Statistical Journal of the IAOS 31 (2015) 203–211
DOI 10.3233/SJI-150902
IOS Press

Measuring output quality for multisource statistics in official statistics: Some directions

Mihaela Agafiţei*, Fabrice Gras, Wim Kloek, Fernando Reis and Sorina Vâju
European Commission – DG EUROSTAT, 5 rue Alphonse Weicker, Luxembourg

Abstract. Many statistical offices have been moving towards an increased use of administrative data sources for statistical purposes, both as a substitute and as a complement to survey data. Moreover, the emergence of big data constitutes a further increase in available sources. As a result, statistical output in official statistics is increasingly based on complex combinations of sources. The quality of such statistics depends on the quality of the primary sources and on the ways they are combined.

This paper analyses the appropriateness of the current set of output quality measures for multiple source statistics, it explains the need for improvement and outlines directions for further work. The usual approach for measuring the quality of the statistical output is to assess quality through the measurement of the input and process quality. The paper argues that in multisource production environment this approach is not sufficient. It advocates measuring quality on the basis of the output itself – without analysing the details of the inputs and the production process – and proposes directions for further development.

Keywords: Quality measurement, multiple sources, big data, administrative data, integration

1. Introduction

Administrative sources have always been used in statistical production. Some areas have been based almost exclusively on administrative sources – birth, crime or bankruptcies – while some other domains have always been based on an integration of sources – national accounts, external trade statistics, balance of payments statistics, education statistics or health statistics. However, most statistical domains are survey based where probability theory is applied to make inferences about the whole population. In this context, administrative registers have been used frequently as sample frames for surveys.

Lately, most National Statistical Institutes (NSI) have been moving towards an increased use of administrative data sources for statistical purposes, both as a substitution and as a complement to survey data. Important drivers have been the need to reduce the response burden on primary data suppliers (i.e. individuals, businesses), to reduce the costs of raw data collection and to provide new or more detailed statistics.

More recently, the emergence of big data has brought a qualitative change to the data sources potentially available for official statistics [5]. The abundance of data brings new opportunities for fulfilling the demand for new statistics, more detailed statistics and more timely statistics alike. These big data sources are very diverse [24], ranging from social networks (Internet searches, blogs, pictures), electronic traces from IT systems supporting business processes (credit cards records, e-commerce transactions, bank transfers) to machine generated data (road sensors, satellite images, mobile phone geo-location data). How these data sources are used depends on the questions that the statisticians wish to answer. Such sources can be used on their own for producing indexes or to model human behaviour. However, it is usually the case that big data need to be combined with other sources in order to produce official statistics. Firstly, official statistics in many areas are drawn as parameters of well-defined standard finite populations (e.g. persons resident in the country, enterprises registered in the country and so

*Corresponding author: Mihaela Agafiţei, European Commission – DG EUROSTAT, 5 rue Alphonse Weicker, L-2721 Luxembourg. Tel.: +352 4301 34372; E-mail: mihaela.agafitei@ec.europa.eu.

on). In this case, the new data often suffer from selectivity bias in relation to those standard populations and therefore need to be corrected by combining them with other non-biased sources. Secondly, official statistics are often designed to support policy making while big data sources, not being designed to support such policy making, only partly answers to the statistical needs.

As a result of all these developments, statisticians increasingly look for ways to combine sources and methods in order to increase their ability to face statistical demands. The result is that more and more statistical outputs are based on complex combinations of sources and methods.

In a world where data are abundant, the comparative advantage of official statistics is the existence of thorough quality checks. However, in this new environment, it is possible that the advantages of integrating administrative and big data sources (e.g. burden reduction, timelines, more detailed statistics, and so on) are offset by possible decreases in the quality of the final output (especially comparability in the case of administrative sources and representativeness in the case of big data sources). On top of that, official statistics face the risk of not being able to properly assess the quality of their output.

While the way official statistics are produced has changed significantly through the use of multiple sources, the way we measure their quality seems not to have kept up. This paper is arguing that the set of quality measures which are currently used in official statistics (European Statistics Code of Practice [10] and related Quality Assurance Framework [9], ABS Data Quality framework [2], Data Quality Guidelines of Statistics Canada [23]) are not sufficient for assessing the quality of multiple source statistics. In this paper we do not question the output quality dimensions; we just argue that some of the dimensions are difficult to measure in the case of complex multisource production processes.

