Big Data for European Statistics (BDES 2018)
14-15 May 2018, Sofia, Bulgaria

Minutes

Monday, 14 May 2018

Opening addresses

Marc Debusschere (Statbel, Belgium), Conference Chairman, opened the conference and greeted the participants.

Sergey Tsvetarsky, President of the Bulgarian NSI, thanked Eurostat, the speakers and everyone who contributed to the project. He stressed Big Data's potential to enrich official statistics, while also noting that it is still in the experimental phase. He underlined the need for the whole international statistical community to develop specific new skills, create a broad framework for cooperation with stakeholders and adapt to this reality and to the challenges and opportunities it offers. In this sense, the project's goal was not only to give an opportunity to exchange ideas, but also to foster new research areas and innovations in Big Data. Mr. Tsvetarsky confirmed his readiness to actively participate in their development in the statistical practice and noted that Bulgarian experts are already taking part in relevant international training courses.

Mariana Kotzeva, Director-General of Eurostat, thanked Eurostat, the BNSI, the CBS and all the people involved in the project. She recalled how the project came to be and some of the administrative and organizational challenges that were identified at the time. Today we know much more about Big Data and we have concrete results, but we also have more questions than before and we have to expand the community of practitioners in order to find the answers. Now we have to implement what we have achieved and think about the future steps. She recalled that there will be a DGINS on digitalization and Big Data in October 2018 and that a way needs to be found to disseminate all the results of the project until September. There will also be another ESSnet where the concept of smart statistics will be introduced.

Dominik Rozkrut, President of Statistics Poland, made a presentation on the topic "Big data in official statistics – where has the velocity gone?". He focused on the need to effectively manage the process of integrating Big Data within the NSIs. The major challenges identified relate to the organizations' internal limitations, such as insufficient resources, and the need to foster a culture of flexibility, innovation and continuous change. The importance of committing to priorities and coming up with a tangible product in the end, even an imperfect one, was also highlighted. One way of doing this is by tracking for accountability and monitoring the process in order to ensure that it delivers. As the NSIs have to integrate a tremendous amount of rapidly changing data

sources in order to stay relevant, the need to enter into strategic partnerships with stakeholders, including the private sector, becomes even more urgent. These partnerships are of mutual benefit because the NSIs are needed to validate the results.

Work package 1 – Webscraping job vacancies

Nigel Swier (ONS, UK), presented the main results of WP1, the participating countries, the methodology used, the approaches to data access, and the six major challenges in using online job vacancy (OJV) data for statistical purposes. The general conclusion was that OJV data is relevant in itself, but is not representative of the labour market and as such cannot replace the job vacancy survey. In addition, agreed access arrangements were recognised as generally better for obtaining statistical data than direct web scraping. It was underlined that collaboration with Cedefop is essential.

Vladimir Kvetan (Cedefop) presented a project, which aims at webscraping job portals from every EU Member State in all official languages and create a data set to help develop vocational training programmes. The idea is to get clear information on the type of skills employers need in order to inform the educational sector accordingly and improve public policies. By the end of 2018 early results will be released for seven countries, while the final version of the project will be finalised in 2020. Cedefop currently has a stock of 13.8 million vacancies that are still being treated.

Michail Skaliotis (Eurostat) introduced the discussion by mentioning the importance of the topic. Firstly, the system of online job vacancies is used actively by employers and job seekers and will become more and more important in the future. Secondly, in the minds of policymakers, this is an easy system to produce statistics, but in reality it is more complex. There are also many questions around data governance, and this conference provides an opportunity to give input to the DGINS meeting in October, when there will be a session on data governance issues and the DGs are expected to choose some priorities.

From the ensuing discussion it transpired that in the future we may see statisticians use a variety of sources according to what is the best one for each type of information that is sought. However, it may be difficult to segment the labour market in that way for OJV data. The need for the NSIs to help improve the general public's statistical literacy and to position themselves within the industry (e.g. Google, Facebook, etc.) was also underlined.

Work package 2 – Web Scraping Enterprise Characteristics

The 6 partners of WP2 (IT, BG, NL, PL, SE, UK) have been investigating the potential of webscraping, text mining and inference techniques to collect general information about enterprises. Compared to WP1, this involves more massive scraping of websites and treating more unstructured data.

