species population indices and trends as well as European and regional indicators for farmland, forest and all common bird species. The outputs are published on annual basis at the website and provided also to EU institutions (e.g. Eurostat and DG Environment). One of the more prominent indicators is the Farmland Bird Indicator produced by PECBMS which has been used by the EU as one of its indicators of Sustainable Development and also as a Structural Indicator. Furthermore, the Farmland Bird Indicator has been adopted as an indicator for EU Rural Development Plans. The common bird indicator produced according to PECBMS' methodology has been accepted as an official indicator of biodiversity by governments in at least 26 European countries.

PECBMS outputs were also successfully used in detecting an effect of climate change and developing an indicator of climatic changes in Europe. Further improvements and potential new policy relevant indicators are under development.

Apart from the widely documented decline of European farmland birds as a result of agricultural intensification, PECBMS data were used in documenting decline of long-distance migrants or the effect of climate change on bird population trends. Since 2002, the PECBMS outputs have been used in 24 scientific peer-reviewed papers.

### **3. SPATIAL DISTRIBUTION SURVEYS – EUROPEAN BREEDING BIRD ATLAS**

However, information on relative population change is not enough for evidence-based biodiversity conservation. Information on the geographical distribution of birds and its changes in time is also needed. In Europe, an atlas of bird breeding distribution was published in 1990s. This first atlas (Hagemeijer & Blair 1997), with its main data collection period in the 1980s, was a milestone in European ornithology, and its data were used beyond the publication of the book for setting conservation strategies at European and national levels, to study the impacts of climate change and for scientific studies on a wide range of topics. However, in the thirty years since data collection, the environment but also the political context in Europe have changed and the bird distribution data are now out of date. Climate and land use changes have altered the habitats of birds across the continent and projects focusing on individual species or regions indicate effects on bird populations at a large scale. However, a continent-wide overview is lacking. Therefore, besides its main role in the PECBMS project, the European Bird Census Council (EBCC) started work on the second European Breeding Bird Atlas (EBBA2)<sup>B</sup>.

The EBBA2 project aims to provide up-to-date distribution maps for birds across the whole of Europe and to document changes in species distribution since the 1980s. The project will also build capacity for conservation and monitoring in areas where this is most needed.

The methodology for field data collation was developed to fit the needs of high quality data standardization and different capacities of (mostly) volunteer fieldworkers across the whole of Europe, from Azores to Russia and including Turkey (Herrando et al 2013). National coordinators are responsible for organising a fieldwork methodology and campaign, which can be either aimed at also producing data for national breeding atlases, or aimed at delivering the data for the EBBA2 project. The fieldwork for the EBBA2 project is scheduled from 2013 to 2017, the results are expected to be available in 2020. The project with 5 years of fieldwork, some 50 countries, 500 bird species, a coverage of 5000 50x50 km squares and an estimated 50 000 fieldworkers is probably the most ambitious biodiversity monitoring programme in Europe.

#### **4. PORTALS FOR DATA GATHERING AND SHARING**

Both projects take advantage of rapidly developing techniques for data analyses and particularly for data collection and sharing. On-line portals for storage, management and sharing data on bird observations have appeared in Europe and worldwide. These portals, if properly managed and organised, are a useful source of data for EBCC projects and they serve public engagement. We aim to integrate and use these technical developments more in future PECBMS and EBBA2 work. Therefore, under the umbrella of the EBCC a new project has recently started called European Bird Portals (EBP)<sup>C</sup>.

#### **5. CONCLUSIONS**

PECBMS, EBBA2 and EBP are citizen science projects based on cooperation between volunteer fieldworkers and professional researchers. Based on our experiences clear synergies in several areas have been identified: on the level of creating and maintaining networks of coordinators and fieldworkers, the use of field data use for multiple purposes, capacity building for conservation and research and policy use of the outputs.

In the near future, we expect a further integration of methodologies, data collection, data management and data analyses concerning bird monitoring (PECBMS), atlas (EBBA2) and birdwatching portals (EBP). This enable us to accomplish a faster data delivery and presentation, to improve the quality of the data, to achieve for instance a better geographical coverage of Europe to better understand bird population changes and to achieve a greater public engagement in biodiversity monitoring and conservation. This integrated approach promises also to deliver more policy relevant information and to trigger further scientific investigations. Although the citizen science approach proved to be cost effective, financing a central coordination, analytical and training capacity remains a challenge.

<sup>A</sup> The Pan-European Common Bird Monitoring Scheme (PECBMS) has commenced in 2002 as a joined initiative of the European Bird Census Council (EBCC) and BirdLife International. Initially supported by the Royal Society for Protection of Birds (RSPB) and Statistics Netherlands, hosted by Czech Society for Ornithology (CSO) in recent years the programme has received substantial support from the European Commission.

<sup>B</sup> The European Breeding Bird Atlas (EBBA2) is a project led by EBCC. Presently the work is coordinated by staff of the Swiss Ornithological Institute, the Czech Society for Ornithology and the Catalan Ornithological Institute.

<sup>C</sup> The European Bird Portal initiative under the auspices of the EBCC is coordinated by staff of the Catalan Ornithological Institute and the Swiss Ornithological Institute and is supported by more than 15 European organisations collecting bird data by online portals.

# Small Area Estimation models with outliers in covariates

Monica Pratesi ([monica.pratesi@unipi.it](mailto:monica.pratesi@unipi.it))<sup>1</sup>, Caterina Giusti<sup>1</sup>, Stefano Marchetti<sup>1</sup>, Nicola Salvati<sup>1</sup>

**Keywords:** M-quantile, Robust estimation, Influence function.

## 1. INTRODUCTION

The objective of small area estimation (SAE) methods is to use survey data for estimating some characteristics, such as means, totals, quantiles of items of interest, in areas or domains where the sample size is not large enough to obtain reliable direct estimates. In the last years SAE methods have received a growing interest, since their use in the formulation of new policies and programs, poverty mapping and measurement of well-being indicators at detailed geographical level highly rely on these methods ([1]).

The most popular approach to model-based small area estimation are linear mixed models, that include random area effects to account for between area variations ([2]). Chambers and Tzavidis [3] proposed a new approach to SAE associated to M-quantile regression methods. M-quantile estimation is free of any distributional assumption, and is robust with respect to the presence of outliers and influential observations in the response variable.

The presence of outliers is a common feature in real data. Chambers [4] classifies outliers into two groups. Representative outliers are correctly measured sample values that are outlying relative to the rest of the sample data and for which there is no reason to believe that similar values do not exist in the non-sampled part of the survey population. Non-representative outliers are gross errors in the sample data, which have nothing to do with the values in the non-sampled part of the survey population. Either type of outlier can have a substantial impact on the estimates if ignored.

In the context of model-based estimation outliers can affect both the response and the auxiliary variables. The special case of outliers in the response variable resulting in the presence of outlying observations has been treated in the literature under both approaches to small area estimation, the mixed model and the M-quantile ones ([5], [6], [7]).

Sinha and Rao [5] suggest a way to extend their robust small area estimator to account for outliers in the auxiliary variables. In this paper we develop their estimator, and we propose a new M-quantile based small area estimator that account for outliers in the covariates. The two estimators are compared by means of model-based simulations both in the situation with representative and non-representative outliers.

## 2. METHODS

In a typical small area estimation problem we consider a population  $U$  of size  $N$  divided into  $d$  non-overlapping subsets  $U_i$  (domains of study or areas) of size  $N_i$ ,  $i = 1, \dots, d$ . With unit level models we commonly assume that a vector  $x_{ij}$  of  $p$  auxiliary variables is known without error for each unit  $j$  belonging to area  $i$ , while the values of the variable of interest  $y_{ij}$  are available only for a sample of population units in each area,  $s_i \subset U_i$  of size  $n_i \geq 0$ . The set  $r_i \subset U_i$  contains the  $N_i - n_i$  indices of the non-sample units in area  $i$ .

---

<sup>1</sup> University of Pisa, Via C. Ridolfi 10, 56124 Pisa (PI), Italy.

