Big Data Quality framework: experiences on real data in Official Statistics

Paolo Righi

Italian National Statistical Institute

Istat
Istituto Nazionale di Statistica

Global Conference on Big Data for Official Statistics
Organized by NBS/UAE, UNSD, ABS and GCC-STAT
20-22 October 2015
Abu Dhabi, UAE

Outline

- **Background**
- **Quality framework**
- **Methodological framework for quality**
- **Quality framework in practice: examples**
- **Global survey: focus on quality and methods**
- **Conclusion**

Background

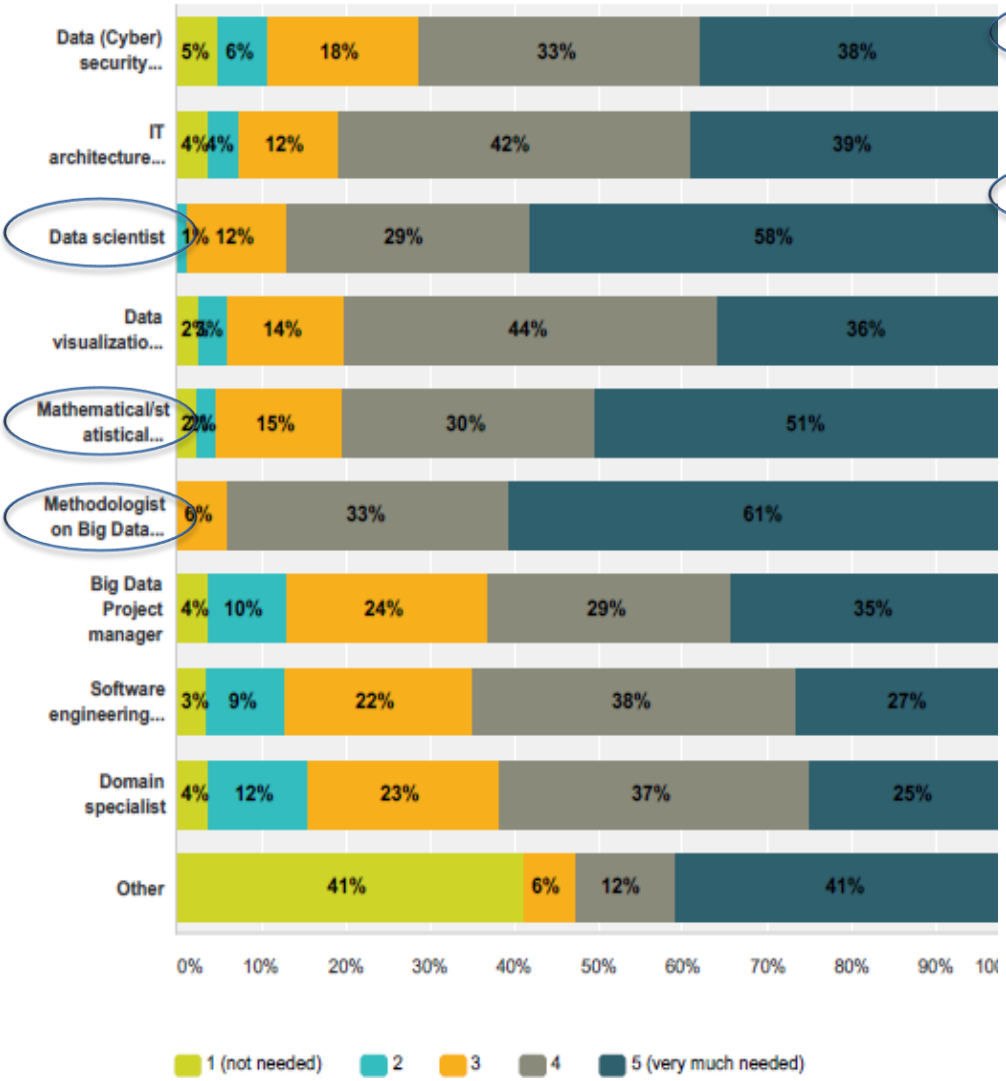
The claim that large amount of information coming from a Big Data Source produces more reliable statistics than survey data is not enough in itself

With very large amounts of data, direct use of standard statistical methods, including simulation-based approaches, will tend to produce estimates of apparently very high precision, essentially because of strong explicit or implicit assumptions of at most weak dependence underlying such methods (Cox, 2015)

In fact quality of the statistics *is very much dependent on how representative a particular Big Data source is of the underlying population, and on the nature of the statistical inference drawn from such data* (Tam & Clarke, 2015a)

