Workshop on the Quality of Multisource Statistics

Report

The ESSnet workshop on the Quality of Multisource Statistics was held in Budapest on 21-22 April 2016 with the aims of facilitating the exchange of experiences and know-how regarding the quality aspects of using administrative data in statistical data production, identifying possible gaps in the existing approaches and of giving inputs for the action plan of the ESSnet.

Three sessions were organized during the one and a half day of the workshop. Each of the sessions begun with presentations on the good practices of specific countries. After that, the Work Package leaders presented the results and the main questions related to their WPs, then participants discussed in small groups the questions related to the WPs that were sent to them prior to the workshop.

After the opening of the workshop, the ESSVIP Admin project was presented by Eurostat.

The ESS.VIP ADMIN project aims to facilitate the use of administrative sources without compromising on quality of the output. Work packages of the project were presented together with the method of their realization.

The ESS.VIP ADMIN project contains the following work packages:

1. The access to and the development of administrative data sources (on-going contract and future Task Force)
2. Statistical methods (contracts)
3. Quality measures for statistics using administrative data (ESSnet)
4. Eurostat as an (in)direct user of administrative data sources held or designed by the Commission (grants)
5. Frames for social statistics (contract, task force and partly included in ESSnet on quality of multisource statistics)
6. Pilot studies and applications (grants and contracts)
7. Methodological support to Member States

The on-going ESSnet on quality of multisource statistics covers the 3rd and partly the 5th WP (Task 5.2 Methodology for the assessment of the quality of frames for social statistics (belonging to WP5 Frames for social statistics))

The ESSnet on quality of multisource statistics was presented by the leader of the ESSnet consortium, Niels Ploug from Statistics Denmark.

Producers of official statistics have to face the issues of the reduction of costs and of respondent burden, while at the same time the need for reliable statistics is increasing. This situation pressured NSIs to use secondary data sources, thus the use of administrative data and multiple
data sources are becoming hot topics in the ESS. However, the NSIs have to guarantee the quality of official statistics independently of the input data sources. The ESSnet on quality of multisource statistics aims to give the framework for quality checks when several data sources are used. The ESSnet has the following objectives:

- Evaluating the quality of input data (WP1)
- Methodology for the assessment of the quality of frames for social statistics (WP2)
- Framework for the quality evaluation of statistical output based on multiple sources (WP3)

The ESSnet on quality of multisource statistics is realized by a consortium with 8 members: Denmark, Norway, the Netherlands, Austria, Italy, Lithuania, Hungary and Ireland.

**Presentation of good practices of specific countries**

*Presentation of “The Integrated production of Population Statistics in Finland” by Jari Nieminen*

The use of administrative data has a long tradition in Finland. Administrative data sources were presented, such as the administrative and statistical registers and the basic factors facilitating the increased use of administrative registers in Finland. Two examples of multisource statistics were the population census and the income distribution statistics. Advantages and disadvantages of the register-based statistical system were explained and a specific example on the quality aspects when using multiple data sources was shown.

*“Administrative data in the business statistics of the FSO” was presented by Claude Macchi.*

The use of administrative data in the Swiss statistical system and especially in the integrated production system of business statistics at the FSO was presented. The example of structural business statistics was explored. Advantages and difficulties of using administrative data were presented, as well as their impacts on quality along the quality dimensions of the ESS.

**Session on WP1 of the ESSnet: Checklists for evaluating the quality of input data**

*Presentation of the tasks and results of WP1 of the ESSnet on quality of multisource statistics: “Checklists for evaluating the quality of input data” by Regin Reinert, Statistics Denmark*

The main objective of this WP is to create a gross list of existing approaches to define methods for input quality check. The most promising methods will be tested by the consortium members. A consolidated checklist will be produced for the measurement of input quality regarding both dimensions and indicators. The final goal is to integrate the best of all existing methods. A commented repository will be created and made accessible for a wider audience on the CROS portal. Any gaps in the existing approaches discovered during the review will be described, and recommendations for necessary further actions to be undertaken will be put forward to Eurostat.