We suppose that  $\zeta\%$  of the  $x_{ij}$  are representative-outliers or that  $\zeta\%$  of the sampled  $x_{ij}^*$  are non-representative-outliers, so that the covariates or the sampled covariates are  $x_{ij}^* = \{(1 - \zeta\%)x_{ij}, \zeta\%(x_{ij} + \eta_{ij})\}$ , where  $\eta_{ij} = C$ .

We propose to deal with outliers in the covariates using a robust method that down-weights on the basis of extreme leverage values ([8]). We choose the trisquared redescending function to down-weight the values of the auxiliary variables - excluding the intercept. This function is defined as follows:

$$w(t) = \{1 - (t/k)^2\}^3 I(|t| \leq k), \quad (1)$$

where  $k$  is a tuning constant. For each observation we define the value  $z_{ij} = (x_{ij}^* - \mu)^T V (x_{ij}^* - \mu)$  where  $\mu$  is a  $p-1$  vector of ‘robust’ estimates of the centers of the  $p-1$  auxiliary variables and  $V$  is a ‘robust’ estimate of the  $(p-1) \times (p-1)$  covariance matrix of the auxiliary variables (without the intercept term). Defining  $u_{ij} = (z_{ij} / (p-1))^{1/2}$ , the weight to be used to down-weight the extreme  $p-1$  values is

$$w(t) = w(x_{ij}^*) = \{1 - (u_{ij}/k)^2\}^3 I(|u_{ij}| \leq k). \quad (2)$$

When the linear M-quantile model ([3], [9]) holds the proposed M-quantile small area estimator for the mean is as follows

$$m_i^{mq-rob} = \left( n_i + \sum_{j \in r_i} w(x_{ij}^*) \right)^{-1} \left[ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} w(x_{ij}^*) x_{ij}^* \hat{\beta}_{\psi,w}(\hat{\theta}_i) + \frac{\sum_{j \in r_i} w(x_{ij}^*)}{n_i} \sum_{j \in s_i} w(x_{ij}^*) r_{ij} \right] \quad (3)$$

where  $r_{ij} = y_{ij} - x_{ij}^{*T} \hat{\beta}_{\psi,w}(\hat{\theta}_i)$  and  $\hat{\beta}_{\psi,w}$  is obtained solving the following estimating equation with the iterative weight least square (IWLS) algorithm

$$\sum_{i=1}^d \sum_{j=1}^{n_i} \psi_q(y_{ij} - x_{ij}^{*T} \beta_{\psi,w}(q)) x_{ij}^{*T} = 0. \quad (4)$$

Equation (4) takes into account both the influence function for outliers on the target variables (in the simulations we use the Huber proposal 2) and outliers on auxiliary variables. Indeed, for fixed  $q$  the estimator of the regression parameter,  $\hat{\beta}_{\psi,w}$ , is  $\hat{\beta}_{\psi,w}(q) = \{X^{*T} W(q) X^*\}^{-1} X^{*T} W(q) y$ , where  $W(q)$  is a diagonal matrix of order  $n$  which contains the final set of weights produced by the IWLS algorithm used to compute  $\hat{\beta}_{\psi,w}$ . The other quantities in equations (3) and (4) are as in Tzavidis et al [9].

Note that the proposed estimator allows for the presence of representative and non-representative outliers in the covariates as well as outliers in the target variable. As a remark, the proposed estimator can account for outliers only for continuous variables. The same apply to our extension to the robust EBLUP estimator presented by Sinha and Rao [5].

The estimator proposed in equation (3) is based on the bias corrected version of the M-quantile estimator, see reference [9]. Using the estimated parameter  $\hat{\beta}_{\psi,w}$  is

straightforward to extend the so-called naïve version of the M-quantile estimator (for the small area mean) to the case of outliers in the covariates. Under some settings the naïve M-quantile estimator can be more efficient than the bias corrected M-quantile estimator (as it is shown in reference [7]).

Using the same down-weight function we can obtain the robust version against outliers in the covariates of the robust (against outliers in the target) EBLUP  $m_i^{reblup}$  originally proposed in [5]. This new version of the REBLUP is as follows

$$m_i^{reblup-rob} = \left( \sum_{j \in r_i} w(x_{ij}^*) \right)^{-1} \left[ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} x_{ij}^T \hat{\beta}_w + \hat{u}_{i,w} \right], \quad (5)$$

where the parameters  $\beta_w$  and  $u_{i,w}$  are obtained solving the following estimating equation

$$\sum_{i=1}^d X_i^T W_i V_i^{-1} U_i^{1/2} \psi(U_i^{1/2} y_i - X_i \beta) = 0. \quad (6)$$

Matrix  $W_i$  is a  $n_i$  diagonal matrix with elements  $w(x_{ij}^*)$ ,  $X_i$ ,  $V_i$  and  $U_i$  are as in [5].

### 3. RESULTS

We use model-based Monte-Carlo simulations under several alternative scenarios to empirically evaluate the performance of the proposed small area robust estimators (3) and (5), comparing them with their original versions (not robust for outliers in the covariates),  $m_i^{mq}$  and  $m_i^{reblup}$ , and with the classical EBLUP,  $m_i^{eblup}$ .

**Table 1. Results under Scenarios A and B, model-based simulations.**

	$\zeta = 1\%$		$\zeta = 3\%$		$\zeta = 5\%$	
	RB%	RRMSE%	RB%	RRMSE%	RB%	RRMSE%
Setting A: outliers in the y, non-representative outliers in the x						
$m_i^{eblup}$	-0,8	14,62	-1,40	14,95	-1,57	15,08
$m_i^{mq}$	-1,01	18,62	-1,65	18,99	-1,7	19,22
$m_i^{reblup}$	-0,38	11,05	-1,24	12,07	-1,51	12,40
$m_i^{mq-rob}$	-0,28	18,05	-0,48	18,26	-0,29	18,46
$m_i^{reblup-rob}$	-0,12	10,94	-0,32	11,33	-0,47	11,83
Setting B: outliers in the y, representative outliers in the x						
$m_i^{eblup}$	0,34	14,71	0,14	14,86	0,22	15,11
$m_i^{mq}$	0,30	18,64	0,21	19,02	0,27	19,31
$m_i^{reblup}$	0,75	11,42	0,65	12,18	0,35	12,54
$m_i^{mq-rob}$	0,00	18,06	-0,20	18,11	0,04	18,55
$m_i^{reblup-rob}$	0,11	11,05	0,05	11,47	-0,28	12,28

Under simulation scenarios A and B we consider the presence of outliers in the target variable and the presence of non-representative or representative outliers in the auxiliary variable, respectively. Population data for  $d=30$  areas with  $N_i=100$  are generated by using a unit level area random effects model with normally distributed random area effects and



unit level errors,  $y_{ij} = 1 + 2x_{ij} + v_i + e_{ij}$ , where the area random effects  $v_i \sim N(0,3)$ , but 10% come from the distribution  $v_i \sim N(0,30)$ , the unit level errors  $e_{ij} \sim N(0,6)$ , but 10% come from the distribution  $e_{ij} \sim N(0,150)$ . For the  $x$  we set  $x_{ij} \sim N(5,1)$ , with  $\zeta\%$  of the sampled  $x_{ij}$  affected by non-representative-outliers (under scenario A) or with  $\zeta\%$  of all the  $x_{ij}$  affected by representative-outliers (scenario B), as follows:  $x_{ij}^* = \{(1-\zeta\%)x_{ij}, \zeta\%(x_{ij}+10)\}$ . The sample is obtained by selecting a within small areas random sample of  $n_i = \{5, 10\}$  units from the corresponding population. To choose the value for the constant  $k$  of the weight function  $w$  we propose a cross validation criterion similar to the one proposed by Rudemo [10]. Results of the model-based simulations under scenarios A and B with  $n_i = 5$  are shown in Table 1 in terms of percentage Relative Bias (RB%) and percentage Relative Root Mean Squared Error (RRMSE%).