Monica Scannapieco from the Italian National Institute of Statistics made the presentation. It was divided in 3 parts: 1) Objectives of the work & Use Cases; 2) Results (Full processing pipeline, Methods, IT solutions and Analysis of legal setting); 3) Final Remarks. The general objective is to access enterprises websites and use the information received by means of webscraping, text mining and inference techniques for statistics production and filling in the National Business register and for Business statistics surveys. The 6 use cases were presented: Use case 1: URLs Inventory, Use case 2: Web sales – Ecommerce, Use Case 3: Social Media Presence, Use Case 4: Job Advertisement, Use Case 5: Sustainable Development Goals and Use Case 6: Economic Activity Classification (NACE).

The main result is full processing pipeline, which consists of 4 parts: Internet access, Storage, Data preparation and Analysis. In the part of Internet access there is an URL searcher, which produces URL retrieved from the enterprises websites. The URLs go to the scraper, which is a software, which produces scraped content in the storage part. In the data preparation part the data is produced by means of methods like URL scorer, Tokenization, Word filters (eg. stopwords), Language specific lemmatization, Text Representation (Bag of Words, Language encoding and Engineered Features). Finally the analysis is done by Machine Learning (Build training & test sets, Training classifier, Application classifier) and Deterministic decision rules.
Regarding results in view of methods there is a distinction between specific and generic webscraping. Specific web scraping means that both structure and (type of) content of websites to be scraped are perfectly known. Generic web scraping means no a priori knowledge on the structure and content is available and the whole websites must be scraped and processed in order to infer relevant information.

The Text representation methods are: Keywords/engineered features: turning the text into features by simply searching for keywords related to the use-case – for example, 'basket' and 'shop' in the e-commerce use-case, subject matter expertise is needed. Bag of words & Language encoding/word embeddings: automatically generate features from the extracted text, no subject matter expertise is needed.

Among the methods deterministic and machine learning methods are used. Deterministic: algorithms are designed from a set of rules with known characteristics of the decision rule in mind. Machine learning (ML): decision rules/classifiers are derived from data (data-driven). Both deterministic and machine learning approaches produced acceptable results. Quality assessment problem is still open. They are not yet at the stage of recommendations for production.

Italy, Poland and Bulgaria developed IT solutions which shared with other countries. The IT solutions of Italy were used by Poland and Bulgaria. The Polish solutions were used by Italy, Netherlands, Sweden, Bulgaria. Bulgaria developed some software.
Regarding legal framework and netiquette the main issues are copyright protection, including protection of data bases as property, privacy and personal data protection and ethical principles for web scraping. From the ethics and legal framework point of view it is necessary to have in hand transparent web-scraping policies in order to allay public concerns about the data collected and the usage of them; national statistical laws and Copyright regulation also play enormous role. The web scraping challenge is to make the web scraping approach lasting in the longer term, it is

necessary to build up new competences, both for web scraping and for data storage at NSIs. Sharing software and knowledge is very important for competence building. From methodological point of view the challenge is representativeness: Large (especially international) enterprises that own multiple domains: need to identify where they conduct e-commerce; Web ordering though portals (e.g. booking.com); there is need for revised quality framework. Concerning data management pipeline whole processing pipeline has been concretely implemented for prototypes (24 prototypes implemented!) and there is need to move to production such efforts.

The discussion on the panel was opened by **Lilli Japec (Statistics Sweden)**. The main questions raised were: Can we webscrape enterprise webpages and gain information that can be used in production of official statistics? Can business registers be improved by using web scraping techniques to get information about enterprises? Can statistical outputs be produced with more predictive power by combining different data sources? It is a good approach to look at case studies in order to address the questions. More specific questions: 1. The survey estimate may not be the ”gold standard”. Have you tried to establish which method is the best one? 2. Would you recommend other NSIs to replicate any of your case studies? 3. How is timeliness affected using web data? 4. Is comparability between countries improved with the webscraped mode? The discussion resulted in unanimous answers that they would recommend other NSIs to replicate the case studies and comparability between countries is definitely improved with the webscraped mode.

