United Nations
Global Working Group on Big Data for Official Statistics
Task Team on Cross-Cutting Issues

Deliverable 3: Further developing work on quality and methodological frameworks and analytical tools

Background document to the GWG meeting 19/10/2015
Draft Version 15/10/2015

1. Introduction

According to its Terms of Reference (TTCC-GWG, 2015), the broad objectives of the Task Team on Cross-cutting issues, classifications, frameworks and taxonomy (TTCC) are "to examine cross-cutting issues related to the integration of Big Data into official statistics' production, such as classification, taxonomy, data methodologies and quality frameworks for collecting, analysing and disseminating statistics derived from Big Data. In this context, the TTCC should develop and share knowledge of methodologies, data analytics and visualisation tools as well as quality assurance frameworks for the use of Big Data in official statistics and to refine the classification of Big Data."
The TOR defines four deliverables. The third deliverable concerns the quality and methodological framework when using Big Data sources for production of Official Statistics, analytics and visualisation tools.
The TTCC will produce a report on the application of the quality framework and on methodologies including appropriate analytical tools applied in various projects using Big Data for producing statistics. The report will most likely take the form of case studies (defining for each case study the estimation goal and methodological framework the quality framework for evaluating the input data and the output of the process) The use cases should allow to develop a generic and comprehensive approach in addressing quality and methodology in the use of Big Data for Official Statistics. The task will be achieved by means of short-term, medium-term and long-term objectives and related activities. In this phase the members of TTCC discussed, in particular, about the short-term objectives and actions. Section 2 introduces some

findings of the Global Strategies on quality and methodological issues. The results highlight that this deliverable can fill up some strategy gaps in adopting Big Data sources in the production process.
Section 3 is devoted to the objectives and activities by term of implementation. Sections 4 gives some conclusion.

2. Results of the Global Survey on quality and methods

The new 2015 edition of the UNSD Global Survey stressed, with many specific questions, the quality and methodological aspects of using Big Data sources in the Official Statistics. In the following, some main conclusions are presented.

1) Quality issues concern the NSOs when using Big Data sources, although Quality frameworks are not developed for most of the observed projects.

Figure 1 shows the results of the general part of the Global Survey, where we observe that the Quality framework is considered the second most urgent topic. Figure 2, extracted from the project part of the Global Survey, shows that more than 60% of the observed projects have quality concerns.

Figure 1

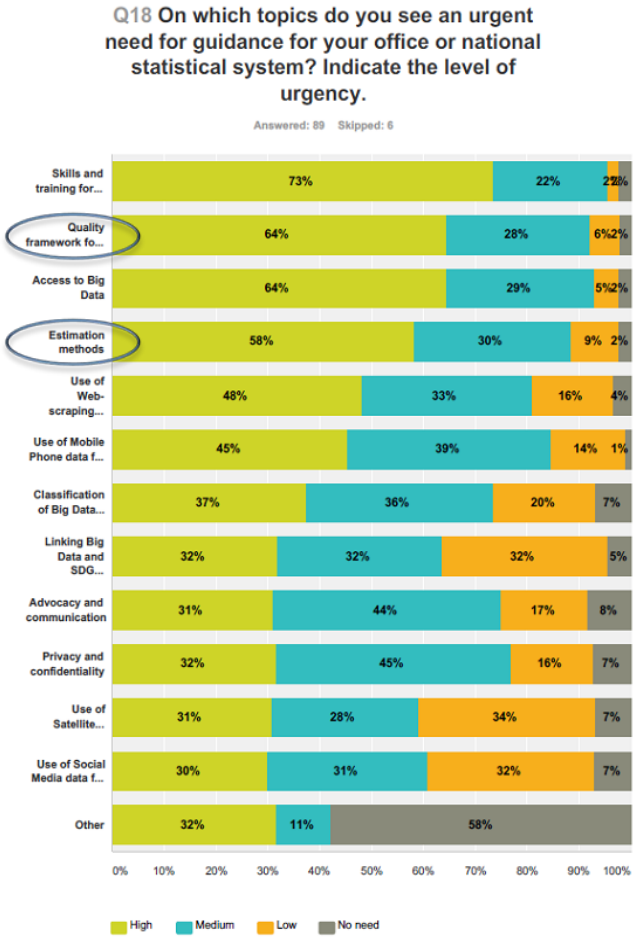
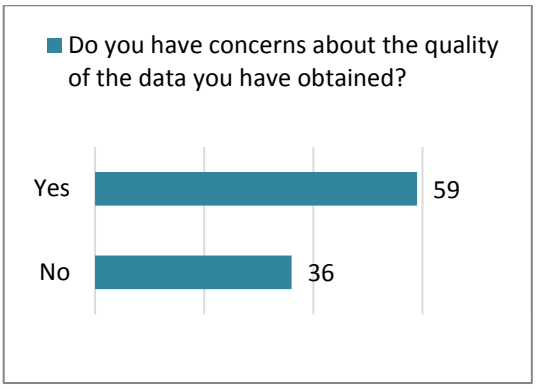