#### 4. CONCLUSIONS

The results of the model-based simulations are encouraging: estimators  $m_i^{mq-rob}$  and  $m_i^{reblup-rob}$  are able to down-weight the effect of the presence of non-representative and representative outliers in the auxiliary variable, both in bias and variability. In further extensions we will develop a naïve version of the proposed M-quantile robust estimators as well as a bootstrap estimators of the mean squared error for the presented estimators.

#### REFERENCES

- [1] Pfeiffermann, D. 2013. New important developments in small area estimation. *Statistical Science* 28, 40–68
- [2] Rao, J. 2003. *Small Area Estimation*. New York: Wiley.
- [3] Chambers R. and Tzavidis N. (2006). M-quantile models for small area estimation. *Biometrika* 93, 255–68.
- [4] Chambers, R. (1986) Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063- 1069.
- [5] Sinha, S. and Rao, J. 2009. Robust small area estimation. *The Canadian Journal of Statistics* 37, 381–399.
- [6] Chambers, R., Chandra, H., Salvati, N., Tzavidis, N., 2014. Outlier robust small area estimation. *Journal of the Royal Statistical Society - Series B*, 47–69.
- [7] Giusti, C., Tzavidis, N., Pratesi, M., Salvati, N., 2014. Resistance to outliers of m-quantile and robust random effects small area models. *Communications in Statistics - Simulation and Computation* 43, 549–568.
- [8] Carroll, R., Pederson, S., 1993. On robustness in the logistic regression model. *Journal of the Royal Statistical Society Series B* , 693–706.
- [9] Tzavidis, N., Marchetti, S., Chambers, R., 2010. Robust estimation of small area means and quantiles. *Australian and New Zeland Journal of Statistics* 52, 167–186.
- [10] Rudemo, M., 1982. Empirical choice of istograms and kernel density estimators. *Scandinavian Journal of Statistics* 9, 66–78.

## **6. REFERENCES**

- [1] Hagemeijer, E.J.M. and M.J. Blair (editors). 1997. The EBCC Atlas of European Breeding Birds: their distribution and abundance. T & A.D. Poyser, London.
- [2] Herrando, S., Voříšek, P., Keller, V. The methodology of the new European breeding bird atlas: finding standards across diverse situations. *Bird Census News* 26/1-2 (2013): 6-14.
- [3] Pannekoek, J. & van Strien, A.J. TRIM 3 Manual. TRends and Indices for Monitoring Data. Research paper No. 0102. (2001). Statistics Netherlands, Voorburg, The Netherlands.
- [4] Van Strien, A.J., Pannekoek, J. & Gibbons, D.W. Indexing European bird population trends using results of national monitoring schemes: a trial of a new method. *Bird Study* 48 (2001): 200-213.

# Accounting for Hyperparameter Uncertainty in a Small Area Application Based on a State-Space Model: the Case of the Dutch Labour Force Survey

Oksana Bollineni-Balabay (o.balabay@maastrichtuniversity.nl)<sup>1</sup>, Jan van den Brakel<sup>1</sup>, and Franz Palm<sup>2</sup>

**Keywords:** bootstrap, hyperparameter, state-space model, true MSE, unemployment

**Abstract:** The sample sizes for the Dutch Labour Force survey (DLFS) are too small to produce reliable monthly figures using design-based estimators. To reduce the variance of the design-estimates, a structural time series model is adopted, which is known as a powerful technique for reducing variance estimates in repeatedly conducted surveys. In small and medium samples, however, the model-based variance estimates of the small area quantity of interest will be underestimated if the uncertainty from replacing the model hyperparameters with their maximum likelihood (ML) estimates is ignored. Several approximation approaches known in the literature are tested for their ability to account for this uncertainty. The results suggest that the relative bias of the signal MSE produced by the Kalman filter can be reduced from about -2 to 2 percent. Even with this slight positive bias, the standard error of the design estimates is reduced by about 22 percent.

## 1. INTRODUCTION

Monthly figures about the labour force are important economic indicators. Most national statistical institutes (NSIs) use a rotating panel design in their Labour Force Surveys (LFS), but in most cases the sample size is not large enough to produce sufficiently precise monthly figures based on design-based estimators. As an alternative, model-based estimation procedures could be considered to increase the effective sample size. At Statistics Netherlands, a multivariate structural time series (STS) model is used, originally proposed by [1]. Variance estimates based on STS models are usually substantially lower compared to the design-based variance estimates. However, applications based on such models often ignore the fact that the true hyperparameters have been replaced by their estimates in the Kalman filter recursions. The obtained mean square error (MSE) estimates are therefore negatively biased. Other techniques frequently used in SAE, like the empirical best linear unbiased predictor and the hierarchical Bayesian approach, usually do take the parameter uncertainty into account, so STS models should be no exception in this respect.

This paper focuses on the true MSE estimation for the DLFS model. To account for the hyperparameter uncertainty, we compare the asymptotic approximation developed by [2] (referred in this paper to as AA), as well as the parametric and non-parametric bootstrapping approaches developed by [3] (PT1 and PT2, respectively) and [4] (RR1 for parametric and RR2 for non-parametric). An extended Monte-Carlo simulation study, where the DLFS model acts as the data generation process, intends to establish the best approach to the MSE approximation for this survey. Finally, the simulation shows how

---

<sup>1</sup> Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4481, 6401CZ Heerlen, the Netherlands / Maastricht University School of Business and Economics, P.O. Box 616, 6200 MD Maastricht, the Netherlands.

<sup>2</sup> Maastricht University School of Business and Economics, P.O. Box 616, 6200 MD Maastricht, the Netherlands.

the validity of the model can be checked. Therefore, such a Monte-Carlo study can be viewed as a tool to assess model uncertainty.

## 2. THE DLFS MODEL

The DLFS is five-wave rotating panel survey with estimates produced on a monthly, quarterly and annual basis (see [5] for details). The series considered in this study are monthly general regression (GREG) estimates  $\mathbf{Y}_t$  of the total number of unemployed labour force from Jan 2001 till June 2010. A 5-dimensional vector  $\mathbf{Y}_t$  with five waves of the target design can be decomposed as  $\mathbf{Y}_t = \mathbf{1}_5 \boldsymbol{\xi}_t + \boldsymbol{\lambda}_t + \mathbf{e}_t$ , where  $\boldsymbol{\lambda}_t$  is a vector containing the rotation group biases (RGB), that are systematic differences between the waves, and  $\mathbf{e}_t$  is a vector with the survey errors modelled as an AR(1) process with an autoregressive parameter  $\rho$ . It is assumed that the true population parameter is  $\boldsymbol{\xi}_t = \mathbf{L}_t + \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t$ , where  $\mathbf{L}_t$  is a stochastic trend,  $\boldsymbol{\gamma}_t$  is a stochastic trigonometric seasonal component, and  $\boldsymbol{\varepsilon}_t$  is an irregular component that is omitted here due to its insignificance. The DLFS model is developed by [5], where the full description of the model can be found.

The design standard errors are used in the model as an input. Parameter  $\rho$  is going to be estimated as in [6] from the input data, whereafter the disturbance variances are estimated by the quasi-maximum likelihood method, treating  $\hat{\rho}$  as given. With a separate survey error hyperparameter per wave, the hyperparameter vector is nine-dimensional:  $\boldsymbol{\theta} = (\rho, \sigma_R^2, \sigma_Y^2, \sigma_\lambda^2, \sigma_{e_1}^2, \sigma_{e_2}^2, \sigma_{e_3}^2, \sigma_{e_4}^2, \sigma_{e_5}^2)$ , where the  $\sigma^2$ -terms correspond to the stochastic term variances of the slope, seasonal, and RGB components, as well as of the five survey error components, respectively.

