**Administrative data and quality -Guidelines towards better quality of administrative data**

Sorina Carmen Vâju, Eurostat, Sorina-Carmen.VAJU@ec.europa.eu

Mátyás Tamás Mészáros, Eurostat, Matyas.MESZAROS@ec.europa.eu

**Abstract**

*Statistical authorities need to produce data faster in a cost effective way, to become more responsive to users´ demands, while at the same time providing high quality output. One way to fulfil this is to make more use of already available data sources, and in particular administrative sources, most typically used in combination with other sources.*

*Depending on the use of the administrative sources and the data configuration different statistical tasks must be applied. Usually it is not only one task but a sequence of different tasks that have to be applied, for example, data integration, imputation and editing or tabulation. For these tasks different methods are available and depending on the input data quality and the data configuration the same method can have limited use or produce lower quality outputs.*

*The use of administrative data sources risks impacting negatively quality on several dimensions, in particular accuracy and comparability. Surveys and administrative sources have both particular strengths and weaknesses. Combining them may overcome these weaknesses, provided that suitable methodology and tools are used. At the same time, harmonised measures of quality for outputs that combine administrative sources with other sources (surveys) are necessary to ensure that European Union official statistics are of sufficient quality and fit for their intended use.*

*This paper looks at the most frequent methodological challenges faced when integrating administrative sources and provides, for typical situations, preferred methods to have the best quality of statistical output. It also introduces the work of ESSnet on the Quality of Multisource Statistics (KOMUSO) to develop quality measures and guidelines related to the use of administrative sources.*

**Keywords:** Administrative data, quality, guidelines, integration

**1. Introduction**

Statistical authorities need to produce data faster in a cost effective way, to become more responsive to users´ demands, while at the same time providing high quality output. One way to fulfil this is to make more use of already available data sources, and in particular administrative sources, most typically used in combination with other sources. During the last years, the statistical offices have made important progress towards using more administrative data in order to replace or complement surveys; however this raises important methodological challenges: namely to integrate administrative data into the statistical process and to maintain proper quality of the output.

Depending on the use of the administrative sources and the data configuration different statistical tasks must be applied. Usually it is not only one task but a sequence of different tasks that have to be applied, for example, data integration, imputation and editing or tabulation. For these tasks different methods are available and depending on the input data quality and the data configuration the same method can have limited use or produce lower quality outputs.

The use of administrative data sources risks impacting negatively quality on several dimensions, in particular accuracy and comparability. Surveys and administrative sources have both particular strengths and weaknesses. Combining them may overcome these weaknesses, provided that suitable methodology and tools are used. At the same time, harmonised measures of quality for outputs that combine administrative sources with other sources (surveys) are necessary to ensure that official statistics are of sufficient quality and fit for their intended use.

Since 2015, the European Statistical System (ESS) has undertaken coordinated efforts to answer these challenges and to improve the use of administrative sources for the production of official statistics. The ESS Vision Implementation Project on Administrative Data Sources[[1]](#footnote-1) (ADMIN) belongs to the ESS Vision portfolio and aims at supporting Member States to make increased or better use of administrative sources, as specified in the Vision for European Statistics[[2]](#footnote-2). The project builds on previous ESS initiatives, like the MEETS program[[3]](#footnote-3) (and in particular the ESSnet Admin Data[[4]](#footnote-4)), the BLUE-ETS project[[5]](#footnote-5) or the ESSnet Statistical Methodology -- Area ISAD" (Integration of Survey and Administrative Data)[[6]](#footnote-6).

During its lifespan (2015-2019), the ADMIN project has a dual purpose: to support the EU Member States to reap the benefits (decrease costs and burden, increase of data availability) of using administrative data sources for the production of official statistics, and to support the quality assurance of the output produced using administrative sources, in particular the comparability of the statistics required for European purposes. It aims to facilitate the use of administrative sources across the ESS, by improving the access to administrative sources, by improving methodological knowledge needed for integrating administrative data in statistical production and by providing tools for assessing the quality of outputs based on administrative sources. The project will also support Member States in implementing these theoretical outcomes in concrete statistical areas. The results of the work on the ESS Vision 2020 ADMIN will also be put at the service of ongoing major projects such as the modernisation of social statistics and the census 2021.

The methodological core of the project aims to answer two questions:

* what statistical methods to use when integrating administrative data in production (covered in Work Package 2. Statistical methods)?
* how to assess the quality of the administrative sources and the quality of the output that is based on such sources (covered in Work Package 3. Quality measures for statistics using administrative data)?

The progress made on these two questions is introduced in sections 2 and 3 of the paper.

**2. Guidelines on the use of estimation methods for the integration of administrative sources**

The increased use of administrative data sources in statistical production reduces the response burden and the cost of data collection, but at the same time it can reduce the quality of the output in several dimensions, like accuracy, timeliness and comparability. Using the appropriate methodology the risk of these negative effects can be reduced.

