**Conceptualising quality for big data**

Giovanna Brancato, Italian National Statistical Institute (Istat), brancato@istat.it

Loredana Di Consiglio, Italian National Statistical Institute (Istat), diconsig@istat.it[[1]](#footnote-1)

**Abstract**

*Whereas the importance of big data for official statistics is widely recognised, the quality of the big data is a concern and the statistical community has soon started to reflect upon a framework to assess their quality. For instance, the United Nations Economic Commission for Europe (UNECE) Big Data Quality Task Team extended an administrative data quality framework; the AAPOR total error framework for big data focused on an extension of the total survey error examining sources of errors specific for big data. Reis et al. (2016) analysed a few alternative approaches on real case applications, concluding for the need to complement the different approaches and to structure the links between input and throughput quality and output quality. In the ESSnet Quality of Multisource Statistics (Komuso), a framework relating the output quality with the sources of errors is being proposed. The proposal takes into account the main quality models for administrative data.*

*This paper aims at investigating how all the factors that have been considered in the literature can be integrated in a global quality frameworks for big data.*

**Keywords:** quality hyper-dimensions, sources of errors, statistical process

**1. Introduction**

The increase in the use of the big data for statistical production has led the scientific community to focus on several quality-related aspects. The literature can be roughly split into three main streams. The first stream aims at extending the standard quality concepts to the big data production (Struijs and Daas, 2014; UNECE, 2014; AAPOR, 2015; Baker R., 2017; Cai and Zhu, 2015). The second one considers the errors arising from specific applications (Lugomer *et al.* 2017; Hsieh and Murphy, 2017; ESSnet Big data, 2018; Wang *et al*., 2012; Barcaroli *et al.*, 2018). Finally, a third stream is focused on a framework in view of the accreditation of the source and the assessment of its usability (Eurostat, 2014). Reis *et al*. (2016) analysed a few alternative approaches on real case applications, concluding for the need to complement the different approaches and to structure the links between input and throughput quality and output quality.

A complementary approach acknowledges the importance of the process quality. Tao and Gao (2016) introduce a set of quality factors for the big data validation systems, e.g. performance, security and robustness. Cai and Zhu (2015) introduce the concept of “audability”. A way to insure the quality along the production process is to apply audit checks in the relevant phases of the data life-cycle. Cai and Zhu (2015) identify these relevant phases in the data generation, the data collection and the data use. They also outline a quality assessment process with a feedback mechanism.

In this paper, we focus on the European Statistics quality dimensions and we attempt to systematise all the elements that have impact on them, with a major attention to the statistical sources of errors generated during the process. This work benefits from the work on multisource statistics carried out in the ESSnet Komuso (Brancato and Ascari, 2018). In addition, here other factors, e.g. the institutional environment and contextual elements, are considered.

The formalisation of the sources of errors influencing the output quality when using big data is also aimed at supporting the process of alternative data sources selection, based on factual quality criteria.

**2. Impact of the input and throughput quality factors on the output quality**

A stream of the literature has focused on a quality framework for the input, the throughput and the output phase (UNECE, 2014), the Eurostat accreditation in assessing the usability and usefulness of a source has shown the relationship between quality factors of the different phases. The formalisation of the impact of the input and throughput on the output quality is a key step to guide in the comparison of different data sources.

In the following, starting from the main factors of quality suggested in the relevant literature we outline these relationship for the European Statistical System quality dimensions. An overview is in Figure 1.

 *2.1. Relevance*

First of all, the relevance quality dimension is influenced by some contextual factors of the **Institutional environment**, that can be identified in the extent to which the **agreement** between the National Statistical Institute (NSI) and the big data holder is binding, and hence in the **sustainability** over time of the source. In addition, **legislation and privacy concerns** may limit the usability and potentially the relevance.

On the one hand, **availability, completeness and clarity of metadata** are a precondition, together with the above mentioned institutional factors, for the usability of the source, since they support the evaluation of the data sources when checking for data availability. On the other hand, they are vital in the “representation” and “measurement” lines (Groves *et al.*, 2004; Zhang, 2012) to define and build the statistical population and the statistical concept. Metadata and content, from the operational point of view followed by the Eurostat accreditation, are analysed in the first stages of the procedure to derive the expected usefulness in terms of relevance of the output (Eurostat, 2014).

The content **validity** (Groves *et al.*, 2004) of the measures in the input, i.e. how they relate to the statistical construct is the most significant statistical factor of the input phase influencing the relevance. However, the use of big data may represent an important contribution to the enhancement of the relevance, since they allow to produce statistical information for unexplored phenomena, i.e. the **newness** of the statistical output.

Cai and Zhu (2015) highlight that even if freshness[[2]](#footnote-2) of big data is very short, the collection, storage and analysis in real time can hardly keep pace with it. They state that **real-time processing** and analysis software for big data is still in development or improvement phases. As a consequence, also the relevance is potentially weakened.