As a first step, more than 500 indicators were listed. Some indicators as undercoverage and overcoverage were present in all lists, but, a series of unique indicators were also enumerated.
The input quality indicators recommended by the ESSnet on Admin data were first mentioned and had high preference among selected indicators. Participants of WP1 selected indicators from the gross list to be tested in detail. The final list contains 16 quantitative indicators – covering 6 dimensions – that are currently being tested by Austria, Lithuania, Hungary and Denmark. The results will be presented at the Q2016 conference in Madrid.

Some participants expressed the need to have some guidelines on how to compare data coming from several sources.

**Work in small groups**

In the small groups, first the participating countries’ experiences with the quality assessment of input data were discussed and the most used tools were enumerated. It was emphasized that the usability of indicators is highly stakeholder dependent, that is, data producers and data users could have different perspectives and interests upon evaluating the quality of datasets. The communication with and the feedbacks sent to data owners, data producers, together with the political commitment on improving administrative data sources were mentioned as key issues. Furthermore, the questions of accessing administrative data and the referring legal issues are important challenges for most of the participating countries.

In many cases, the lack of metadata, coverage issues, the delays in data takeovers and the outliers cause difficulties. Referring to the coverage issues, a distinction should be made between the coverage from the data owners’ perspective and from NSIs’ perspectives. It became clear that in a number of countries there is no standard quality assessment procedures used for evaluating input data, instead, non-standardized methods used. A significant part of the participants uses the checklist developed by Daas et al.[1] for the assessment of administrative data sources or follow the recommendations of the BLUE-ETS project[2].

**Presentation of good practices of specific countries**

“Example of Slovenia: use of administrative data sources in statistical surveys” by Barbara Kutin Slatnar.

The presentation posed the question of why NSIs should combine different data sources and explained the characteristics of multisource statistics. Quality aspects of using administrative sources were discussed and some suggestions were made to solve emerging problems. Influences of a multisource model on the business processes were also discussed.

**Two presentations were made by the experts of ONS:**

„Development of a tool for quality assurance of administrative data” by Carolyn Watson.

The presenter first placed admin data in context with special attention to the quality aspects, then introduced the Administrative Data Quality Assurance Toolkit developed by the ONS. The

---


[2] [https://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf](https://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf)
toolkit makes possible to define the position of a dataset in the risk/profile matrix and in the quality assurance matrix that determines the necessary assurance level to be used for a given dataset. Finally, the operational context of using administrative data was traced, emphasizing the importance of communicating with data suppliers when assessing input data quality.

“Assuring statistical quality of Administrative data” by Charlotte Gaughan

The example of linking data collected by the ONS and administrative data for structural business statistics was presented, using data from the British National Tax Office (HMRC). Legal issues and the limitations of usability were traced and followed by the details and challenges of data linking. A special emphasis was given to data cleaning by means of editing/imputation methods. The main conclusion of the presentation was that data linking, cleaning and imputation methods are effective although further development of methods are required.

Session on WP2 of the ESSnet: Methodology for the assessment of the quality of frames for social statistics

Presentation of the tasks and results of WP2 of the ESSnet on quality of multisource statistics: “Methodology for the assessment of the quality of frames for social statistics” by Johan Fosen, SSB

WP2 aims to produce a guideline that facilitates the measurement of and the reporting on the quality aspects of frames for social statistics, including both the frames of persons and/or households, and the frames of addresses to access persons and households. In addition, quality measures and the associated methodology, which are beyond the scope of this WP, will be identified for future work.

In order to develop a theoretical framework for the assessment of the quality of frames for social statistics, the deliverables will cover, among others, a review of the existing literature and the current practices in the NSIs regarding the quality assessment of frames for social statistics; a gap analysis; an action plan detailing the tasks to be carried out (based on the gap identification).