## 3. TRUE MSE APPROXIMATION

If the estimated hyperparameter vector  $\hat{\boldsymbol{\theta}}$  is used in the Kalman filter recursions, the true conditional mean square error is defined as  $\mathbf{MSE}_{t|t} = E_t(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}_t)^2$ , where the expectation is taken with respect to the distribution of the quantity of interest (a state variable or the signal)  $\boldsymbol{\alpha}_t$ . This true MSE can be decomposed as the sum of the filter uncertainty and parameter uncertainty:

$$\mathbf{MSE}_{t|t} = E_t(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t)^2 + E_t(\hat{\boldsymbol{\alpha}}_{t|t}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}))^2$$

The first term, filter uncertainty, is approximated by the naive MSE-estimates delivered by the Kalman filter. The second term, the parameter uncertainty, can be approximated using the methods mentioned in Section 1.

## 4. RESULTS

The central data generating process of the present simulation study is the DLFS model. The performance of the five MSE-approximation methods is going to be examined on samples of the original length  $T = 114$ , as well as for  $T = 80$  and  $T = 200$ . For each of these sample lengths, a Monte-Carlo experiment is going to be set up with  $S = 1000$  multiple series. The focus of this simulation study is the true MSEs of the population signal consisting of the trend and seasonal components.

The simulation has shown that the variances of the seasonal and, in particular, RGB components are often estimated to be close to zero. This causes bi-modality in the distribution of these variance estimates, as well as the normality distortion in the

distribution of the other hyperparameters. Therefore, four DLFS model formulations are considered in this simulation study: Model 1 (the original model); Model 2 (with  $\sigma_Y^2 = 0$ ); Model 3 (with  $\sigma_\lambda^2 = 0$ ); and Model 4 (with  $\sigma_Y^2 = \sigma_\lambda^2 = 0$ ). The simulation evidence suggests that the preference in modelling the DLFS series may be given to the more parsimonious Model 3.

The performance of each of the five approximation methods  $b$  is evaluated with the help of the MSE relative bias:  $RB_t = \left( \frac{MSE_{t|t}^b}{MSE_{t|t}^{true}} - 1 \right) \cdot 100\%$ . Table 1 presents the relative biases averaged over 1000 simulations and over time, discarding the first 30 observations. This is the time needed for the diffuse part of the state covariance matrix to decay. MSE simulation averages for Model 3,  $T = 114$  are presented in Fig. 1.

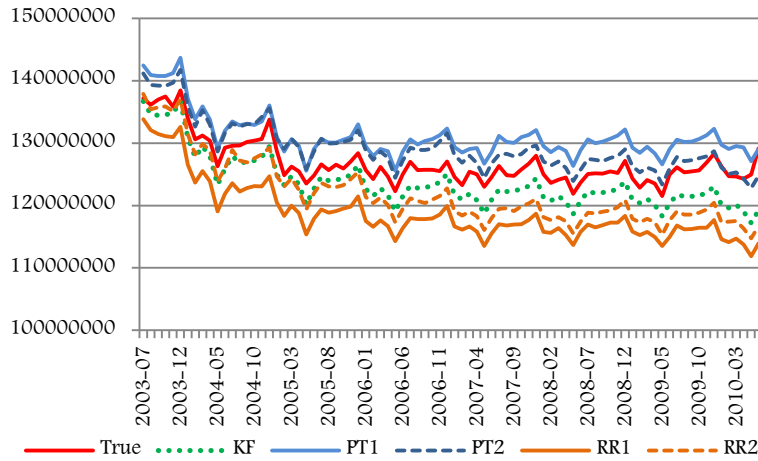
The AA-method it turned out to be inapplicable to the models with marginally significant hyperparameters. When some of the hyperparameters are estimated close to zero, conditional on a  $\rho$ -draw, the information matrix becomes numerically either singular, leading to a failure in the procedure, or nearly singular. In the latter case, the asymptotic variance becomes excessively large and thus not reliable. Taking this into account, the AA-method could only be considered for Model 4. As expected, the method performs poorly for the short sample length with positive biases of about 15 percent. The performance for  $T = 114$  and  $T = 200$  is comparable to that of the PT1-bootstrap, but significantly worse than the PT2 performance.

The simulation results for  $T=80$  suggest that the time average of the Kalman filter (KF) relative bias of the signal MSE ranges between -3.2 and -2.1 percent for the four models considered in the paper. This bias tends to decrease as the sample length increases. The KF-biases are quite small for the case of  $T=200$ , such that none of the approximation methods offers a smaller bias in absolute terms. One could still apply the best approximation method with positive biases in order to get a range of values containing the true MSE.

**Table 1. Signal MSE relative bias in the DLFS model, percent, d=30**

	<b>T=114</b>				<b>T=80</b>				<b>T=200</b>			
<b>Models</b>	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>
<b>KF</b>	-2.1	-2.6	-2.4	-2.2	-3.0	-3.2	-2.1	-2.2	-1.3	-1.6	-1.3	-1.3
<b>AA</b>	NA	NA	NA	5.2	NA	NA	NA	14.9	NA	NA	NA	5.9
<b>PT1</b>	8.1	5.7	3.3	5.5	8.6	6.7	4.9	6.2	<b>6.3</b>	6.2	6.3	5.5
<b>PT2</b>	<b>2.2</b>	<b>3.2</b>	<b>1.9</b>	<b>1.5</b>	<b>4.8</b>	<b>3.7</b>	<b>1.4</b>	<b>2.1</b>	6.8	<b>4.0</b>	<b>3.0</b>	<b>2.3</b>
<b>RR1</b>	-8.3	-7.8	-6.4	-6.5	-7.2	-9.0	-7.3	-7.2	-8.0	-8.0	-4.9	-5.9
<b>RR2</b>	-1.1	-6.0	-3.9	-3.5	6.7	-3.5	-3.9	-3.7	-5.1	-5.6	-4.5	-5.0

What one immediately sees is negative biases for the RR-bootstrap and positive ones for the PT-method. Against the claim of Rodriguez and Ruiz in [4] that their approach has better finite sample properties compared to the approach of Pfeiffermann and Tiller [3], the case of the DLFS suggests that the RR-estimates, both parametric and non-parametric ones, are even more negatively biased than the uncorrected KF-estimates across all the models and sample lengths (except for RR2 in M1,  $T = 80$  and  $T = 114$ ). The PT-methods never produce negative biases. While the PT-bootstrap is proven to have satisfactory asymptotic properties, Rodriguez and Ruiz illustrate the superiority of their method in small samples based on a simple model (a random walk plus noise). The present simulation study reveals that the RR-method may not behave well in some more complex applications.



**Figure 1. Signal MSE comparison for Model 3, T=114 months**

## 5. CONCLUSIONS

The present work aimed at establishing the best approximation approach to the true MSE in a state-space time series model used for the production of official estimates on the total number of unemployed in the Netherlands. A simulation study conducted for this purpose reveals that the asymptotic approximation is not applicable to cases with hyperparameters close to zero due to failures when inverting the information matrix of the hyperparameter estimates. The simulation results suggest that the Pfeiffermann-Tiller bootstrap approaches with their positive biases consistently outperform the corresponding approaches of Rodriguez-Ruiz, where the biases are as a rule negative and larger than those of the Kalman filter in absolute terms. Further, the non-parametric bootstraps, being free of normality assumptions about the error distribution, perform better than their parametric counterparts in both methods. Another result of this simulation study within the scope of the model uncertainty check revealed that it might be worth considering a more robust version of the DLFS model which restricts the variance of the RGB component to zero. For this model, the relative bias of the signal MSE produced by the Kalman filter can be reduced from about -2.4 to 1.9 percent with the non-parametric Pfeiffermann-Tiller bootstrap approach. Even with this slightly positive bias, the standard errors of the design estimates are reduced by about 22 percent.

## REFERENCES

- [1] Pfeiffermann, D., Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9 (1991), 163-175.
- [2] Hamilton, J., A standard error for the estimated state vector of a state-space model. *Econometrics*, 33 (1986), 387-397.
- [3] Pfeiffermann, D. and Tiller, R., Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26 (2005), 893-916.
- [4] Rodriguez, A. and Ruiz, E., Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis*, 56 (2012), 62-74.
- [5] Van den Brakel, J. and Krieg, S., Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology* 16 (2009), 177-190.