In order that the various administrative sources have positive effect on the quality of the different statistical outputs, in the framework of the second work package of the ESS Vision 2020 ADMIN project, Eurostat coordinated the work on drafting guidelines for the use of estimation methods for the integration administrative sources in the statistical production processes. The draft guidelines try to provide some advice on how the different data sources, including administrative sources, can be integrated in the statistical production to have the best outcome under several conditions[[7]](#footnote-7).

The ideal situation would be to provide a consistent decision tree which classifies all possible situations, and provide the best methods for integrating the administrative data with other sources. Unfortunately, the topic is very complex and covers a large number of cases which depends on many conditions, therefore only the most relevant cases are covered by the guidelines. In order to handle the different options the structure of the guidelines follows the Generic statistical business process model (GSBPM) adapted to the specificities of administrative sources:

2*.1. Define purpose of using administrative data*

First, in the data integration process, it is important to determine the planned usage of administrative sources. It can be understood in a broad sense. It may be a statistical table, an indicator, or a quality measure for a certain variable. The envisaged result defines so called usages of administrative sources. The statistical question has to be seen in connection with the subject-domain in which the question is asked because the domain(s) can have important influence on the use of the different estimation methods under consideration.

2*.2. Check the type and availability administrative data and other data sources*

The next step is to identify the available data configuration(s). For obtaining the result there can be different data sets available. In the guidelines, only those cases are taken into account where the data are at least partially administrative. The possible data configurations used in the guidelines are based on the ESSnet on quality of multisource statistics (KOMUSO)[[8]](#footnote-8) which is described in section 3 in more detail.

The guidelines use eight possible data configurations:

* Configuration 1: Combining non-overlapping microdata sources without coverage problems, “split-variable” case;
* Configuration 2: Combining non-overlapping micro-data sources without coverage problems;
* Configuration 3: Combining non-overlapping micro-data sources with part of the variables is in a single source, without coverage problems;
* Configuration 4: Combining overlapping micro-data sources without coverage problems;
* Configuration 5: Combining overlapping micro-data sources with under-coverage;
* Configuration 6 Combining macro-data sources;
* Configuration 7: Combining a micro-data source with a macro-data source;
* Configuration 8: Combining longitudinal data sources.

Depending on the data configuration different statistical tasks must be applied. Usually it is not only one task but a sequence of different tasks has to be applied, for example, data integration, editing and imputation or tabulation. For these tasks different methods are available and the decision about the method depends on properties of the data.

These different configurations are of utmost importance for classification of the application of statistical methods and for defining a precise workflow for the different estimation methods. For example, the application of editing and imputation methods has to be seen in relation to the data configuration. Also the application of different kinds of data integration procedures/methods depends essentially on the data configuration.

2*.3. Check the quality of the input dataset*

Before the different data sources are integrated into a unique data structure for answering the question stated at the beginning, when defining the purpose of using administrative data, the quality of the available input data sources has to be evaluated based on the available metadata. Depending on the information collected, it may be that the workflow has to be modified and additional steps have to be taken to address the quality issues of the input datasets.

The input data quality defines the essential conditions for the selection of the different statistical methods to define how to use administrative data in statistics. The measurement of administrative data quality can be addressed by the checklist delivered by the ESSnet KOMUSO. The [Checklist for Evaluating the Quality of Input Data](https://ec.europa.eu/eurostat/cros/system/files/essnet_wp1_report_final_version4.pdf)[[9]](#footnote-9) of the ESSnet KOMUSO is described more in details in Section 3. The guidelines are quite general and applicable to any statistical source. The ESSnet makes references to previous results (like ESSnet Admin Data or BLUE-ETS project), where further indicators, both quantitative and qualitative, are described, according to the general paradigm of data quality. The set of indicators, of course, is too large to be applied entirely in a quality check of the input datasets, but some of them could be chosen where the objective of the quality measurement has been established. In some cases specific problems of administrative data can require ad hoc indicators and methods to evaluate how to transform administrative data into statistical one.

2*.4. Quality issues with administrative data*

Regarding the acquisition of administrative datasets, the agencies responsible for a certain statistical product have to face the problem that generally they do not control the production of the administrative data that can be affected by misclassification errors, missing items and inconsistencies. While for missing data and inconsistencies standard methods can be used for measurement and correction, misclassification errors generally requires ad hoc methods. Hence once the agencies have decided to use the administrative data, it is assumed that they know the problem connected to the input data and develop appropriate procedure to handle with these errors, for example by “modelling” the difference between the original data and the reference statistical ones.