Workpackage 3 – Smart meters

Toomas Kirt (Statistics Estonia) presented the main objectives and the results of the smart meters pilot study. The aim of the pilot is to evaluate whether smart meter data can be used to produce official statistics. In particular, the goal is to show whether current business statistics surveys can be replaced by smart meter data, to produce new household statistics, and to identify vacant or seasonally vacant dwellings. There are wide variations among the participating countries in terms of access to data. However, the main remaining challenges are to link the observed unit to the statistical unit, develop a methodology to distribute energy consumption where there are multiple consumers at the same address, and find information on own produced and consumed electricity. Electricity data has the potential to be used as a complementary data source for the production of new statistical products or to improve current survey-based statistics.

Anders Holmberg (Statistics Norway) said that smart meter data is a promising new source, while stressing the importance of finding a reasonable business case to invest in the production of new Big Data sources. In particular, it is important to evaluate the possibility of doing something better, faster or cheaper than when using existing practices and to think about which task is nearest to becoming a reliable product. Big Data also offers the possibility to compare old and new sources and add value or even replace existing statistical practices. The importance of having enough people working on this was also underlined. NSIs were urged to bring in their public relations departments in communicating the use of these new sources because of the confidentiality issues involved and the risk of receiving backlash. Lastly, he concluded that

surveys would not disappear, because the new sources contain less variables and it will take time to make tangible products.

Work package 4 – AIS data

Anke Consten (CBS, Netherlands) introduced the concept of AIS data and the aims of WP4, namely to investigate whether AIS data can be used to improve current statistics and to develop new statistical products. The sources used for this pilot initially included one European data source and land-based stations, while access to satellite and national data was received at a later stage.

Eleni Bisioti (ELSTAT, Greece) made an overview of some of the challenges related to the use AIS data, such as the unreliability of manually entered variables, possible data loss due to adverse weather or magnetic conditions, land-based receivers' limited capacity to detect signals beyond 40 nautical miles, and the fact that ships can turn off their AIS transceiver. The importance of using satellite data to complement land-based data when following ships across oceans was also stressed.

Christina Pierrakou (ELSTAT) presented the advantages of using AIS data to improve maritime and inland waterway transport information, including on the number of ships in a given port, their destination, the distances travelled, the routes, and port location. The possibilities to produce new statistics and economic indicators were also outlined, after which two examples of possible visualisations of ship movements were shown.

Anke Consten briefly presented the pros and cons of two possible scenarios for producing national and European AIS statistics. In the first case scenario, NSIs would make use of their own AIS national data sets, whereas in the second one NSIs would get access to a worldwide dataset and decode, clean and store data in a central database. The possible future areas of work on AIS data in the framework of the next ESSnet and the recommendations for taking full advantage of AIS data were also mentioned. The overall conclusion was that AIS can be a useful source to improve current statistics and generate new statistical products.

Faiz Alsuhail (Statistics Finland) expressed his conviction that AIS data can be useful for producing new, faster and timelier statistical products that can be particularly useful. The WP members have proved that this data can be stored and processed, and the project has built a critical mass of knowledge. One additional way of disseminating this knowledge would be through ESTP courses.

Work package 5 – Mobile Phone Data

WP5 has two main objectives: investigate and enable access to mobile phone data, and start developing use cases and appropriate methodologies. It is carried out by 9 partners (ES, BE, DE, FR, IT, FI, NL, RO, UK).

David Salgado (INE, Spain) made the presentation. There are 5 deliverables: Deliverable 5.1: Current status of access to mobile phone data in the ESS (July, 2016); 2. Deliverable 5.2:

Guidelines for the access to mobile phone data within the ESS (May, 2017); 3. Deliverable 5.3: Proposed elements for a methodological framework for the production of official statistics with mobile phone data (Feb, 2018); 4. Deliverable 5.4: Some IT elements for the use of mobile phone data in the production of official statistics (March, 2018); 5. Deliverable 5.5: Some quality aspects and future prospects for the production of official statistics with mobile phone data (May, 2018). He gave a definition for Big Data: Big Data for Official Statistics - refer to third people and not to data holders; are central in their economic activity; lack statistical metadata (since they are generated for very different purposes). UNECE (wider) Definition for Admin Data: "Data collected by sources external to statistical offices."