- [6] Pfeffermann, D., Feder, M., and Signorelli, D., Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics*, 16 (1998), 339-348.

# Reliable poverty estimates in groups with small samples

Agne Bikauskaite, Dario Buono – Eurostat

**Keywords:** Small area estimation (SAE), parametric bootstrap, Empirical best predictor, Nested Error Model, poverty indicators

## 1. INTRODUCTION

The European Statistical System's (ESS) objectives is to produce and disseminate the highest quality of statistics. Data has to be precise and comparable. A very important issue is development and implementation of a framework enabling the production of small area estimates for ESS social surveys (for instance at risk of poverty or social exclusion, unemployment rate, etc.).

One of the main aims of the Europe 2020 strategy is the reduction of poverty in European Union Member States (MS). The EU target is to lift at least 20 million people out of the risk of poverty and social exclusion by 2020 compared to the level in 2008. A qualitative estimation of poverty technique in EU MS is needed to better implement, monitor and determine areas where support is most required.

The issue of small area estimation is the production of reliable estimates for groups covered by a small sample. The purpose of this paper is to present the findings of the analysis on techniques efficient for producing high reliability poverty estimates based on small samples.

## 2. DATA AND METHODS

Our estimation of poverty indicators is conducted on the micro-data from a statistical Survey in Income and Living Conditions (EU-SILC 2012) collected in several EU MS. The sample of individuals is divided in large and small groups by most frequent activity status. In addition individuals from each group are divided into several groups by age and by gender.

Estimates of poverty indicators are obtained using the Empirical best prediction method based on a Nested Error Model proposed by Molina and Rao (2010) for nonlinear small area estimation, as available within the "R" routine library SAE.



To estimate the variances of the empirical best estimators the parametric bootstrap method is used.

The results of the direct estimation available using the EU-SILC calibrated sampling weights (currently applicable by chosen EU MS) are then compared with ours, using the variance measure.

### **3. CONCLUSIONS**

The practical purpose of this paper is to find out how the Empirical best estimator proposed by Molina and Rao (2010) treats with various data set having different size of groups and to compare obtained results by applying aforementioned method and using methodology applied by Eurostat.

We expect that Molina and Rao proposed model's estimators of the poverty indicators having small samples will compute better results comparing to results gotten by using direct estimation methods chosen by MS as it is the case for Spanish data (see paper [1]).

Our aim is to provide suggestions and/or advices for the improvement (in terms of variance reduction) of the reliability of the nonlinear parameters estimation methodology obtaining small size samples.

### **4. REFERENCES**

[1] I. Molina and J. N. K. Rao, Small area estimation of poverty indicators, The Canadian Journal of Statistics Vol. 38 No. 3 (2010), 369-385

# **The Seasonal Adjustment Center of Competence**

## **Missions and First Achievements**

Dominique Ladiray (dominique.ladiray@insee.fr)<sup>1</sup>

**Keywords:** Seasonal adjustment, Jdemetra+

### **1. THE SEASONAL ADJUSTMENT CENTER OF COMPETENCE**

Created under the auspices of Eurostat, a new Seasonal Adjustment Center of Competence (SACC) was launched on April 9<sup>th</sup>, 2014.

This Center, devoted to Seasonal Adjustment and created for 2 years, has several missions:

- The testing and documentation of JDemetra+, the new software developed by the National Bank of Belgium and the National Bank of Germany;
- The promotion of this software throughout the European Statistical System and the assistance to users for the migration from Demetra+, or other seasonal adjustment software, to JDemetra+;
- The dissemination of knowledge on the development of JDemetra+ to assure not only the development of new modules or plug-ins, but also the maintenance of the software in the coming years;
- The assistance to National Statistical Institutes (NSI), Central Banks (CB) or other institutes in the domain of seasonal adjustment, with the objective to foster the application of the European Statistical System (ESS) guidelines on seasonal adjustment.
- The dissemination of knowledge on seasonal adjustment by all means, in particular documentation, methodological papers and trainings.

To achieve its mandate, the SACC works in close cooperation with the Seasonal Adjustment Steering Group (SASG) which it reports to, Eurostat and the ECB, the developers of JDemetra+ (National Bank of Belgium and National Bank of Germany) and the Seasonal Adjustment User Group (SAUG) which groups several NSIs and CBs.

The SACC is constituted of 4 NSIs: Insee (France, coordinator; leader: Dominique Ladiray), Istat (Italy, leader: Anna Ciammola), ONS (United Kingdom; leader: Duncan Elliott) and Statec (Luxembourg; leader: Véronique Elter). Two international experts, Sylwia Grudkowska (Poland) and Agustin Maravall (Spain), complete the team. The global investment represents a minimum of 1000 person-days.

Any institute in the ESS can use the services proposed by this Seasonal Adjustment Center of Competence, through the helpdesk or by direct contact.

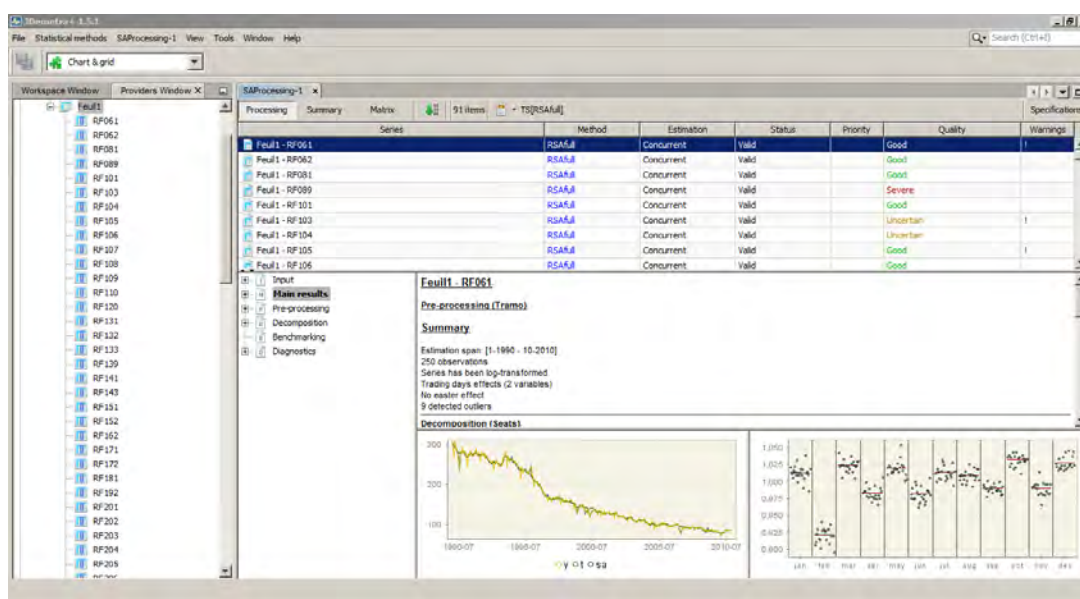
---

<sup>1</sup> INSEE, Methodology Department, 18 boulevard Adolphe Pinard, 75014 Paris FRANCE.

## 2. JDEMETRA+, THE EUROPEAN SOFTWARE FOR SEASONAL ADJUSTMENT

JDemetra+ is a new open source tool for seasonal adjustment (SA) that enables the implementation of the revised ESS Guidelines on SA. It has been developed to provide a set of reusable and extensible components, following a standard technology, compatible with the environment of most European statistical institutions. JDemetra+ is not only a user-friendly graphical interface, comparable to its predecessor, Demetra+, but also a set of open Java libraries that can be used to deal with time series related issues like the SA processing of large-scale data sets, non-standard SA methods, the development of advanced research modules, temporal disaggregation, benchmarking and business cycle analysis.

JDemetra+ is built around the concepts and the algorithms used in the two leading SA methods, i.e. TRAMO/SEATS and X12-Arima/X13-Arima-Seats. They have been reengineered, following an object-oriented approach that allows for easier handling, extensions or modifications.