One approach to measuring the quality of the statistical output is to assess quality through the measurement of the input and process quality. The underlying idea is that high quality input and high quality processing should guarantee high quality statistical output. The Total Survey Error (TSE) framework [3] decomposes the total error by identifying the main sources of error in the process and by describing their contribution to the total error. Similar approaches have been proposed to statistical production based on administrative sources [18,27]. The framework has been further extended to big data by considering the generic steps involved [1]. These approaches reflect the concerns of the producers of official statistics. An assessment of the statistical production process provides not only an indication of the quality of the final statistical product, but also assists the producer in improving the process.

Users have different needs when it comes to measures of quality of the statistical products [27]. Their concern is not whether better sources or better processes should have been used for the production of a particular statistical product, but whether the final output is suitable for their purposes.

The current paper argues that in multisource production this approach of looking at the input and process is generally not feasible. In theory, the approach could be further extended to the output resulted through combining multiple sources; however the situation becomes very complex when several sources are integrated. This article gives an overview of the current status of the work related to measuring the quality of output based on multiple sources and proposes some directions of further investigation. It advocates measuring quality on the basis of the output itself without analysing the details of the inputs and the production process. A summary measure of output quality is not only relevant for managing the statistical production process but it is especially relevant for quality reporting to the users of official statistics.

2. Quality of statistics – general discussion

2.1. *Quality facets*

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. During the last few years, there have been attempts to develop specific quality indicators for assessing the usability of an administrative source [4,6] and of the big data [21].

Process quality can be interpreted in several ways. In a first approach, the process quality refers to the quality of the various intermediate steps that are necessary to produce statistics. For each of these steps, process quality assessment can be envisaged to detect possible weaknesses of the process. Through such process quality analysis, statisticians could envisage some measures in order to improve the output quality. For example, if a non-response-rate is too high, it could be envisaged to improve the training of the interviewers or the mode of data collection. For a complete picture on the different processes one can refer to Generic Statistical Business Process Model (GSBPM) [25]. Nevertheless, even if a quality indicator can be defined at each step of a given statistical process, the set of all indicators will not provide any clear insight on the quality of the final statistical output. They will mainly provide information for monitoring the quality of the production process. All the same, the quality indicators of the statistical process could help assessing the output quality. For example, the non-response rate in itself will not give an overall measurement of statistical output accuracy, but estimating the variance introduced by using some imputation methods in order to deal with the non-responses or data non-availability would give information on overall accuracy.

The second approach for assessing process quality is to analyse whether the statistical production process projects the desired quality features over the statistical outcome or whether it needs to be improved by making different use of the available sources. This second assessment is very relevant when administrative and big data sources are integrated, because it indicates how they can be used to improve the statistical production process. These two perspectives can be distinguished when speaking about quality process but they are complementary and respond to different needs: the first perspective is mainly relevant from a producer point of view (total quality management) whereas the second is mainly pertinent from the user point of view (influence of the statistical process on the final output quality). Laitila et al. note [18] that the literature overlooks the first perspective of the process quality assessment.

Output quality refers to the final statistical product and it should provide the user with easy to understand information on the quality of the final data. The users need (i) to gather a direct and global assessment of the output quality and (ii) to have the possibility to compare it across domains, time and/or space. For survey based statistics, the underlying sampling theory is well developed and there is a sufficient consensus on measuring survey errors. Statistical institutes have developed quality frameworks for measuring the output quality which are largely derived from the particular context of survey based statistics.

For the outputs based on big data or administrative sources, some indicators can be redefined and implemented using the same reasoning (e.g. under/over coverage rates, non-response rates, etc.). The situation becomes very complex when several sources are integrated. In this environment, assessing quality at input level and process level is still relevant but it is not enough. There are many ways to combine sources and the quality assessment should capture not only the impact of the sources but also of the statistical technique used for integration. If we refer to international/regional statistical aggregates, an additional layer of complexity is given by international comparability and the wider range of combinations of sources and methods. This paper focuses on the evaluation of output quality and tries to identify strands for progress.