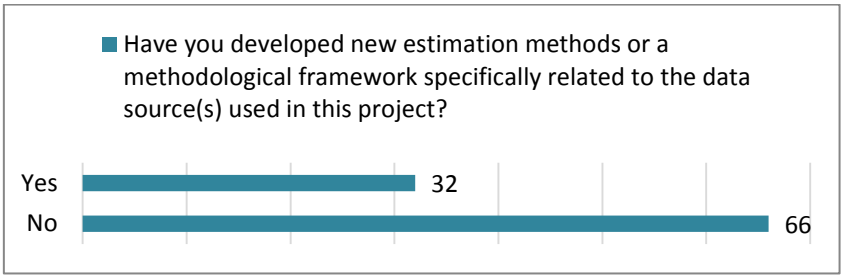
Figure 2



2) Methodological frameworks for Big Data sources are still an exception in the set of the observed projects although the NSO considers the issue quite relevant.

Figure 1, put the estimation method in the fourth place in the list of the topics with urgent need for guidance, and probably this weakness is reflected by the number of projects (less than 1/3) in which a new estimation method or a specific methodological framework has been developed (Figure 3).

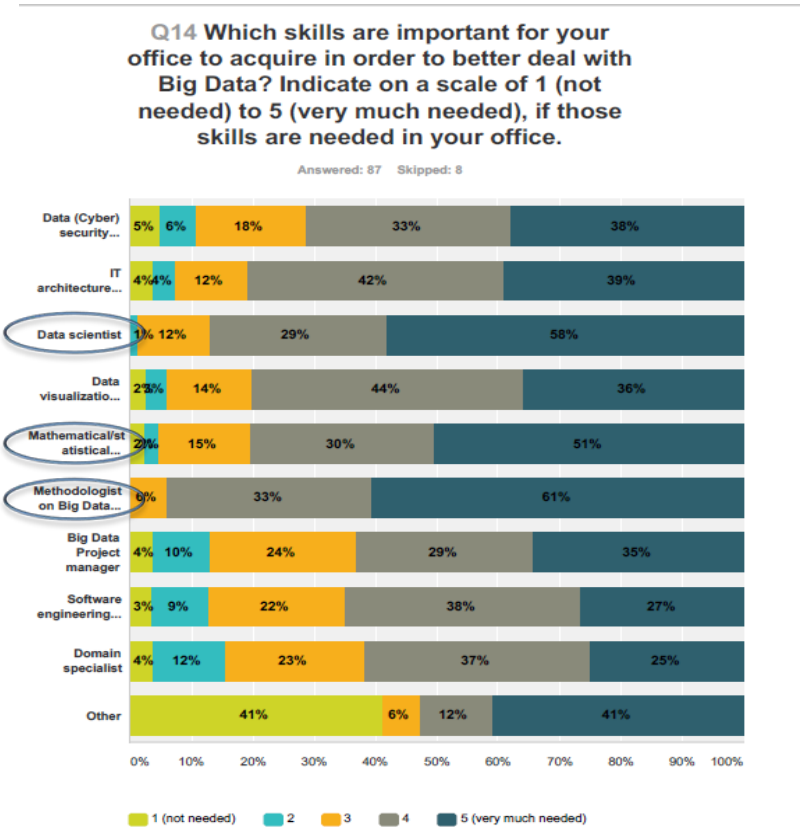
Figure 3



3) NSOs clearly identify professionality to deal with the quality issues and to set up a methodological framework.

The distribution of the skills important to acquire underlines that quality and methods are felt as relevant and urgent topics . Methodologist on Big Data together with Data Scientist is the most important figure to deal with the Big Data sources. Also mathematical and statistical modelling specialist competences are well considered by the NSOs (Figure 4).