The process consists of 3 phases. For mobile phone data they are: Phase 1: Raw Telco Data Generation; Phase 2: Statistical MicroData Generation; Phase 3: Aggregated Data Generation; More phases: Inference. For administrative data the phases are: Phase 1: Admin Data Generation; Phase 2: Statistical MicroData Generation; Phase 3: Inference to Target Population.

As regards Geolocation of network events, the first approach is the Best Service Area Approach, which means which area is best served by each antenna. Other approach is the Bayesian Approach. From data to target population inference is done. The probabilistic sampling is not possible – this is the curse of representativity. Adaptation of species abundance problem in ecological sampling: Hierarchical model; Two working assumptions: At t0 individuals are assumed to be physically in the territorial cell of auxiliary admin/survey data. Mobility patterns of individuals do not depend on the concrete MNO they are subscribed to. The Bayesian approach has computational scalability and integration with other data sources.

About Quality issues CoP is affected by two generic facts: MNOs active part of the production process; Change of inferential paradigm; Higher degree of breakdown. Example: accuracy dimension - From confidence intervals to credible intervals and model checking and model assessment.

Main conclusions: Access is blocked: further work on perceived risks and collaboration is needed; Total Survey Error paradigm is still valid; Geolocation; No probability sampling: are hierchical models possible? Computational complexity for storing, accessing, and computing in situ; Quality Assurance Framework needs revision in view of active role of data holders and change of inferential paradigm.

Anders Holmberg from Statistics Norway opened the discussion in the panel. He raised the main question about official statistics using Mobile Phone Data and more specifically about data environment, production environment and methodological and technical environment. Further questions are about the access, assurances and resource trade-offs. What statistics are the most promising to make with Mobile Phone Data and why? A conclusion was made that mobile phone data are more mature than smart metres and AIS data.

Work package 6 – Early estimates of economic indicators

WP6 with 6 partners (IT, PT, FI, NL, PL, SI), aims to investigate multiple data sources of different types in order to produce early estimates, particularly for a consumer confidence index

and for nowcasting turnover indices. The WP6 coordinator **Tomaž Špeh, SURS** made the presentation. The presented issues were: WP6 objectives; Summary of activities carried out and results achieved; WP6 pilots: Early estimates of economic indicators ; Impact of big data sources on economic indicators (Correlation, Time lag, Selectivity, etc.); Quality assessment of the input, throughput and output phase of the process; Lessons learned - Data sources used and their specifics (Appropriateness, Accessibility, Availability in time); Data cleaning strategies; Publication lags & realistic vintages of data; Data pretreatment; Recommendations and future perspectives.

The main goal of the WP6 was to explore how a combination of (early available) big data sources, administrative and existing official statistical data could be used in creating an existing or new early estimates for official statistics: Exploration of big data sources and statistical areas where those sources could be used; Other administrative and statistical sources which could be combined with investigated big data sources; Implementing business cases identified in SGA-1 period; Data collection, data linking, data processing, methodological and IT issues; Examples of calculated concrete estimates for economic indicators with quality assessment of results.

The work done was: Investigate multiple big data, administrative and other existing sources; Collaboration with WP7 team -Joint meeting in Warsaw; Nowcasting the turnover indicators (April 2016) - One of the pilots that was started in WP6 SGA-1, Statistics Finland (basic proposal) , Statistical Office Slovenia; Finalizing the proposal for SGA2 (November 2016); Successes to have access to big data sources; First estimates of early economic indicators using also big data source were produced; WP6 coordination meeting, Ljubljana, October 2017, WP6&WP7 face to face meeting, Lisbon, April 2018. Draft deliverables were prepared.

Proposed pilot for WP6 SGA-2 - Title of the pilot: Early estimates of economic indicators. Potential economic indicators: Gross domestic product (GDP), Consumer price index (CPI), Industry production index (IPI) Retail sale, Balance of payments, Economic sentiment indictors, New leading economic indicators. The tasks were: to create and test the methodology of creating early estimates for at least one of the main economic indicators; to test the quality measures which assess quality of the sources, statistical production and statistical results; to investigate multiple Big data and other existing sources for purposes of early estimates of at least one of the main economic indicators. Many of the sources are available in most of the countries so it is possible to test them and create the results for more than one country.