2.2. *Quality frameworks and quality indicators*

In a modern statistical production environment where several sources are integrated, it is not an easy task to quantify or describe output quality in a simple way. There are many ways of combining sources. A survey can make use of administrative registers for building the frames and for designing the sample. Data coming from registers can be used as such at micro level, either replacing individuals or variables in the sample. Other registers can be used as auxiliary information at micro and macro level (imputation, estimation, post-stratification). It is also possible to apply sampling techniques to get estimations from administrative sources and to combine them with estimations from a survey. Big data sources can be combined with survey data for small area estimation [22], used for temporal and spatial disaggregation of estimations from other sources or for now-casting [14]. Also, data from several surveys can be combined, either pooled [15] or matched (statistical matching [19]); that means that administrative sources and/or registers are used as auxiliary information/ benchmark for the success of the technique.

Statistical institutes rely on quality frameworks for measuring and communicating the quality of their statistical outputs. They identify dimensions of quality, translate the quality dimensions into quality indicators and specify how the indicators should be calculated. Most quality frameworks cover the same aspects of quality, although they may be classified in different dimensions. The European Statistical System (ESS) framework [9] for quality is discussed in the following paragraphs, as it is more familiar to the authors.

The ESS quality framework [9,10] identifies five quality dimensions to describe output quality: (a) relevance, i.e. the statistics meet the needs of users; (b) accuracy and reliability, i.e. statistical outputs accurately and reliably portray reality; (c) timeliness and punctuality, i.e. statistical outputs are released in a timely and punctual manner; (d) coherence i.e. statistical outputs generated by process that use the same concepts (coherence between annual and sub-annual statistics, coherence between labour force and with National Accounts etc.) and comparability i.e. it is possible to combine and make joint use of statistical outputs coming from different sources but referring from the same data items (comparisons over time, across regions or domains); (e) accessibility and clarity, i.e. statistical outputs are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance [12]. Other quality frameworks, for example the ABS Data Quality framework [2] or the Data Quality Guidelines of Statistics Canada [23] identify the following dimensions: relevance, accuracy, timeliness, accessibility, interpretability and coherence. Interpretability is virtually equivalent to the ESS concept of clarity, while coherence incorporates the two related ESS concepts of coherence and comparability. All these frameworks have been an input to the recent United Nations' generic National Quality Assurance Framework [26].

These dimensions are translated into quality indicators which are provided in quality reports. The ESS quality reports [9] consider in detail the quality indicators suitable for survey based statistics, as well as some indicators for data collection based purely on administrative data. Big data is not even considered, while the quality measurement of outputs based on multiple sources needs to be further developed.

The impact of integrating multiple sources on how to measure the quality dimensions is diverse. On the one hand, this integration has no effect on some quality dimensions and how they are measured; this is the case for relevance and accessibility. Other dimensions, such as timeliness and punctuality may be affected, but the way they are measured is still appropriate. All these quality dimensions describe the statistical product irrespective of the statistical process behind (i.e. irrespective of the choice of data sources to be used, of statistical processing and of the integration approach). On the other hand, some quality dimensions require incorporating the effect of sources, methods and the integration approach; this is the case for accuracy and reliability, and coherence and comparability. The more complicated production process has an impact on how these dimensions are measured. Finally, the quality dimension clarity is not affected in the description of the final output, as this is irrespective of the sources and processes. Clarity is affected, however, as regards the description of the more complicated production process; for this reason statistics based on multiple sources will generally be less clear to the user.

As explained above, the deviations and the correct use highly depend on the quality of sources and of the way they are combined. At each step of the production process [8], accuracy and comparability appear as the quality dimensions that are actually at stake. For instance, the error measures are very much dependent on the statistical method applied and the statistical assumptions on which they are based: coefficient of variation for sampling theory; false match and false non-match rates for record linkage, Frechet bounds for statistical matching; fitness of the model for statistical modelling, etc. Moreover, even if some accuracy related quality indicators (e.g. coverage rate, edit failure rate, imputation rate, average size of revisions, etc.) can apply to different types of data sources, it is very difficult to assess their impact on the accuracy of the final statistical output. 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/or are, regardless the data source, produced using different methodological approaches/combinations. Generally speaking, comparability is often reduced by conceptual and methodological differences of the statistics under consideration and the lack of comparability is evaluated using relevant metadata that thoroughly describe the concepts and methods used. At international level, this translates into a huge number of possible sources of lack of comparability, given by combinations of: (i) national legal and institutional environments – that translates into a variety of ways of dealing with the shortcomings of the administrative data and big data sources relative to both concepts and quality dimensions; (ii) acceptable trade-off between quality dimensions at national level – that could not converge into an acceptable trade-off at international level; (iii) appropriate trade-off between costs and benefits in terms of output data quality at national level – that could hamper international comparability; (iv) methodological choices to integrate the several data sources – applying common methodological approaches is a strong guarantee for international comparability. Applying different national methodological approaches requires us to check their comparability and whether the aggregates are meaningful.