Figure 4



3. Deliverable Further developing work on quality and methodological frameworks and analytical tools

The deliverable "Further developing work on quality and methodological frameworks and analytical tools" is explained by the TTCC as follows:

a) improving of developing current Quality frameworks on Big Data. This topic has been studied and formalised by previous projects and task forces. The TTCC starts from the proposal of the UNECE Task Team. The objective can be obtained with an overview of other proposals and collecting examples of applications of quality frameworks with real data for Official Statistics;
b) organizing and formalising a methodological framework for Big Data. This topic is less mature than the previous one and no main references are given in literature, although several proposals can be found. The objective can be obtained with an overview of the theory and collecting examples of applications with real data for Official Statistics. Note that in some applications the methodological framework is often implicit.
c) suggesting a proper use of analytical tools. The improvement or development of analytical tools are out of the scope of the deliverable.

3.1 Short-term objectives and actions

The short- term objective is to define the state of art, in a broad sense, of theory and practice of Quality and Methodological frameworks. The TTCC considers the two topics strictly related, being the methodological framework the tool for: assessing the quality of the produced statistics by using Big Data sources; improving the quality of the final statistics.
The TTCC has defined the following list of actions to complete in the first months of the 2016.

1. **Literature Review on existing methodological frameworks on Big Data.** A useful starting point is the Project Summary Report of the Sandbox Task Team of the UNECE Big Data Project (2015).
2. **Literature Review on existing quality frameworks on Big Data.** Since the deliverable of UNECE Big Data Quality Task Team is well-established, the action should investigate if there is room for improvements.
3. **Collect practical applications describing methodology, analytical tools, and quality measures**. This is a literature review to collect applications but also to understand: i) the level of development of methodology, quality and of using of the analytical tools; ii) the degree of satisfaction of the NSOs about quality when dealing with the Big Data sources. The review is mainly focused on applications and pilots in Official Statistics.
4. **Analysis of the results of the UNSD Global Survey.** Focus on the questions related to methodologies, analytical tools, and quality measures. The analysis completes the previous actions giving a large scale vision of these topics, even though less detailed.
5. **Draft a simplified methodological framework according to some dimensions of interest**. Examples of dimensions: context of use of Big Data source (replicating or reproducing existing statistics/ supporting/ producing new, etc.) approach to produce statistics, quality dimensions involved, etc. The framework could be enriched by useful information such as target parameter and domain of interest, analytical tools and quality indicators.
   5.1. **Possible improvement of UNECE Quality framework**.

Currently, the TTCC produced a draft version of literature review concerning the Quality and Methodological frameworks and examples of applications.

Regarding the Quality Framework two evidences appear:

1) there are quality dimensions not included (at least explicitly) in the UNECE proposal (ABS, 2010; OECD 2011; Couper, 2013; Brackstone, 1999). Such conclusion comes to light also considering the practical applications of Quality framework (see below);
2) quality in the Big Data context, more than in other contexts, should be evaluated strictly on business need, and in terms of objective costs–benefit criteria, i.e. by comparing the costs with the benefits of using the new data source, such as reduction in provider load, sustainability of new source and the traditional quality dimensions (Tam and Clarke, 2015a).

Regarding the Methodological Framework, attempts to define a general framework have been proposed (Tam and Clarke, 2015b; Lavallée, 2015; Cox, 2015). Nonetheless, the experiences observed in the literature review do not show a uniform approach to the methodology.

The review collecting examples is mainly organised to identify the following aspects:

a) quality dimensions considered in the application;
b) quality indicators used to measure the quality dimensions;
c) effects of the lack of quality on the final statistics;
d) methodological frameworks (or part of) used to reduce the effect of lack of quality.

An example of the output for each application is given in figure 5.

Finally the findings of the Global Survey have been analysed as well. Some conclusions are shown in section 2. Nevertheless, other valuable information emerges from the survey.

Figure 5

Satellite imagery – ABS Australia

Tam et. al (2015). Small steps towards Big Data - Some initiatives by the Australian Bureau of Statistics. International Statistical Review.
Tam (2015). A STATISTICAL FRAMEWORK FOR ANALYSING BIG DATA. The Survey Statistician.
Tam & Clarke (2015) Big Data, Statistical Inference and Official Statistics. ABS Research papers.

Remote sensing for Agricultural Statistics – investigate the use of satellite sensor data for the production of agricultural statistics such as land use, crop type and crop yield. [Paper introduces an exhaustive methodological framework dealing with, sampling design, selectivity (under-coverage) and non-response of Big Data Source and setting up the super-population model for making valid inference].