The WP6 SGA-2 Pilots: IT - Use of electronic transactions of System of payments and of the Anti-Money Laundering Reports data on estimating private household consumption; FI - Machine learningapproaches for nowcasting GDP and TIO using firm-level data and traffic loops data; SI - Estimating early GDP and IPI using traffic loops data; PT - Predicting exports, based on Nights spent in tourism establishments; PL - Using internet data sources about the property market and job offers to forecast coincident and leading indicators; NL - Using Monte Carlo Markov Chain (MCMC) to clean the data, remove noise and solve the problem of missing data.

Future perspectives: NSIs can address a major quality issue, namely the timeliness, to form an initial quick estimate of the target indicator by using a range of micro level data sources accumulated well before the official release is made, by employing large dimensional econometric models. This does not necessarily lead to too large revisions, but adds significantly

to the quality of official statistics through timeliness dimension. This work can proceed in multiple directions: other data sources can be explored with the methodologies we have presented and possibly in relation to other indicators; other modeling frameworks are possible - quality measures with focus on precision criteria; a real-time application can be programmed, especially relying on the traffic loops data (daily estimates); these methodologies and their added value is not limited to nowcasting the GDP or some aggregate final indicators, but could be explored in order to impute some missing components of the aggregated figures; in case of dealing with a large data (traffic loops), a more novel approach for data cleaning and dimension reduction should be considered. By training subsets of data in different neural networks, the dimensionality problem is solved. Traffic loops inspected further by exploring clusters of the measurement points around designated areas (borders, manufacturing clusters, mining fields, etc.).

Lilli Japec, PhD, Statistics Sweden opened the discussion panel. The general question was: Can a combination of early available big data sources, administrative data and existing official data be used in creating early estimates of economic indicators? More specific questions were raised: Based on your experience, what are the main challenges for implementing early estimates in production? How do you communicate these types of early estimates with users? How do you justify e.g., the use of traffic density data in the model? Does the improved timeliness increase or decrease costs of production? How do we make sure that this great result is spread to all NSIs? The improved timeliness at first raises costs of production, but later the costs become constant. The discussion was closed with an opinion that this is the work package which produces the most effective data in the sphere of early estimates and that makes a big difference for governments and banks.

Tuesday, 15 May 2018

Prof. Dr. Diego Kuonen, CStat PStat CSci Statoo Consulting, Berne, Switzerland delivered a keynote speech: Production Processes of Officia Statistics & Data Innovation Processes Augmented by TrustedSmart Statistics: Friends or Foes? He cited Walter Radermacher that "official statistics is also confronted with a rapidly changing context and needs". He gave a definition for Big Data - The term "big data" applies to an accumulation of data that can not be processed or handled using traditional data management processes or tools. Big data are a data management IT infrastructure which should ensure that the underlying hardware, software and architecture have the ability to enable "learning from data" or "making sense out of data", i.e. "analytics" ("data-driven decision making" and "data-informed policy making"). The 5 Vs of Big Data are volume, variety, velocity, veracity and value. He stressed that the "Veracity" (i.e. "trust in data"), including the reliability ("quality over time"), capability and validity of the data, and the related quality of the data are key. Existing "small" data quality frameworks need to be extended, i.e. augmented. The value means the usefulness of the data. Prof. Kuonen considered the Internet of Things and underlined that IoT is about data, not things. He continued with a review what makes a smart city. He cited W. Edwards Deming that "Data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action." Prof Kuonen underlined that data are the fuel