3. Output quality assessment on the basis of input and process

The natural approach for identifying the possible impact of combining several types of data sources on quality is to look at each step of the production process and assess the impact of such integration.

Kloek and Vâju [16] discuss the uses of administrative data in the context of such an integrated production system and propose a typology composed of the following 5 types of uses: (i) direct use of administrative data at micro level (statistical unit). This includes two basic cases: "using specific records from the administrative source to replace other ways to collect the data (horizontal) and using specific variables not included in the survey (vertical)"; (ii) use of administrative data in combination with other sources at micro level (statistical unit). It is important to note that, in a highly integrated system, the distinction between primary data and auxiliary data is becoming less and less relevant; (iii) use of administrative data in combination with other sources at aggregated level (groups of statistical units). Here "the strength of a survey is combined with the strength of an administrative source"; (iv) use of administrative data as source for the population frame; (v) use of administrative data as circumstantial evidence for validation purposes. The use of administrative sources for providing aggregate data is implicitly included in (i).

Florescu et al. [5] discuss how big data can be combined with other sources. The uses do not differ substantially from the possibilities available for administrative sources. Big data sources can be used in statistical matching, record linkage, estimation (e.g. nowcasting), calibration and small area estimation. One additional motivation for source combination in the case of big data sources is often – even if not always – its selectivity bias, which is not that prominent in the case of administrative sources. In cases where the selectivity bias cannot be shown to be small, big data needs to be combined with other non-biased data sources.

Combining several data sources mainly impacts the measurement of accuracy and comparability. The way of assessing the other quality dimensions is not affected by the type and number of integrated data sources. Comparability assessment can be to a large extent reduced to structural error generated by some statistical bias. Possible outliers/breaks in data sets can be detected based upon existing methods. This will, as illustrated in paragraph 4.1, 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. 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. Indeed, possible dependencies between all the risks that could occur at each step make this task difficult to achieve. Moreover, it is possible that the errors in different sources cancel each other out. If one focuses only on assessing the output quality, some feasible approaches can be pointed out. Focusing on accuracy essentially means to deal with problems related to the estimation of the measurement error, which could be broken down into bias and variance error estimation.

Contrary to survey data, the treatment of administrative data and big data from a probabilistic point of view appears to be less natural. This is because randomness does not play an important role when dealing with non-survey data, whereas when dealing only with survey data, the random mechanisms are explicit. Therefore, accuracy assessment of the combination of sources should most likely focus on aggregating random mechanisms effects with the bias effects which are present both in surveys (e.g. non-response bias) and non-survey data.

4. Direct output quality assessment

The diversity of ways to combine sources and relevant methods to produce official statistics makes it difficult to make a summary assessment of output quality. This difficulty is even higher for international aggregates, as at European level, by aggregating and comparing over 28 separate production processes in the Member States. The standard quality report produced within the Quality Assurance Framework of the European Statistical System is a rich source of information, but it is still not enough to cover the complex production processes involving multiple sources and methods.

In this section we discuss possibilities to assess accuracy and comparability of statistical outputs without the need to analyse the data integration process behind it. This is not to argue that such analysis will lose its relevance. The information on the quality of the different process steps will always be needed (i) for the identification of the main sources of inaccuracy and incomparability; (ii) for the design of the process: how to use

Table 1
Risk and corresponding impacted quality dimension when combining multiple data 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 concepts/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

the scarce resources in an optimal way; (iii) for monitoring the process in order to detect potential risks and direct corrective actions and (iv) for contributing to the overall assessment of the statistical output quality.

Three cases will be envisaged in the following sections: 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.

4.1. *Direct assessment of output quality from the output itself*

There are several ways to assess the output quality on the basis of the output itself. Here we list the simplest examples.

In a time series with sufficient length, the deviations from the estimated deterministic part (i.e. 'trend') can give an indication of the total variance of the process. Unfortunately, this variance does not just correspond to the measurement error, since it contains part of the actual variability of the time series which follows a stochastic process. Nevertheless, estimation of the overall process variance can give an upper bound for the variance error component. This is useful i) for checking any design based estimates of variance, ii) as a rough estimate in case that design based estimates are missing and iii) as a reference value for the outlier detection. The same method can be applied for cross-sectional data after standardising the values with a relevant variable (e.g. per capita) to correct for differences in scale.