Quality issues

Quality dimension	Quality indicator or qualitative evaluation	Threats	Methodology for dealing with lack of quality
Selectivity: the coverage of satellite data is the same as the coverage of land parcels	Coverage	-	In case of under-coverage to make valid inference ignorability condition (missing at random conditions) must hold
Linkability: of pixels observed by ground trothing and by satellite	Quality of linking variables	[With bad quality linkage the subsequent model relating crop yields with satellite data produces biased estimates]	[Probabilistic record linkage models]
Missing Data: missing data is due to a bad weather (persistent cloud cover)	Rate of missing values	Biased estimates	Testing if the non-response mechanism is missing at random. Ignorability condition. [Nevertheless] persistent cloud cover, as a result of moisture in the atmosphere, which may affect the type of crops being grown, or yields [Non-ignorable non-response] This issue may, however, be bypassed by using traditional data collections e.g. statistical surveys, instead of using satellite data, for these areas.

Parameters: Crop yields by type
Objective: Support traditional survey. BD as auxiliary variables for model based inference;
Analytical tools: parametric generalized linear models, machine learning tools, supervised classification

Methodological framework to make inference:
Ignorability conditions (selectivity and non-response) need to be checked (missing data) and satisfied. The satellite data may be treated as a random sampling, and model assisted or model based inference can be implemented after a supervised classification of the satellite imageries.

3.2 Medium-term objectives and actions

To obtain the final deliverable, the successive step should enhance the short-term objective.

The aim is to give a more precise overview of the cross-cutting issues on quality and methodology and analytical tools.

The principal action in the medium-term (2016) is the collaboration with the Task Teams on Mobile Phone, Satellite Imagery, Social Media, and SDG. The task teams should share their experiences, evidences and conclusions on these three topics.

The task team on cross-cutting issues will integrate these experiences with the short-term evidences to outline the cross-cutting issues in the medium-term.

The form of cooperation and the type of outputs must be planned among the task teams. This step is quite important and challenging for obtaining an organic output in the GWG.

3.3 Long-term objectives and actions

The overarching aim of developing a generic and comprehensive approach in addressing quality and methodology in the use of Big Data for Official Statistics will be completed in the long-term period.

At this stage the final output and the actions are necessarily not well defined.

The TTCC will produce a first report on the applications of the quality framework and on methodologies including appropriate analytical tools applied in various projects using Big Data for producing statistics.

The report will most likely take the form of case studies (defining for each case study the estimation goal and methodological framework the quality framework for evaluating the input data and the output of the process).

A second report should illustrate the above mentioned overarching aim by integrating the evidences of the first report, theory on quality and methods found in the literature and the input coming from the other task teams devoted to the Big Data sources.

4. Final Discussion

The TTCC with the deliverable “Further developing work on quality and methodological frameworks and analytical tools” aims to tackle three issues of different levels of maturity. Quality framework on Big Data is currently well founded and the proposals in the literature converge. On the other hand, the Methodological frameworks on Big Data are less uniform and some proposal has a weak formalisation. Finally, the analytical tools are well developed and the output of the deliverable seems to be how to use them in efficient way. The overall deliverable is planned to summarize the convergence of a bottom-up (examples) and top-down (theory) approach. The potential output should test and strengthen the work already done on the Quality framework and organise a Methodological framework according to each specific Big Data source. The latter output seems to be innovative in the Big Data field and for such a reason further feedback based on practical applications will contribute to improve the deliverable with further versions.

References

ABS (2010). The ABS Data Quality Framework. Available at: https://www.nss.gov.au/dataquality/aboutqualityframework.jsp. Accessed November 2014.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25, pp. 139–149.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7, pp. 145-156.

Cox D.R. (2015). Big data and precision. *Biometrika,* 102, pp. 712–716

Lavallée P. (2015). Sample Matching: Toward a probabilistic approach for Web surveys and Big Data?. *Plenary Session*. ITACOSM 2015 4th ITAlian Conference on Survey Methodology Rome, June 24-26, 2015.

OECD. (2011). Quality dimensions, core values for OCED statistics and procedures for planning and evaluating statistical activities. Available at: http://www.oecd.org/std/21687665.pdf. Accessed November 2014.

Tam, S.M. and Clarke, F. (2015a). Big Data, Official Statistics and Some Initiatives of the Australian Bureau of Statistics, *International Statistical Review* (to appear).

Tam, S.M. and Clarke, F. (2015b). Big Data, Statistical Inference and Official Statistics. *Research Papers* – ABS, Catalogue no. 1351.0.55.054, March 2015.

TTCC – GWG (2015). Terms of Reference of the Task Team on Cross-cutting issues, classifications, frameworks and taxonomy, Revision 8, 3 October 2015.

UNECE (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team.

UNECE (2015) Sandbox Task Teamongoing project. Available at http://www1.unece.org/stat/platform/display/BDP/Sandbox+Task+Team. Accessed October 2015.