Administratively created definitions and classifications may change over time causing new and different misclassification. Therefore, another problem is the "stability" of the classifications and definitions used by other administrations, and the NSIs should be informed about any changes. If they are stable, the adopted methods to deal with adjustments hold in the following survey rounds, the adjustments are more or less constant over the time and can be cancelled in calculating variations. The methods to measure the stability are related to metadata; for this reason it is suggested to collect and to manage metadata in a structured way developing "a regime" of procedures for analysing them from one round to another round.

Not only classifications and definitions can change but the changes could be made without informing the statistical agencies because of different reasons: the external administrations do not recognize their importance for statistics, especially for the less visible ones, or the official documentation is lacking or no stable relations are established between agencies and other administrations.

2*.5. Integrate data and check output*

After integrating the datasets the remaining tasks can be executed to achieve the purpose defined at the beginning of the workflow (see *2.1 Define purpose of using administrative data)*. The output of the workflow has to be evaluated as well according to predefined quality criteria, which highly depends on the quality of the production process. If the results are not satisfactory, then other options have to be chosen for performing the above mentioned tasks.

For evaluation of the results one can use the quality framework for the evaluation of statistical output based on multiple sources which is under development in Work Package 3 of the KOMUSO project, discussed briefly in the following section.

**3. Quality of multisource statistics**

Complementary to the work package 2 on statistical methods, Work Package 3 of the ESS Vision 2020 ADMIN deals with the assessment on quality when statistical output is based on multiple sources, e.g. when administrative data are combined with surveys or with other administrative sources in the production of statistics. An increased use of administrative data sources entails the risk of impacting negatively on the output quality. Of particular concern are the following quality dimensions:

* The relevance of statistics might decrease, if administrative concepts replace statistical concepts, as the administrative concepts may not always answer users' needs.
* Depending on the type and quality of administrative sources, the accuracy of the statistics can be affected positively or negatively. Different factors come into play and they should be taken into account. While information from administrative data sources can be less impacted by memory effects or effects of social desirability, the accuracy of statistics might also decrease, in particular, when the owner of the administrative data source has little incentive for recording and updating of some information or the incentive for registration are not good.
* While generally administrative data sources have the benefit of producing timelier results (e.g. register-based censuses report results considerably earlier than traditional censuses), using these sources may also decrease timeliness in particular, if the administrative data are not of required quality when statisticians need them (e.g. tax data).
* The comparability of statistics risks could be reduced considerably, if the national administrative concepts widely replace the statistical concepts. Such conceptual discrepancy could reduce the comparability of the data at national level and EU level.

It is often the case that administrative data need to be transformed before being used as input for official statistics. Most of these transformations are related explicitly or implicitly to the use of estimation methods. For example, one can quote data linkage and matching, data classification, data crossing, data forecast, or data modelling techniques as possible statistical methods that enable the use of administrative sources. The use of these methods is both a solution addressing the deficiencies of the administrative data (different concepts, classifications, lack of timeliness etc.) and a source of variation in the quality of the statistical output. Making sure that the right methods and tools are available and correctly applied will bring great benefits when integrating administrative sources in statistical production. While the work on statistical methods is covered in the "Guidelines on the use of estimation methods for the integration of administrative sources", the ESSnet on Quality of Multisource Statistics complements it by offering guidance on 1) the evaluation of the input data (administrative data); 2) on the evaluation of the output (statistical product), after the methods for integration have been used

*3.1. ESSnet KOMUSO*

In order to make progress on quality, the ESSnet on Quality of Multisource Statistics (KOMUSO) was set-up to implement the third work package of the ADMIN project. This work package aims to develop and promote quality measures for evaluating the quality of administrative data and of the statistical outputs that use a combination of sources, among which administrative sources. Big data is excluded from the scope of this work package. This work package has evolved in parallel with the work package on statistical methods, and the progress in one work package has progressively fed back in the other one.

The broad objectives are to gather existing knowledge on quality assessment and reporting and review it critically; provide up-to-date guidelines based on expert consensus on quality assessment in statistical production (input and output); develop new indicators for the quality of the output based on multiple sources and a methodology for reporting on the quality of output; and finally, to produce recommendations for updating the ESS Standard and the ESS Handbook for Quality Reports. Additionally, the ESSnet covers also tasks linked to the quality of the sampling frames for social statistics.

A framework partnership agreement[[10]](#footnote-10) is working towards achieving the following objectives during a four year period (08/2015 –08/2019):

* Take stock of the existing knowledge on quality assessment and reporting and review it critically in order to produce recommendations on the most suitable approaches;
* Develop new indicators for the quality of the output based on multiple sources;
* Produce a methodological framework for reporting on the quality of output;
* Produce indicators relating to the quality of frames themselves and the data whose production is supported by frames;
* Produce a methodological framework (guidelines) for assessing the quality of the frames used in social statistics; draft a proposal for minimum quality requirements for sampling frames for EU social statistics;
* Produce recommendations on updating the ESS Standard and the ESS Handbook for Quality Reports.