Reported breaks in series can give a hint on the size of possible bias. In this context, a break in series is understood strictly as resulting from changes in sources and/or methods, for instance: changes in the questionnaire, in sample design, in the coverage of administrative sources or in the definition of the variables. The change in sources and methods is taken as a ready-made experiment showing the sensitivity of statistical output to changes in the process. These breaks in series are normally flagged by statistical offices and information about the changes is reported in the metadata. The quantitative impact is sometimes assessed over a certain period, when a double production system is kept. However, this is not always possible (changes in external sources) or feasible (too expensive). In such cases, the quantitative impact has to be estimated, for instance with the help of extrapolation.

The use of breaks for the assessment of bias is not straightforward as they might not be reported, for instance when the quantitative impact is low and/or the process owner is unaware of the break. This will give the impression that the bias is always significant. Moreover, the interpretation of a break is not always straightforward, as in a redesign often several changes are done to sources and methods, which makes it difficult to establish the impact of the separate elements.

Revisions concern production processes that deliver provisional data in order to meet timeliness requirements and that are followed by a second or even third scheduled release. Obviously, subsequent releases take into account more data (e.g. higher response rate, additional sources, etc.). Revisions due to errors should rather be treated as breaks in this context and not as revisions as these errors are the result of an unforeseen variation of the production process and give information on the sensitivity of final output. The series after revision is generally expected to have a lower variance. The difference in variance between the different series is an indication of the impact of further data on the variance. In case the revisions show systematic corrections, this would be an indication of bias. Once the bias is detected, it can be used to improve the methods for deriving the provisional estimates.

Outlier detection techniques can be applied to both time series and to cross-sectional data, for instance, to the comparison over countries at the European level. The EU Member States are very different in size and economic structure. Meaningful comparison requires a suitable way of rescaling (for instance by inhabitant or by Gross Domestic Product [GDP]). An example is Luxembourg as an outlier in GDP per capita which is mainly due to the high proportion of cross-border workers that contribute to Luxembourgish GDP [11]. Outliers can have a real interpretation, but they might also be the result of hidden lack of comparability. The outlier detection method is not entirely neutral. Extreme values are more likely in smaller countries due to higher variance and in case of imbalanced economic structures depending on only few dominant branches of industry.

The main advantages of the direct assessment methods 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. That is why any knowledge on the statistical production process remains important.

4.2. Assessment of output quality with a common reference data source

Under the header of assessment of output quality with a common reference data source, two cases are distinguished. The first case is the quality survey; the second is any other reference source.

The quality survey is deliberately designed to measure quality and therefore drawing conclusions about accuracy and comparability is straightforward. Among the advantages 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 production process; (iii) it is easy to summarise into an overall assessment. However, it requires running in parallel a completely separate process in order to avoid that errors that occur in the normal statistical production process will be repeated, which implies important additional costs and burden. Additionally, the variance often risks being too high to be conclusive. Concluding, the quality survey is an important instrument in theory, but 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 related statistics with considerable conceptual harmonisation, as administrative sources or big data sources. The related statistics are commonly used in the analysis phase of the statistical production process to offer a benchmark to the statistics produced. For instance, the number of unemployed persons obtained from a statistical survey could be benchmarked against the number of unemployed persons based on an administrative register. The investigation of administrative sources that could serve as a reference point has only recently started and it will almost certainly be extended to big data sources. 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 last disadvantage can be partly compensated if information on the separate sources before integration remains available.

4.3. Methods derived from bootstrapping

A European collaborative project proposed several possibilities to quantify the error measure for multiple sources statistics [7]. We refer here to the most common case where survey data is combined with administrative data at micro level, i.e. some data in the sample are directly surveyed while some other data are coming from other sources, basically administrative data.