*2.2. Accuracy and reliability*

The accuracy of the statistical output is the result of a large number of different components. Many of these components are common to the survey and administrative production processes and fall in analogous categories.

For traditional structured data, the AAPOR (2015) shows the errors due to the quality of the sources and the production process. The sources of errors can be classified in row, column or cell errors, where rows and columns represent units and variables, respectively. Row errors represent **undercoverage**, **overcoverage** or **duplication** errors; column errors are mainly **specification/validity** errors. Finally cells errors are due to **missingness** and **measurement errors**.

Coverage problems are generated by the coverage of the input source, that can represent a risk of **selectivity error** (non-representativeness) of the source. As an example failure to capture the signals/phenomena impacts on the coverage. Coverage problems can arise also due to the **complexity** of the data, as well as due to problems in the availability and quality of linkage variables used to identify the statistical units, that make difficult to **transform** them into statistical units.

It is worth noting that, linkage errors are even more relevant when the scope of the research is to look to relationships among units, typical in social network studies (Wang *et al*., 2012). These network studies have greatly benefitted from the growing availability of large, complex networks derived from big data. However, erroneous identifications of units (contacts) produce measurement errors on the relationships to be measured (Wang *et al.*, 2012).

When data are received by the NSI in a structured form, despite the gain in time and resources, the validity of methods and procedures applied by the data provider to transform the data, is out of the control of the statistical organisation, which can only rely on the available documentation.

**Measurement errors** in the input data are, similarly to the standard TSE approach, a function of the **validity** of the construct of the big data source measurement (corresponding to the questionnaire design) and the capturing tools (low signal/noise ratio in sensors). The latter can result also in missing measurements (lost signals in sensors). With respect to the throughput, similarly to the coverage, the statistical processes needed to obtain the statistical variables from big data as well as coding and editing produce measurement errors. **Complexity and diversity** of the big data increase the need for processing before attaining usable data for statistical purposes, including for example errors introduced by the algorithms, e.g. misinterpretation of sentiment from tweets (among the others, Hsieh *et al.*, 2017).

Finally, at the analysis stage, they depend on **reduction/sampling**, **matching** and **modelling errors** and on errors due to the **estimating strategy** (also for correcting selectivity).

When metadata are lacking, absent or erroneous the definition and identification of the statistical units and variables become more complex, and limit the source usability.

*2.3. Timeliness and punctuality*

A first important component on the output **timeliness** and **punctuality** is due to the timeliness and punctuality of the input source and to the **agreements/cooperation** with the data provider on the frequency of the delivery.

However, the **complexity** can require a long time for processing the data as well for integrating the big data in the statistical production system (e.g. if the **structure/nesting** of the records requires transformations to obtain statistical units and variables).

With respect to the throughput and also related to the abovementioned complexity of the input, the components of the **system performance** (Tao and Gao, 2016) – availability, response time, throughout, scalability - play a relevant role.

Big data are characterised by high **freshness**, however we can expect that even if big data are continuously updated, in practice, the NSIs have to set timeframe for their acquisition, thus decreasing their timeliness (see also section 2.1). Recall that timeliness in rapidly changing phenomena is also a key element of the relevance. In fact, the content changes so quickly that timeliness become more and more essential.

*2.4 Coherence and Comparability*

Coherence and comparability are enabled by the use of **standardised concepts** for the key variables of the input source or by the possibility to align the input classifications with those adopted by the NSIs.

Besides, the existence of **metadata** is a precondition for the usability of the big data and to assess the coherence and comparability.

Eurostat defines coherence statistics as their adequacy to be reliably combined in different ways and for various uses. Coherence is guaranteed with statistical techniques widely relying on models (e.g. seasonal adjustment, reconciliation, benchmarking) therefore during the processing the most relevant sources of errors identified are **model assumption errors** and **estimation errors** (e.g. model based estimators that do not guarantee hierarchical additivity/benchmarking).

Sampling and non sampling sources of errors (the latter including frame, selectivity, measurement and item non response and processing errors) have direct impact on accuracy, and then indirectly on coherence errors due to the lack of accuracy.

With respect to the comparability over time, the **sustainability** of the data provider, the changes through time of the source and of the legislation related to their use are the most relevant input quality factors.

When data are obtained from different providers for different regions, an impact on space comparability is awaited.

*2.5 Accessibility and Clarity*

Accessibility and clarity refer to the ease and modalities by which users can access, use and interpret statistics. It’s worth noting that the input quality dimension accessibility affects the big data usability. However, it does not influence directly the difficulty users may encounter in obtaining the statistical output. Here, we exclude that a Public Use File for research is produced from big data.

On the contrary, the lack of **clarity/transparency** of the methods applied for the big data source limits the NSIs capacity to support the statistical results with proper documentation for (expert) users.

**Figure 1. Input, process and output quality dimensions for the big data quality framework**



**3. Final remarks**

In this paper we aim at defining a structured quality framework for the use of big data, by identifying: the main sources of statistical errors, the process development and the contextual factors that have a relevant impact on the widely spread and adopted output quality components. We focused on the situation in which the statistics are produced using big data as unique input data source.