*3.2. Input quality assessment*

The ESSnet has already accomplished a review current practices on input quality assessment and a proposal of an input quality checklist. This work builds on previous projects and collected and reviewed existing methods for assessment of input quality. Additionally, a checklist for quality assessment was developed on the basis of the most promising indicators from the reviewed methods. The checklist has been tested with respect to usefulness and computability on several types of data by several NSIs participating in the ESSnet. These results[[11]](#footnote-11) can be used in the process of combining administrative sources with other data, at the step of checking the input data quality (see section 2.3 of the current paper).

*3.3. Output quality assessment*

The ESSnet has also reviewed existing knowledge on output quality assessment, and has conducted suitability tests for existing measures and indicators for output quality. The ESSnet is currently working on developing guidelines for integrating administrative data in statistical production, as well as associated quality measures and indicators. The reader is referred to the other papers in the Special Session: Quality of multi-source data at the Q2018 conference, which present the preliminary results of the ESSnet on output quality.

**4. Conclusion**

When several sources are used in statistical production, it is not easy to quantify or describe output quality in a simple way, as there are many ways of integrating the various sources. Both Member States and Eurostat should be able to assess the quality of these new possible sources before integrating them in the statistical production system, as well as the quality of the final statistical output. Quality assessment, both at input and output level, is necessary to certify that European Union official statistics are of sufficient quality and fit for their intended use, in particular in terms of comparability and accuracy. However, the set of ESS quality indicators currently defined is not always applicable or feasible for assessing the quality of multiple source statistics. Most current indicators are survey-oriented, while the particular quality issues typical to integrating administrative sources in statistical production are not fully accounted for. The multiple-source integration is a new statistical paradigm that needs to be reflected in quality standards and quality reporting and assessment, and the ESS.VIP ADMIN project has taken key actions in this direction.

**5. References**

European Statistical System, ESS Vision 2020. Available at: <http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020>. (Accessed: 07 May 2017).

Essnet on Quality of multisource statistics. Available at: <https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en>. (Accessed: 07 May 2017).

ESSnet Statistical Methodology -- Area ISAD (Integration of Survey and Administrative Data). Available at: <https://ec.europa.eu/eurostat/cros/content/isad-0_en>. (Accessed: 07 May 2017).

[ESS Vision 2020 ADMIN project (Administrative data sources)](https://ec.europa.eu/eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources_en). Available at

<https://ec.europa.eu/eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources_en>. (Accessed: 07 May 2017).

MEETS programme – results. Available at: <http://ec.europa.eu/eurostat/statistics-explained/index.php/MEETS_programme_-_results>. (Accessed: 07 May 2017).

BLUE-ETS. Available at <https://ec.europa.eu/eurostat/cros/content/blue-ets_en>. (Accessed: 07 May 2017).

Use of Administrative and Accounts Data for Business Statistics. Available at: <https://ec.europa.eu/eurostat/cros/content/use-administrative-and-accounts-data-business-statistics_en>. (Accessed: 07 May 2017).

Guidelines on the use of estimation methods for the integration of administrative sources. Available at: <https://ec.europa.eu/eurostat/cros/content/task-212-guidelines-use-estimation-methods-integration-administrative-sources_en>. (Accessed: 07 May 2017).

1. <https://ec.europa.eu/eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources_en> [↑](#footnote-ref-1)
2. <http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020> [↑](#footnote-ref-2)
3. <http://ec.europa.eu/eurostat/statistics-explained/index.php/MEETS_programme_-_results> [↑](#footnote-ref-3)
4. <https://ec.europa.eu/eurostat/cros/content/use-administrative-and-accounts-data-business-statistics_en> [↑](#footnote-ref-4)
5. <https://ec.europa.eu/eurostat/cros/content/blue-ets_en> [↑](#footnote-ref-5)
6. <https://ec.europa.eu/eurostat/cros/content/isad-0_en> [↑](#footnote-ref-6)
7. <https://ec.europa.eu/eurostat/cros/content/task-212-guidelines-use-estimation-methods-integration-administrative-sources_en> [↑](#footnote-ref-7)
8. <https://ec.europa.eu/eurostat/cros/system/files/final_report_wp3.pdf> [↑](#footnote-ref-8)
9. [Checklist for Evaluating the Quality of Input Data](https://ec.europa.eu/eurostat/cros/system/files/essnet_wp1_report_final_version4.pdf) [↑](#footnote-ref-9)
10. The partnership agreement was signed with a consortium made of eight ESS members (Denmark –coordinator, Ireland, Italy, Lithuania, Hungary, Netherlands, Austria, Norway). [↑](#footnote-ref-10)
11. [Checklist for Evaluating the Quality of Input Data](https://ec.europa.eu/eurostat/cros/system/files/essnet_wp1_report_final_version4.pdf) [↑](#footnote-ref-11)