For survey-based statistics, the mean square error (MSE) is a common measure of accuracy, combining structural error (bias) and random error (variance). As the sample survey estimators are especially designed to be unbiased or approximated as unbiased, the variance estimate is seen as an estimate of the MSE. Therefore, sample based statistics are accompanied by variance estimates (e.g. coefficients of variation), which give users an indication of the impact of sampling error on the accuracy of the statistics. In practice, it should nevertheless be pointed out that non-responses and possible lack of representativeness of the sample can also introduce a bias in surveys. When non-survey data, such as administrative data or big data sources, are intended to be integrated into survey data, it can lead to the introduction of (i) bias through modelling, linkage, conceptual differences, and (ii) non-sampling errors like coverage, selectively bias and misclassification errors. Other non-biased measurement errors in administrative data or in big data sources should also be taken into account. Therefore, when combining various statistical sources, a global error measure should include all these possible influences. Identifying and quanti-

Table 2
Application of bootstrap methods by type of combination of administrative data with other sources

Possible use	Application of bootstrap	Remarks	Main practical problem
In statistics production as a replacement for primary and/or complementary data	Yes	Existence of overlapping survey data is welcome and 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.
As a sample framework and source of auxiliary information in sample design or as input for statistical registers	Partially	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
As a source of additional variables to be used for estimates or as auxiliary information to support processing of primary data for example: editing, imputation, calibration of estimates	Yes	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.

fying each possible source of error related to a multisource statistic is the first step to achieve. Once this first step is reached, if combination of the different error sources is linear or could be linearised under plausible hypotheses (mainly independence between the possible sources of error), variance can be analytically estimated. Otherwise, in case of complex non-linear combination of possible error sources, bootstrap and other related simulation techniques consist as a natural candidate in order to estimate uncertainty of the statistical output.

The collaborative project [7] proposed ways to adapt the bootstrap re-sampling methods in order to estimate the root mean square error (RMSE) that includes both sampling variance, other non-biased measurement errors and possible biases coming from the various used sources. First step consists in incorporating the effect of interaction between sources (survey data, big data and administrative data) in a statistical model that needs to capture all possible error sources. At a second stage, bootstrap methods enable inserting randomness through the simulation or replication of samples [13,17] in order to estimate the probability density related to the estimation of the parameter of interest. An example on poverty rate variance estimation through the use of bootstrap methods is given in [20]. Two main constraints have nevertheless to be filled in before using this type of methods. The first one deals with the ability to model correctly the uncertainty, in particular for administrative data, before simulating it or replicating it, whereas the second one concerns the computation costs that could hamper the use of such techniques. The decreasing cost of computing is making this easier; at the same time, the availability of more data makes it tempting to apply the techniques to an ever increasing amount of data.

The degree of feasibility of bootstrap methods depends on the type of use of the administrative data. Table 2 presents, for the three main uses of administrative data, an assessment of whether it is feasible to apply this approach.

Concluding, once randomness of non-survey data can be modelled and/or replicated, the use of bootstrap methods can lead to the calculation of root mean square error that gives an accuracy assessment. The major advantage of bootstrapping is that it can be applied to almost all of the possible cases with a minimal prior knowledge of the underlying distribution. Even if bootstrapping is a relative simple statistical technique, it has the major disadvantage of requiring very high computing resources. At European level, this problem of computing requirements becomes even more serious due to the huge volume of data. Moreover, the possible required micro data might not be available at the European level due to confidentiality concerns.

5. Conclusions

The paper argues that the way quality is measured in official statistics should evolve in order to offer adequate quality measures when multiple sources are integrated in the statistical process. Our concern is the measurement of output quality in order to provide a relatively simple and easy to understand message on the output quality to both the external users and the statisticians themselves.

The usual approach for measuring the quality of the statistical output is to assess quality through the measurement of the input and process quality. The paper identifies the main areas of risk in the process when multiple sources are integrated. This approach is useful when trying to improve the production process. The process knowledge helps to build an integrated framework necessary for the assessment of the output quality of multisource statistics. However, we argue that focusing on the process is not enough in a multisource production environment.

The quality information should be summarised in such a way that data users can assess the accuracy and comparability. The complexity is increased by the number of sources and the steps needed to integrate and process them. On top of that, errors at each step cannot be assumed to be independent. In order to provide a meaningful assessment of the output to the user, the paper recommends measuring quality on the basis of the output itself, without analysing the details of the inputs and the production process. The paper distinguishes three alternative approaches that do not depend on the design of statistical process: (a) direct output assessment; (b) a common reference source; (c) bootstrapping. The suggested approaches need further development. It is also important to reflect on how difficult their implementation is in order to allocate scarce resources between measuring quality and improving it. This is beyond the scope of this paper and needs further investigation.