and analytics, i.e. “learning from data” or “making sense out of data”, is the engine of the digital transformation and the related data revolution. Next he reviewed the two approaches of analytics statistics, data science and their connection. He focuses on the fact that both approaches of analytics, i.e. deductive and inductive reasoning, are complementary in order to enable data-driven decision making, data-informed policy making and proper continuous improvement. The essence of it is that it is primary analytics or top-down (i.e. explanatory and confirmatory) analytics. It is “Idea (hypothesis) evaluation or testing” and the analytics’ paradigm is “deductive reasoning” as “idea (theory)”. On the other hand, data science is typically concerned with analysing secondary data that have been collected (and designed) for other reasons. It is secondary analytics or bottom-up (i.e. exploratory and predictive) analytics. It is “Idea (hypothesis) generation” and the analytics' paradigm is “inductive reasoning” as “data first”. The conclusion is that we need both confirmatory and exploratory analytics. He stressed on the key importance of the veracity of data, i.e. the trustworthiness of data, including the related data quality, in a world of (big) data and IoT (data). Another conclusion made was that the analytics is an aid to thinking and not a replacement for it and that analytics should be envisaged to complement and augment (official) statistics, and not a replacement for it. Prof. Kuonen stated that technology is not the real challenge of the digital transformation and that digital is not about the technologies. He also mentioned his contributions: Glocalised smart statistics and analytics of things – Core challenges and key issues for smart (official) statistics at the edge, as well as Data Innovation Strategy of Swiss Federal Statistical Office. He underlined that culture change is key in the digital transformation. The professor ended the presentation with a review of process models for continuous improvement: The “Plan-Do-Check-Act” (PDCA) cycle, the related “Plan-Do-Study-Act” (PDSA) cycle, a process model for the production of official statistics - the “Generic Statistical Business Process Model” (GSBPM), which is consistent with the PDCA or PDSA cycles. GSBPM is a key conceptual framework for the modernisation (and standardization of the production) of official statistics. He concludes that it is necessary to move emphasis from measuring quality to improving quality. Moreover, the GSBPM is a deductive reasoning and a sequential approach. This process model needs to be adapted to incorporate data innovation by taking into account both approaches of analytics (i.e. inductive and deductive reasoning) and through the usage of, for example, data-informed continuous evaluation at any GSBPM step. He summarizes that the current production processes of official statistics need to be augmented and empowered by data innovation. According to him a process model for data innovation – the CRISP-DM (“CRoss Industry Standard Process for Data Mining”) process, which was initially conceived in 1996, is also consistent with the PDCA or PDSA cycles. Prof. Kuonen stated the difference and links between GSBPM (“current statistical production”) and the “Data innovation process model” (the complementary cycles of developing & deploying “analytical assets”). He concluded that new computational models must be adopted between private and public actors to guarantee mutual trust in the process. It is necessary to guarantee that data are processed for the agreed purpose, by the agreed method, respect of user privacy & business confidentiality, compliancy with legal provisions.

Work package 7 – Multiple Domains

It is carried out by five partners: Poland (leader), Ireland, Netherlands, Portugal (second wave) and United Kingdom. It was presented by **Jacek Maślankowski, Statistics Poland and Sónia Quaresma, INE Portugal.** They made brief overview of the results, discussed main findings and outlined future perspectives. There are 3 statistical domains: population, tourism and agriculture. They were reviewed separately, as well as intra- vs. inter-domain data combining. In the domain of population were searched: life Satisfaction by Twitter, everyday Citizen Satisfaction by Facebook and morbidity areas and personal well-being with Google Trends. In the tourism domain were searched: tourism accommodation establishments (data sources – various portals) and border movement (road traffic, air traffic, train traffic). The main findings in the tourism accommodation establishments were that it was difficult to determine if a given object meets the mandatory requirements related to the official classification of an object for a given type of facility and that they had to combine this data with official register of Tourism accommodation establishments. In the domain of agriculture it was made crop types identification by PL and IE. The used methodology was raw satellite image processing, segmentation of processed data and object based image classification (machine learning algorithms). There were three data sources: satellite data from Sentinel-1 and Sentinel-2 and in-situ survey. The intra-domain data combining gave the following results: Agritourism showed the impact of rural areas on development of agritourism places. There were three data sources: web scraping agritourist lodgings, satellite data on agricultural fields and land and buildings register on NUTS 5. The output was that the agritourist lodgings are mostly located on hilly areas ; the number of them is moderately correlated with area of forest land, meadows, and pastures, while weakly correlated with area of arable land, lakes and rivers; the number of agritourists is mostly related to area of lakes and rivers, then to forest land. It is negatively correlated with area of urban areas and arable areas and we may conclude that conditions of agritourism development are not the same as for agriculture. The main findings are that in the population domain the data sources are according to data quality and accessibility; two use cases in the tourism domain can be implemented with success – tourism accommodation establishments, border movement; the agriculture domain has one successful project – crop types identification that is compliant with in-situ survey on crop types; the population domain in Big Data is strongly related to the social statistics. The conclusions made by the package coordinator how to use WP7 Experience in Official Statistics were: life satisfaction pilots are shared on Github; tourism accomodation establishments can be implemented with the use of web scraping methods; border movement pilot can be implemented with a set of scripts with entropy-econometrics; agriculture domain can be implemented with two approaches by Statistics Poland and Statistics Ireland. The main future perspectives outlined were that Mobile Call Records and Call Detail Records are reliable information on Population (e.g., day and night population, number of commuters) as well as Tourism domain (e.g., number of tourists); tourism accomodation establishments – classification; border movement – to add additional data sources, e.g., AIS data; agriculture – extending to all regions.