The presentation clarified the definition of frames, the types of frames in social statistics, and the differences between frame errors and output errors. A description of different types of frame errors were given and related to sources for quality measurement.

Work in small groups

First of all, the question “what are the characteristics of a good frame?” was discussed. Major differences between social and business statistics were traced, such as the differences in the use of unique identifiers (in many countries unique identifiers are only available in the case of business statistics thus making the social statistics frames the main problem, whereas in the “register countries” the social statistics frames are considered the easier ones due to the changes in the Business population over time) and in the level of documentation and feedbacks. A
suggestion was made to modify the distinction of business and social statistics to institutional and population statistics.

The coverage issues were mentioned as the most important dimension of the quality of frames, that countries are aware of, however they lack the adequate tools for the measurement as well as the treatment of them (e.g. in the case of population statistics, international migration is one of the most challenging difficulties). Intentions to measure the coverage problems usually are based on post enumeration surveys and on the census. Another type of frame errors measured by the participants is the domain classification error. The frame error type ‘contact information’ (due to e.g. multiples addresses, mixed business and residential buildings) were mentioned as a field where there are challenges and effort is made to improve the quality in cooperation with phone companies. Additional problems are represented by temporary addresses (e.g. students living abroad), and information on mobile phones.

**Presentation of grant results within the VIP Admin project**

*“The Challenges in compiling an Education Register in Iceland from multiple sources” by Ásta M. Urbancic*

The presentation started with explaining the need for a new register of educational attainment and with the enumeration of the sources on which the new register was set up. Among the challenges affecting the quality aspects of the data, the coverage issues, the lack of a single code for educational attainment, missing information and changing qualification requirements were mentioned together with some recommendations on how to resolve these. As regards the quality of the register, the aspects of relevance, accuracy, coverage, coherence and comparability over time and space were examined. The suggested method for dealing with quality issues is the development of deterministic data editing processes.

*“Use of administrative data in Polish agriculture statistics” by Artur Łączyński*

The presentation offered insights into the experiences of using registers for agricultural statistics in Poland. Administrative data sources as well as the quality related issues were discussed in details. As regards the latter, data sources were evaluated along the dimensions of content and coverage, conceptual congruities of definitions, update frequencies and the questions of accessing data (legal aspects, timeliness, channels of data transfers, etc.). Data cleaning was emphasized highlighting the need for comparing the data before and after data cleaning. Recommendations for data owners were also detailed. The presentation ended with the challenges and the future of using agricultural registers.

*“Possibilities to use administrative data sources for register based Census in Latvia” by Pēteris Veģis*

The presentation started with the legal issues of accessing admin data in the production of official statistics and then covered the use of administrative data in two main areas: use of administrative data in the estimation of the number of population on one hand; and
administrative data sources that characterize the economic activity of the population on the other. As regards the former, a logistic regression model – based on the population register and other administrative sources – was presented for the estimation of the status of the actual place of residence for every registered resident in Latvia. Regarding the latter, data from the LFS and a series of administrative data sources were compared on the current activity status, on occupation, on the branch of economic activity and on the status in employment.

Session on WP3 of the ESSnet: Framework for the quality evaluation of statistical output based on multiple sources

Presentation of the tasks and results of WP3 of the ESSnet on quality of multisource statistics: “Framework for the quality evaluation of statistical output based on multiple sources” by Ton de Waal, CBS

The aim of WP 3 is to produce a set of quality indices that measures the final output and incorporates the impact of all steps, based on a source of error framework and coherent with GSBPM. The first step is to identify different situations for which quality is needed to be measured. ESSnet is reviewing critically the existing quality measures and approaches to evaluate and compare the quality of the output based on several sources (among which at least one is administrative source however, in the perspective of quality we generally have to face very similar problems using statistical data set collected by an organization other than the NSI.). Suitability tests of existing and proposed quality measures and approaches will be carried out in several domains. An important result of these tests will be an analysis describing the gaps between the current situation with respect to the quality evaluation of statistical output based on multiple sources and an envisaged “ideal” situation for the near future. Suitability tests will include also a cost-benefit analysis to help Member States to determine which quality measures and quality approaches should be used for a given situation.