Disclaimer

The contents of this paper do not necessarily reflect the opinion or position of the European Commission.

References

[1] American Association for Public Opinion Research (2015), “AAPOR Report on Big Data”, AAPOR Big Data Task Force.

[2] Australian Bureau of Statistics (2009), “*ABS Data Quality framework*”

[3] P.P. Biemer, Total survey error: Design, implementation, and evaluation, *Public Opinion Quarterly (Oxford Journal)* **74**(5) (2010), 817–848.

[4] BLUE-ETS, Deliverable 4.1, List of quality groups and indicators identified for administrative data sources.

[5] D. Florescu, M. Karlberg, F. Reis, P. Rey Del Castillo, M. Skaliotis and A. Wirthmann, Will ‘big data’ transform official statistics? *Q2014 – European conference on quality in statistics* (Vienna, 2014).

[6] ESSnet “Use of Administrative and Accounts Data in Business Statistics” (2013), Deliverable 2.4: “*Guide to checking usefulness and quality of admin data*”.

[7] ESSnet “Use of Administrative and Accounts Data in Business Statistics” (2013), Deliverable 6.3: “*Guidance on the accuracy of mixed-source statistics*”.

[8] ESSnet Data Integration (2011), Report on WP1: *State of the art on statistical methodologies for data integration*.

[9] Eurostat (2009), ESS Standard for Quality Reports, *Office for Official Publications of the European Communities*, Luxembourg.

[10] Eurostat (2011), European Statistics Code of Practice, *Office for Official Publications of the European Communities*, Luxembourg.

[11] Eurostat (2013), GDP per capita in purchasing power standards in 2012, *Eurostat NewsRelease*, 190/2013–12 December 2013, Luxembourg.

[12] Eurostat (2014), ESS Handbook for Quality Reports, *Office for Official Publications of the European Communities*, Luxembourg.

[13] C. Girard, The Rao-Wu rescaling bootstrap: from theory to practice, *Federal Committee on Statistical Methodology Research Conference*, 2–4 November, Washington DC, 2009.

[14] H. Choi and H.R. Varian, Predicting the present with google trends, *Economic Record* **88**(s1) (2012), 2–9.

[15] M. Karlberg, F. Reis, C. Calizanni and F. Gras, A European toolbox for a modular design and pooled analysis of social survey programmes, *NTTS 2015* (New Techniques and Technologies for Statistics), 10–12 March 2015, Brussels, Belgium.

[16] W. Kloek and S. Vâju, The use of administrative data in integrated statistics, *NTTS 2013* (New Techniques and Technologies for Statistics), 5–7 March 2013, Brussels, Belgium.

[17] L. Kuijvenhoven and S. Scholtus, Bootstrapping combined estimator based on register and sample survey data, *Discussion paper 201123 of Statistics Netherlands*, The Hague/Heerlen, 2011.

[18] T. Laitila, A. Wallgren and B. Wallgren, Quality Assessment of Administrative Data, *Research and Development – Methodology Reports from Statistics Sweden*, Örebro, Sweden, 2011.

[19] A. Leulescu and M. Agafiţei, Statistical matching: A model based approach for data integration, *Office for Official Publications of the European Communities*, Luxembourg, 2013.

[20] I. Molina and J.N.K. Rao, Estimation of Poverty Measures in Small Areas, http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S3P3_ESTIMATION_OF_POVERTY_MEASURES_MOLINA_RAO.pdf.

[21] M. Petrakos, G. Sciadas and P. Stavropoulos, *Accreditation procedure for statistical data from non-official sources*, (study commissioned by Eurostat), 2013.

[22] M. Pratesi, C. Giusti, S. Marchetti, N. Salvati, F. Giannotti and D. Pedreschi, Area level SAE models with measurement errors in covariates: an application to sample surveys and Big Data sources, *SAE 2014* (Small Area Estimation 2014), 3–5 September 2014, Poznan, Poland.

[23] Statistics Canada (2009), “*Statistics Canada Quality Guidelines*” – the fifth edition.

[24] UNECE (2013), “*Classification of Types of Big Data*”.

[25] UNECE (2013), “*Generic Statistical Business Process Model*”, the fifth version.

[26] UNECE, “*National Quality Assurance Frameworks*”.

[27] A. Wallgren and B. Wallgren, *Register-based statistics: administrative data for statistical purposes* (Vol. 553). John Wiley & Sons, 2007.