Work package 8 – Methodology

Piet Daas (CBS, Netherlands) made an overview of the aim of WP8, which is to generalise the findings of the pilots and relate them to the conditions for future use of Big Data sources within the ESS. WP8 focused on the most important topics in the areas of IT, quality and methodology

based on the state of play in January 2017. One topic emerged in each of these areas, i.e. data processing life cycle for IT, process chain control for quality, and data process architecture for methodology. The three ways in which Big Data can be used for official statistics were also presented, namely to improve survey- and census-based statistics and as the main source for other statistics.

Jacek Maślankowski (Statistics Poland) presented the IT report. The report covers information on the Big Data processing life cycle, metadata management, format of Big Data processing, data hubs and data lakes, data source integration, choosing the right infrastructure, list of secure and tested APIs, shared libraries and documented standards, speed of algorithms, and training, skills and knowledge. There is also information on the types of data storages, GitHub repositories, and on how to clone software. The main findings are that there is no unified framework for big data metadata management and for data source integration, but there is a common point in all WPs on tools and data storage and there are best practices on using different APIs. In addition, data lakes and data hubs require further exploration. The literature study was also presented.

Magdalena Six (Statistics Austria) presented a report on the quality aspects of Big Data. 7 quality aspects were identified, including 5 in relation to the cause(s) of errors and 2 in relation to changes in the composition of the source. The quality aspects were assessed according to UNECE's Quality Framework for Big Data. The report includes 7 chapters, with one chapter for each quality aspect. The main challenges are that some error sources cannot be unequivocally classified and the comparability of Big Data sources over time becomes difficult because of the non-stable access to them and the changes in the technological process. Another major challenge is to develop a standardised quality framework due to the diversity of Big Data sources.

Valentin Chavdarov (NSI, Bulgaria) presented the reasons for having a specific methodology for Big Data and the main methodological issues involved. These include assessing the accuracy, changes in data sources, machine learning, data linkage, secure multi-party computation, sampling, data processing architecture, unit identification and others. These issues are different in terms of scope: whereas some are specific to Big Data, others cover almost all stages of the statistical production process.

Faiz Alsuhail congratulated the participants and said that the WPs have shown how to store data, preprocess it, remove noise from it and make it fit for analysis. However, the way to gain more information on these issues is through practice, mistakes and continuous learning. He praised the literature review as a very useful and timesaving tool to promote capacity building. The work done within the WPs has a very strong European dimension, but further consideration should be given to the best ways to disseminate the knowledge and promote its reuse. It is also needed to think about the technology issue at an early stage.

Big Data and official statistics: what's next?

Peter Struijs presented the current ESSnet project and the plans for the future one. The call for proposals for the new ESSnet is expected to be published very soon, and the idea is to launch the

new project around November-December 2018. The new ESSnet would include 3 WPs: implementation, new pilot projects and trusted smart statistics. The foreseen budget breakdown for each WP is roughly 40% for implementation, 40% for the pilots and 20% for smart statistics. The same countries plus Slovakia have expressed preliminary interest in taking part in the new ESSnet, but the numbers vary for each subtopic. An overall positive evaluation of the current ESSnet was made in terms of meeting its objectives and expectations, usability of the outputs and benefit-cost ratio, but there are still many areas to improve. Peter thanked the Review Board, Eurostat and all the participants.