For the different ways of combining data sources, different quality measures (and methods to apply them) have to be used. The presentation explained the basic data configurations identified in the ESSnet and gave examples and some of the most relevant challenges for each configuration.

Work in small groups

As regards the proposed data configurations, the participants mostly agreed that they cover quite well the possibilities with some exceptions. However, the first type was considered too simplistic and optimistic. Other mentioned configurations were: combining two administrative sources (with partial overlap or not), repeated cross-sectional surveys, time series at the micro level. It was added that for the classification, the cooperation of the statisticians responsible for a given subject matter would be also needed. Among the methods for the measurement of the output quality of multisource statistics, the comparisons of different sources and of the
estimations based on them seem to be the most salient. Users’ feedback was mentioned as of crucial importance. Moreover, the use of editing/imputation and validation methods, as well as the checks of time series were discussed. According to some of the participants, there should be indicators for each step of the data production, of which a global measurement would be calculated and, on the other hand, before starting the quality measurement processes, it should be clarified if the quality evaluation will be user or producer oriented. A key question regarding the quality of the output is what the bias and the variance will be when sources other than surveys are used.

**Conclusions of the workshop**

Participants had more comments and answers for the questions of WP1 than for the questions related to WP2 and WP3. Models and questions referring to frames should be commented by methodologist dealing with sampling, while models related to WP3 could be assessed by the statisticians who have a deep knowledge on the subject matter areas. Models and questions will be circulated in different working groups, task forces and on the cross portal for comments.

For WP1 the most important outcome is that in most countries there is some input quality check however not standardised nor systematic. For the input quality assessment a 2 steps process would be appropriate. The first step refers to the checks to be carried out before a decision is made on the specific use of the input dataset and it should consist of only a few indicators with special attention on the details of the accessibility of the data source. The second step would be a deep assessment of the data source.

As far as the frames are concerned, the existence of a population register is of crucial importance. The most common problem with frames is the coverage which is known, but difficult to quantify. In most cases NSIs have qualitative information on frames.

The models described in WP3 were assumed to be useful for the assessment of output quality. However to have a more complete view on the output quality, processes should also be assessed.

Accessibility of administrative data was discussed in each small groups even though it was not included in the questions. This topic is very relevant for most of the countries. Accessibility and usability of administrative data have different levels. The first level is the physical accessibility of datasets. The question is whether the NSI actually receives the datasets or not. The legal background is highly important, but in several cases, although statistical laws prescribe the accessibility of administrative data for NSIs, the transmission of datasets cannot be realised. Cooperation with data owners is crucial from several aspects.

The second level is the quality of the data. It covers the availability of metadata, coverage, accuracy aspects, existence of identifiers and legal and technical possibilities of linking different data sources. Regular communication between data owners and statistical institutes is essential.

The third level is the integration of statistical needs into the registers. In this phase there is a close cooperation between the owners of registers and the statistical institute. Register owners
consult statisticians before setting up a new register or modifying or terminating an existing one. Statistical needs are taken into consideration.

The level of accessibility to administrative data varies not only over countries but over data owners or datasets. Some countries meet difficulties to get a dataset while for another administrative register their needs are satisfied.

**AOB**

1. *Data integration project of UNECE*

Some on-going projects are related to the topic of the ESSnet. One of them is Data Integration project of UNECE which aims to gain experience in data integration by pooling resources in joint practical activities and to translate experiences into general recommendations for data integration and to provide initial guidance for a quality framework. Participants of the workshop were asked to send comments.

2. *Data validation tool*\(^1\) of Eurostat was presented in the workshop.

---

\(^1\) [https://webgate.ec.europa.eu/fpfis/mwikis/ESSValidServ/index.php/Main_Page](https://webgate.ec.europa.eu/fpfis/mwikis/ESSValidServ/index.php/Main_Page)