## 3. FIRST ACHIEVEMENTS

### 3.1. The Helpdesk on Seasonal Adjustment

The ESS Helpdesk on Seasonal Adjustment has been technically opened on the CROS portal by Eurostat on June, 20<sup>th</sup> 2014 and is located at the following URL:

<http://www.cros-portal.eu/content/ess-seasonal-adjustment-helpdesk>

This helpdesk replaces two other helpdesks related to Seasonal adjustment already available from the CROS portal.

Any institute in the ESS can use the services proposed by this helpdesk and ask for any question related to seasonal adjustment methodology or seasonal adjustment softwares (mainly Demetra+ and JDemetra+).

Since its opening, and up to the end of 2014, the helpdesk received 26 requests from 12 different institutes.

### 3.2. In-depth Testing of JDemetra+

The SACC made an in-depth testing of JDemetra+ before its official release, and the tests were organized in 4 main steps.

- First step: Crash test and large scale evaluation of the software  
This test consisted in adjusting thousands of time series (in fact around 100,000) with JDemetra+ using the auto-modelling feature (the most demanding for the software) and the cruncher facility.
- Second Step: Test of the core engines
- Third step: Reproduction in JDemetra+ of the production processes  
Each SACC member, and some members of the SAUG, reproduced a complete production process using JDemetra+ and compared the results with the ones obtained with its current SA software.
- Fourth step: Complete functional testing  
This functional testing mainly concerns the interactive mode.

### 3.3. Highlights of the large scale test

#### 3.3.1. Methodology

Various softwares were compared, using about 100,000 real time series coming from the Euro-Indicators database:

- Tramo-Seats version 197, Tramo-Seats version 891;
- Demetra+ 1.0.4 (Tramo-Seats and X12);
- X12-ARIMA version 0.3 and X13-ARIMA-SEATS Release Version 1.1 Build 13 (Seats and X11 specifications)
- JDemetra+ 2.0.0 (Tramo-Seats and X12).

#### 3.3.2. “Crash” test

During the first tests, we experienced a few problems with Tramo-Seats and X-13-ARIMA-SEATS: for some specific series, the softwares just stopped. This problem was due to some minor bugs. In particular the DOS version of Tramo-Seats had not been as well tested as the TSW version which did not crash. These small bugs were very quickly corrected by Agustin Maravall (Bank of Spain) and Brian Monsell (USCB).

Apart these minor problems, none of the softwares crashed.

#### 3.3.3. “Success” of the modelling process

One important output of the automatic modeling of the series is of course the ARIMA model. But, in some occasions the software cannot find a model for the series. The following table reports, by periodicity, the number of series without a final model.

It clearly appears, from this point of view, that the algorithms used by the USBC are not as reliable as the ones used by Tramo-Seats or JDemetra+

**Table 1. Number of cases when a model was not found**

Period	NbSeries	DemTS	TS891	JdemTS	JdemX12	X13seats	X13	X12	TS197	X12
4	59519	374	407	376	437	695	695	967	1726	103
12	30944	64	32	65	64	372	372	524	35	524
All	90463	438	439	441	501	1067	1067	1491	1761	627

### 3.3.4. Speed

Period	NbSeries	Nobs	JdemTS	DemTS	TS197	TS891	JdemX12	DemX12	X13	X13seats	X12
4	59519	66	2.3	3.3	6.8	8.5	4.7	4.6	13.8	17.8	16.4
12	30944	176	4.0	7.8	11.1	12.0	20.2	31.6	50.1	56.7	84.8
Total	90463	104	2.9	4.8	8.3	9.7	10.0	13.8	26.2	31.1	39.8

The “TS family” appears to be much faster than the “X12 family”. This is mainly due to the modelling routine: the exact algorithms used by X13 are definitely slower than those used in Tramo-Seats.

And you have on average the following ranking:

JdemTS < DemTS < TS197 < TS891 < JdemX12 < DemX12 < X13 < X13seats < X12

### 3.3.5. Automatic modeling and discrepancies between SA series

Five global aspects have been considered to check the automatic Reg-Arima modeling process:

- The percentage of similar decomposition scheme (additive or multiplicative)
- The percentage of similar ARIMA model
- The percentage of similar number of TD regressors
- The percentage of similar number of outliers.
- The discrepancies between BIC measured by the MAPE (on similar models).

The discrepancies between SA series are measured by the MAPE.

	JdemTS				JdemX12		
	DemTS	TS197	TS891	X13seats	DemX12	X12	X13
% of similar decomposition scheme	89.99	95.18	99.09	83.59	94.55	95.36	95.37
% of similar Arima models	39.47	40.04	82.56	30.07	82.07	82.10	82.29
% of same number of outliers	49.31	61.30	88.65	56.19	87.99	88.00	87.93
% of same # of TD regressors	88.37	73.05	98.05	82.00	86.34	86.25	86.33
Discrepancies in BIC (MAPE)	1.00	0.67	0.42	1.08	0.03	0.03	0.04
Discrepancies between SA (MAPE)	0.86	1.03	0.30	0.90	0.10	0.28	0.28

From this very synthetic table, it clearly appears that:

- JdemX12 is very close to the X11-family: models are often similar and even if they are not, the BIC statistics remains very close from a model to another.
- JdemTS is very close to the last version of Tramo-Seats (TS891) with more than 80% of similar ARIMA models and average discrepancy in BIC very small.

In any case, JD+ reproduces very well the last versions of the genuine Tramo-Seats and X-12 programs

# The use of statistical services in the European System of interoperable statistical Business Registers

Susanne Maus ([Susanne.MAUS@ec.europa.eu](mailto:Susanne.MAUS@ec.europa.eu))<sup>1</sup>

**Keywords:** Statistical Service, interoperability, business register, ESBRs

## 1. INTRODUCTION

In 2009 Eurostat started a reflection on the production methods of European statistics. The reflection leads to a program aiming at re-engineering the statistical production process in the European Statistical System (ESS) with increased efficiency, improved coherence and comparability of data. Tools and administrative mechanisms will be developed to reach these objectives. Within an ESS-wide programme, sharing of information, statistical services and costs in selected statistical domains are piloted.

In this context, a five year project "*European System of interoperable statistical Business Registers*" (ESBRs) was launched in 2013 focussing on the interoperability of statistical business registers in the ESS.

The purpose of the project is to obtain better business statistics, through the interoperability of business registers resulting in improved data availability on globalisation issues, reduced inconsistencies in European business statistics and a more efficient production of business registers. The development of Statistical Services is considered as fundamental for data quality management, efficient production and a coordinated use of interoperable Business Registers.

The paper focusses on the path foreseen for the technical side of interoperability in the ESBRs and outlines the plans for developing and using Statistical Services in the production of business registers.

## 2. METHODS

Statistical business registers play a key role in creating abilities for linking and sharing information needed for producing high quality European statistics. This requires not only common concepts and harmonized operational rules for maintaining a register, but also processes, organisations and systems which are able to interact with each other. Moreover, the development and the operations of business register systems are costly. Noting the fact that statistical business registers have similar or equal functions, the opportunities will be investigated for defining generic solutions or services for these functions which could reduce the development, maintenance and exploitation costs.

According to the guidelines for describing statistical services definitions [2] a service is a representation of a real world business activity with a specified outcome. It is self-contained and can be reused by a number of business processes (either within or across statistical organizations).

A statistical service allows performing a task in the statistical production process. Using statistical services will allow statistical organizations to create flexible business processes and systems for statistical production more easily.

---

<sup>1</sup> Eurostat

In practice, a shared statistical service is an identified generic functionality representing a business function that can be used by different NSIs inside the NSIs business process to produce statistics on a given topic. Identification of generic functionality can for example take place on the level of an NSI, as a result of cooperation between NSIs or at Eurostat level. A statistical service can be made available to NSIs by exposing it in a Service Oriented Architecture (SOA) being accessible via standardised interfaces. Alternatively an NSI might replicate the service and integrate it into their business process.