Our belief is that to anchor the big data quality framework, as much as possible, to the known sources of errors can bring benefits in other quality-related matter, at the planning, executing and evaluation phases.

First of all, to have a preliminary idea of the main sources of errors that will influence the statistical production, e.g. to be able to have a quality assurance plan, can orient the researchers towards better decisions in the planning stage of the production process.

It is known that one of the main characteristics of big data is their complexity and that this highly affects the performance and quality of the production process. This has to be considered when using big data for the statistical production and efforts should be addressed towards the increase of efficiency, stability and reliability of the systems used for the storage, treatment and analyses of the big data.

Finally, once it is ascertained that some kinds of error in the big data are equivalent in nature and impact to those already studied for the traditional production processes, similar approaches and methods can be attempted to try to estimate the potential bias and variance on the final statistical results. Of course, we are aware that the nature of the big data makes more difficult the adoption of traditional errors models currently used for estimating the undercoverage or the measurement error, and this represents a challenge for quality assessment.

The reinforcement of the quality framework for big data usage and the ability to document and increase the quality can contribute to increase the credibility of the big data source an hence users’ trustfulness on official statistics when produced by new data sources.

In order to achieve the above mentioned objectives, the outlined framework should be operationalised and equipped with a set of indicators (limited, relevant and informative), supporting the quality management in the use of big data for statistical production.

**4. References**

AAPOR (2015), Report on Big Data (No. 4eb9b798fd5b42a8b53a9249c7661dd8). Mathematica Policy Research.

Baker R. (2017), Big Data. A Survey Research Perspective. In Total Survey Error in Practice by editors P.P. Biemer, de Leeuw E., Eckman S., Edwards B., Kreuter F., Lyberg L.E., Tucker C.N., and Brady T. West, Wiley series in Survey Methodology, Hoboken, New Jersey.

Barcaroli G., Golini N., Righi P. (2018) Quality evaluation of experimental statistics produced by making use of Big Data, European Conference on Quality in Official Statistics, Q2018

Brancato, G. and Ascari G. (2018), Guidelines on the quality of multisource statistics. European Conference on Quality in Official Statistics Q2018

Cai L., and Zhu Y. (2015), The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp. 1-10.

Essnet Big Data (2018), Some Quality Aspects and Future Prospects for the Production of Official Statistics with Mobile Phone Data. Documents available at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5\_Reports,\_milestones\_and\_deliverables](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Reports%2C_milestones_and_deliverables) (accessed 21/05/2018)

Eurostat (2014) Accreditation of big data sources as input data for official statistics https://ec.europa.eu/eurostat/cros/system/files/D5\_Accreditation%20procedure%20for%20statistical%20data%20from%20non-official%20sources\_20140206\_0.pdf (accessed 20/05/2018)

Groves R. M., Fowler F.J.Jr, Couper M., Lepkowsky J.M., Singer E., Tourangeau R. (2004), Survey Methodology. Wiley, New York.

Hsieh Y.P. and Murphy J. (2017), Total Twitter Error. Decomposing Public Opinion Measurement on Twitter from a Total Survey Error Perspective. In Total Survey Error in Practice by editors P.P. Biemer, de Leeuw E., Eckman S., Edwards B., Kreuter F., Lyberg L.E., Tucker C.N., and Brady T.West, Wiley series in Survey Methodology, Hoboken, New Jersey.

Lugomer K., Soundararai B, Murcio R., Cheshire J., Lingley P. (2017). Understanding sources of measurement error in Wi-Fi sensor data in the Smart City http://huckg.is/gisruk2017/GISRUK\_2017\_paper\_95.pdf (accessed 20/05/2018)

Reis F., Di Consiglio L., Kovachev B., Wirthmann A., Skaliotis M. (2016) Comparative assessment of three quality frameworks for statistics derived from big data: the cases of Wikipedia page views and Automatic Identification Systems, European Conference on Quality in Official Statistics, Q2016

Struijs P. and Daas P. (2014), Quality Approaches to Big Data in Official Statistics. Q2014.

Tao C. and Gao J. (2016), Quality Assurance for Big Data Application – Issues, Challenges, and Needs. In Proc of the Twenty-Eighth International Conference on Software Engineering and Knowledge Engineering, Redwood City, San Francisco Bay, California, 2016.

UNECE (2014) A Suggested Framework for the Quality of Big Data available at

<https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2> (accessed 20/05/2018)

Wang D.J., Shi X., McFarland D.A., Leskovec J. (2012), Measurement error in network data: A re-classification. Social networks.

1. The views expressed in this paper are those of the authors and do not necessarily reflect the position of Istat. [↑](#footnote-ref-1)
2. Cai and Zhu (2015) call this “timeliness”, meaning real-time data collection at the source [↑](#footnote-ref-2)