Mariana Kotzeva started with an overview of the way official statistics have evolved from a process under the full control of the NSIs to the Internet of Things and today's "datafication" where many different sources are used. With the advent of smart technologies, the devices become subjects, not just objects, and this completely changes the relationship between NSIs and data sources. This trend has several major implications for official statistics, including the need to establish data partnerships, find new solutions to ensure privacy and security, create trust, and develop new skills. As the traditional business model cannot provide solutions to the new challenges, the new model should focus on trusted smart statistics. This new model involves more input data, more analytics and artificial intelligence, and more statistics embedded in smart systems. The concept of trust in all stages of the statistical production process is central to the new model.

Panel discussion – Big data and the future of the ESS

Peter Struijs opened the discussion by asking the panellists to comment on what would constitute success for the ESSnet.

Lilli Japec said that implementation is key to success. She underlined the need to combine Big Data with more timely economic indicators and adopt a more user-oriented perspective. She also mentioned that HBS and TUS are old-fashioned and could be improved with the use of Big Data, for instance, through credit card and transaction data and mobile phone data, respectively.

Dominik Rozkrut agreed that implementation is key and it should be explicitly set as a goal. The goals in general have to be clearly defined, practical and concrete and should be monitored through performance indicators. Statisticians within the NSIs should be encouraged to take advantage of the project's deliverables, although the biggest challenge might be to convince them that Big Data is not different compared to what they have been doing so far. Big Data should be seen as a tool that helps to increase efficiency, timeliness and accuracy in the process of producing the same statistics without changing the paradigm. The need to communicate this to users and stakeholders was also underlined.

Diego Kuonen also stressed the importance of continuously communicating to stakeholders and users what Big Data is really about and mentioned that this element was missing from the first phase of the ESSnet. It would be useful to prepare sales pitches to get the message both internally and externally. The other important element that should be kept in mind is blendedness, i.e. pilots that complement existing statistical products.

Mariana Kotzeva suggested creating a list of statistics produced with the help of Big Data sources at the end of the project. It is not necessary to produce new statistics, but to show what the concrete output is. The main challenge is that ESSnets have been implemented like research projects in the sense that the developers start to think about the implementation phase only after getting the results. With regard to the second package of the future ESSnet, it is important to think what the elements of success are. On the third package – smart statistics – the main challenge is to define the use cases, which need to be as concrete as possible.

Peter Struijs summarised the discussion by saying that the NSIs need to be outward-oriented and as concrete as possible.

On the issue of involving the **academic community** in the validation of specific results, **Mariana Kotzeva** agreed that this is a good idea, whereas **Diego Kuonen** expressed his preference for external experts because of academia's propensity to promote their own models.

When asked about what can be done to change the **organisational culture** of NSIs, all the participants agreed that this process needs continuous internal communication. According to **Dominik Rozkrut**, this is the most challenging task ahead because it is easier to come up with ready-made solutions than to integrate them. He gave the example of Statistics Poland, where the directors have to report to him on the innovations they have implemented in their area. In this sense, change needs to become natural to people, and the public service should not differ from the private sector. **Diego Kuonen** stressed the importance of understanding the organisational culture and of advertising new technologies as a way to complement current work. **Mariana Kotzeva** underlined the importance of setting clear goals, involving all levels and communicating on each step. In addition, people should be held accountable as a way to incentivise change. Because of the enormous scale of this task, stakeholders should be involved and experimenting should become part of doing business.

Asked about the **relationship between Big Data and surveys**, **Lilli Japec** said that surveys would always be needed because they study many different variables in comparison to Big Data sources. However, surveys may need to be carried out differently. This is why it is important to find the right methodology to combine data sources. She suggested adapting the Code of Practice to Big Data sources, but cautioned against devoting too much time and resources pondering the wording as this does not create value for taxpayers.

On the possible use of **co-funding** by private partners for other ESSnet projects, **Mariana Kotzeva** said that it is important to look for financial opportunities provided by other EU funds because Eurostat's budget is limited. Using research funds would also allow for statistics to become part of the EU's research agenda.

With regard to **international partnerships**, all the participants agreed that the NSIs need to partner with data owners because data access and data ownership are becoming strategic issues with which the NSIs cannot cope on their own. The advantages provided by the ESS and the EU, including in terms of ensuring coordination between the NSIs and global partners, were also

highlighted. Standards and practices should be exchanged within international organisations and statisticians should be trained to share knowledge.

Marc Debusschere thanked all the participants, the panelists and Eurostat and closed the meeting.