On the one hand, statistical services allow sharing best practices across European countries and contribute to the overall efficiency of the European Statistical System. On the other hand, they contribute to a better consistency of statistics by standardising the methods to produce statistics across European countries.

### **3. RESULTS**

As general results of a functioning ESBRS, it is expected that all National Statistical Institutes, as producers of business registers, are able to access shared services for register management and that a catalogue of certified services exists that covers all or even most of all relevant functionalities of the business process.

The realization of sharing services is planned stepwise:

- In a first phase, which is planned for 2014 to 2015, the current situation is investigated with regard to the feasibility of common statistical services and to identify the already existing distribution of using statistical services in the ESS. Furthermore this phase serves for the identification of all functionalities that are needed in the various steps for the production and maintenance of statistical business registers and that should be developed as a statistical service in order to share them among the producers of business registers.
- The second phase is planned for 2016 to 2017. Main challenge within this phase will be the insertion of the results of phase 1 in the general ESS catalogue for statistical services so that the functionalities are available to use. Such catalogue will be a list of services described using a standardized description.
  - Currently, only a few such shared services exist or are being developed. However, a broader use of shared statistical services would increase data quality and harmonisation, and thus interoperability, and will as well reduce production costs.
  - The goal of this task is to develop a list of relevant statistical business register services, including appropriate criteria and descriptions. These statistical services do already exist or might be made available. In the other cases the catalogue will be a list of desirable statistical services.
- The subsequent phase is planned from 2018 on and will be characterized by using, improving and complement the list of statistical services in the ESBRS.

### **4. CONCLUSIONS**

Changed requests for European business statistics require modifications of the existing production of European business registers. The premise is to replace the separated national business registers with a European System of interoperable statistical Business

Registers that follow common concepts, harmonized methods and where the components of the system are able to interact with each other.

The paper points out the way forward for the business registers domain and drafts the steps that are currently undertaken to prepare European business registers to share statistical services in future within the ESBRS.

## **REFERENCES**

- [1] European Interoperability Framework (EIF) for European Public Services
- [2] ESS.VIP Programme Cross-cutting project on sharing statistical SERVICES “Guidelines for describing statistical Services definitions” Eurostat, August 2013, based on the recommendations described in the Common Statistical Production Architecture (CSPA), V1.0, UN-ECE, December 2013



# Standardisation in the European Statistical System: inventory of normative documents and the standard- setting process – results of the ESSnet on Standardisation

Mr. Csaba Ábry ([csaba.abry@ksh.hu](mailto:csaba.abry@ksh.hu)), Mr. Zoltán Vereczkei ([zoltan.vereczkei@ksh.hu](mailto:zoltan.vereczkei@ksh.hu))

**Keywords:** standards, standardisation, inventory

## 1. INTRODUCTION

The European Statistical System (ESS) Vision 2020 identifies „efficient and robust statistical processes” as one of the five key areas to deliver the vision of the ESS. In this context, the Vision 2020 document highlights that *„we will improve our efficiency through systematic collaboration within the ESS, while fully respecting the subsidiarity principle. We will intensify our collaboration by further intensifying the sharing of knowledge, experiences and methodologies but also by sharing tools, data, services and resources where appropriate. The collaboration will be based on agreed standards and common elements of technological and statistical infrastructure”*.

The European Statistical System already uses a lot of tools in different forms enhancing collaboration across the ESS, such as statistical legislation, methodological handbooks, gentlemen’s agreements, code lists. The implementation of the system envisaged by Vision 2020 should be based on a system of common tools, normative documents enabling common production, common use of integrated data and sharing common tools and infrastructure.

The term ‘standard’ is meant to be a guarantee for the ESS members that the concerned normative documents have been set according to the five principles of standardisation: consensus, transparency and openness, balance, due process, proportionality.

Standardisation is a key area in the European Statistical System for the conceptual and methodological harmonisation and further integration of data, methods and processes in the ESS. This goal asks for a procedure providing these guarantees. This process is the standard-setting process.

### 1.1. Past

The issue of standardisation at the ESS goes back to the Workshop on Standardisation held in Brussels, on 14<sup>th</sup>-15<sup>th</sup> October 2010. The workshop concluded that an ESSnet project (later called: ESSnet STAND-PREP (ESSnet on Preparation for Standardisation)) was to be launched in the same year in order to prepare actions to be carried out later by the Sponsorship on Standardisation.

The Sponsorship on Standardisation, started in 2011 and finished in 2013, worked out the framework of the ESS standardisation system. One of the main deliverables of the Sponsorship on Standardisation was the ‘recommendations document’, adopted by the European Statistical System Committee (ESSC) in September 2013. The recommendations contain the suggestions, guidelines on the operation of the ESS standardisation system, including the frame of reference for the inventory of normative documents and the standard-setting process.

In the framework of the Sponsorship on Standardisation it was also decided that an ESSnet project had to be launched, which task would be to examine the usability of the deliverables of the Sponsorship on Standardisation in practice. This ESSnet project is the ESSnet on Standardisation with Hungary (coordinator), France, Italy, the United Kingdom, the Netherlands, Latvia and Lithuania forming the ESSnet partnership.

## **1.2. Present**

The ESSnet on Standardisation started its activities in December 2012 and is working on the practical implementation of the future ESS standardisation system, based on the ESSC-adopted recommendations of the Sponsorship on Standardisation. The ESSnet is in its second phase (SGA-2), with the current action closing in March 2015.

The work of the ESSnet is built around four key areas: inventory of normative documents, the standardisation process, impact assessment and Business Architecture-related activities. The current document summarises the results of the ESSnet on Standardisation on two areas, highlighted as new items in the field of European statistics: the inventory of normative documents and the standard-setting process of the ESS.

## **1.3. Future**

The inventory of normative documents and the standards-setting process of the ESS are cornerstones of the future standardisation system in the ESS. The ESSnet on Standardisation is working on some but not all recommendations of the Sponsorship on Standardisation. Additional elements of the future ESS standardisation system are still to be developed or fine-tuned before the final system can be operational. Final decision on the establishment of the standardisation system in the ESS is still to be made in the near future.

# **2. METHODS**

## **2.1. Inventory of normative documents**

The inventory of normative documents is a publicly available storage location of normative documents relevant for the ESS, providing detailed up-to-date information on them. The aim of the inventory is twofold: it provides information for stakeholders and users in the ESS and beyond and it is also meant to support decision-making on standard development.

After receiving the necessary input from the Sponsorship on Standardisation, the ESSnet on Standardisation started to populate the inventory with descriptions for selected normative documents. The starting list of normative documents was derived from the RAMON database and other ESS-related websites. The Sponsorship also provided a template for the assessment of normative documents as ESS standards and a template of the inventory. These inputs were used by the ESSnet to provide the normative document descriptions.

It was agreed that, given the number of normative documents identified by the Sponsorship on Standardisation (593) and the limited amount of human resources devoted to the concerned Work Package of the ESSnet, to rationalise the list.

The ESSnet concentrated on the latest editions of Eurostat handbooks (or handbooks co-published by ESTAT), on international normative documents relevant for the ESS (latest

editions), on IT tools recommended at ESS level and on fundamental legal acts. In addition, only Eurostat handbooks, international handbooks co-published by Eurostat and IT tools were assessed (a dedicated template was designed for the assessment of IT tools).

## **2.2. Standard-setting process**

A core element in the recommendations of the Sponsorship on Standardisation is the process of standardisation. The goal of the standard-setting process is the development and implementation of the right standards to support quality and efficiency of the ESS through smooth interoperability and common services. The process guarantees the involvement of the relevant stakeholders, clarity on the rules, clarity on the status of the (potential) standards and the adherence to the five principles of standardisation: consensus, transparency and openness, balance, due process, proportionality.

Based on the work of the Sponsorship on Standardisation, three types of normative documents exist: regulations, standards and other normative documents. The regulation-setting process is itself enshrined in legislation: regulations provide binding legislative rules that are adopted by an authority. Standards in the ESS are meant to be distinguished normative documents that are established by consensus and approved by a recognised body that provide for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context.

The framework of the standard-setting process was worked out by the Sponsorship on Standardisation. The ESSnet on Standardisation examined the usability of this model in practice and has significantly fine-tuned it to make the model applicable in practice. The fine-tuning activity was undertaken not just from theoretical point of view (e.g. simplification and further rationalisation) but also based on lessons learned from test cases by involving initiatives at international level in the testing of the fine-tuned process model.

## **3. RESULTS**

### **3.1. Inventory of normative documents**

As a deliverable of the ESSnet on Standardisation, 103 normative documents were assessed and attributes were provided for 197 normative documents. The result of the assessment was that 27 normative documents in the inventory already meet the criteria for being ESS standards. These results are already publicly available on the CROS portal.

### **3.2. Standard-setting process**

The result of elaboration and fine-tuning the standard-setting process is a proportional and transparent standard-setting process for the ESS, based on theoretical rationalisation and practical tests.

## **4. CONCLUSIONS**

Standardisation has been an important issue in the ESS for years. In the form of the deliverables of ESSnet on Standardisation, the work of many initiatives materialise, including the ESSnet STAND-PREP, the Sponsorship on Standardisation and the ESSnet on Standardisation.

The ESSnet on Standardisation, working on four key areas of standardisation in the ESS, provides the starting set of descriptions for the inventory of normative documents and a streamlined model for the standard-setting process of the ESS. These developments, among other key issues, such as impact assessment and Business Architecture-related activities, are cornerstones of the standardisation system in the ESS.

Some elements of the system of standardisation in the ESS are still to be developed or further fine-tuned but the inventory of normative documents and the standard-setting process for the ESS are ready for use and are available for future actions.

## **REFERENCES**

- [1] ESSnet on Standardisation deliverables, 2012-2014, <http://www.cros-portal.eu/content/standardisation>
- [2] Sponsorship on Standardisation: recommendations of the Sponsorship of Standardisation, September 2013, <http://www.cros-portal.eu/content/recommendations-sponsorship-standardisation-september-2013>
- [3] Sponsorship on Standardisation: stocktaking pillar final report, 21 May 2013 [http://www.cros-portal.eu/sites/default/files//Pillar1\\_closing\\_21May.pdf](http://www.cros-portal.eu/sites/default/files//Pillar1_closing_21May.pdf)

# Manual for statistics on energy consumption in households (MESH)

Duncan Millard<sup>1</sup>, Cristian Fetic [Cristian.Fetic@ec.europa.eu](mailto:Cristian.Fetic@ec.europa.eu)<sup>2</sup>

**Keywords:** Household energy use, surveys, administrative data, data-linking, European cooperation.

## INTRODUCTION

The household sector represented around 27% of the entire consumption in the EU in 2010. Various factors explain the upward trend in energy consumption, such as an increase in the number of households, greater comfort demanded, and an increase in electrical appliances in homes. The significant consequences of energy use, in terms of energy dependence, security of supply and environmental impacts, lead to a need for more detailed knowledge on the elements of energy demand, and the factors that have a bearing on it. Hence there is a requirement for further development of energy statistics to help monitor and understand these issues. To solve these issues Eurostat supported the development of the Manual for statistics on energy consumption in households (MESH) as a part of the Eurostat ESSnet Program. The main objectives of the project are:

- Identifying the current status of the energy statistics in residential sector at European level and the MS users' needs.
- Drafting a global inventory of best practices on methods and statistics in the residential sector, both at European level and abroad.
- Producing a manual stating to provide greater insight into the use of energy in the residential sector, the various statistical techniques to be applied, the variety of practices and methodologies used.
- Carrying out a training session for all the users in the EU's MS to enable them to meet the detailed information requirements for the residential sector.
- Dissemination of the manual and the good practices found in all the EU NSIs.

---

<sup>1</sup> Chief Statistician, Department of Energy and Climate Change (DECC), UK

<sup>2</sup> Administrator, Eurostat, European Commission

The paper will explore the key features of the manual and how it explores new approaches to data productions such as the use of administrative data, including that owned by private business and the benefits of data-linking. It will also cover key learning from running a successful multi-country project covering strong communication and engagement, and how the results of a manual can be effectively brought to life through effective dissemination and training.

### Contents of the manual

The aim of the manual was to help all Member States improve their data on the household sector generally, learn alternative techniques and specifically to meet the new requirements of the Energy Statistics Regulation.

This manual is a reference document, a guidebook, and hopefully a source of inspiration for new ideas that can help statisticians provide comprehensive and comparable data on household energy use. Through this it is hoped that it will support the development of policies and monitoring of them at individual member state and EU level.

The manual is set out in 7 key chapters plus this Background. After this background, Chapter 1 gives an overview of energy statistics in the European Union, the current approaches used. Chapter 2 sets out the boundaries of the household sector and definitions of the key elements that need to be measured. The aim of this chapter is to establish definitions valid throughout the European Union that will enable comparison of energy statistics for the households sector among the countries of the European Union.

The theory of the different systems of acquisition and production of data such as surveys, administrative data, in situ measurements, modelling, etc., is presented in Chapter 3. This theoretical information is complemented in Chapter 4 with the presentation of good practices of different member states for the acquisition and production of information.

Chapter 5 presents the integrated approaches used in several Member States which show how the various methods are combined to produce statistical information in particular countries. Chapter 6 looks beyond the present and future reporting requirements to show different possibilities to disaggregate the information that go beyond the minimum levels set out in the regulation.

The manual concludes with Chapter 7 which shows methods on the production of renewable energy statistics in households, how household energy statistics can help to

improve the knowledge of fuel poverty, and the benefits of data matching whereby different datasets can be linked together, offers a new means of exploiting the maximum potential of data and at the same time provides a route to understanding wider issues around energy use in households, for example the energy savings from retrofitting energy efficiency features. The section covers the issues to be addressed in undertaking this work and provides examples of work undertaken. Increasingly looking to maximise the use of data, rather than collecting new data, is a major opportunity for all statisticians and thus this section aims to provide some valuable insight on the topic.

The project leader was IDAE (Spain) and the other contracting partners are ST - AT (Austria), CBS (The Netherlands), SORS (Slovenia) and DECC (United Kingdom), however all NSIs have provided inputs to the project.

## **REFERENCES**

The manual is published at:

<http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-13-003>



### NTTS 2015 Scientific Committee

Chairman: **Martin Karlberg**, European Commission – Eurostat

**Rein Ahas**, University of Tartu

**Silvia Biffignandi**, University of Bergamo

**Carola Carstens**, European Commission – Directorate-General Communications Networks,  
Content & Technology

**Piet Daas**, Statistics Netherlands

**Patrick Deboosere**, VUB

**Anders Holmberg**, Statistics New Zealand

**Beat Hulliger**, University of Northwestern Switzerland (FHNW)

**Risto Lehtonen**, University of Helsinki

**Ralf Münnich**, University of Trier

**Marianne Paasi**, European Commission – Directorate-General Research and Innovation

**Giuditta de Prato**, European Commission – Institute for Prospective Technological Studies

**Pilar Rey del Castillo**, European Commission – Eurostat

**Evelyn Ruppert**, Goldsmiths University of London

**Fritz Scheuren**, NORC at the University of Chicago

**Natalie Shlomo**, University of Manchester

**Marina Signore**, Istat

**Roxane Silberman**, Réseau Quetelet

### NTTS 2015 – Reliable Evidence for a Society in Transition

New Techniques and Technologies for Statistics (NTTS) is an international biennial scientific conference series, organised by Eurostat, on new techniques and methods for official statistics, and the impact of new technologies on statistical collection, production and dissemination systems.

The purpose of the conference is both to allow the presentation of results from currently ongoing research and innovation projects in official statistics, and to stimulate and facilitate the preparation of new innovative projects (by encouraging the exchange of views and co-operation between researchers – including the possible building of research consortia) with the aim of enhancing the quality and usefulness of official statistics and to prepare activities related to research in statistics within the European Framework Programme for Research and Development (Horizon 2020).