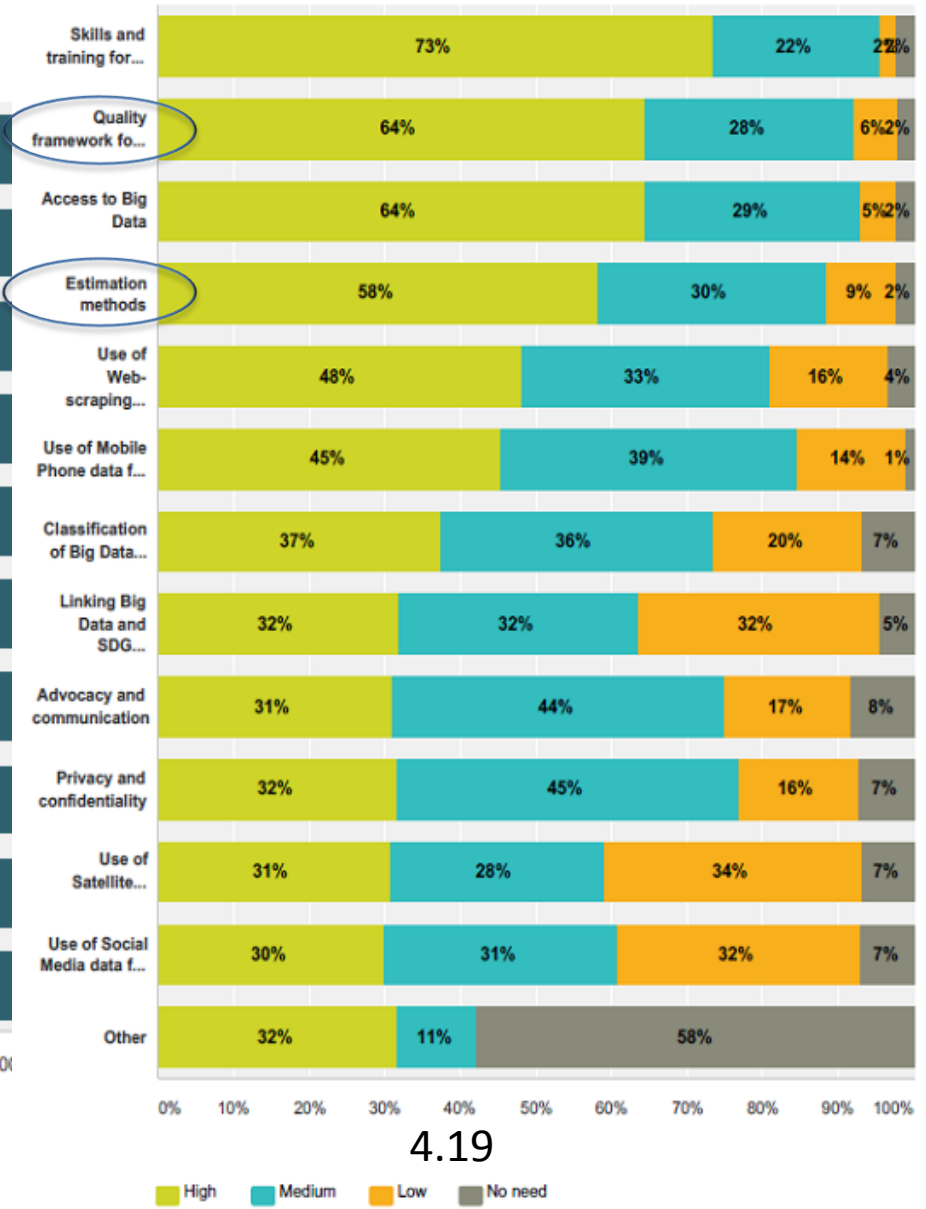
Q14 Which skills are important for your office to acquire in order to better deal with Big Data? Indicate on a scale of 1 (not needed) to 5 (very much needed), if those skills are needed in your office.

Answered: 87 Skipped: 8



Q18 On which topics do you see an urgent need for guidance for your office or national statistical system? Indicate the level of urgency.

Answered: 89 Skipped: 6



Quality framework

Consider the proposal of the UNECE Big Data Quality Task Team (2014)

Quality investigated by the phase of business process

- **Input**: acquisition, or pre-acquisition analysis of the data;
- **Throughput**: transformation, manipulation and analysis of the data;
- **Output**: the reporting of quality with statistical outputs derived from big data sources.

Quality framework: Dimensions

- Institutional/ Business Environment
- Privacy & Security
- Complexity
- Completeness

(UNECE, 2014)

- Usability
- Time factor (Timeliness, frequency)
- Coherence: Linkability; Consistency

- Validity
- Accessability and clarity
- Relevance
- Accuracy:
  Input: Selectivity/representativeness
  Output: bias+ variance (Systematic and random error)

Other Quality dimensions (found in literature): interpretability, cost and sustainability (ABS 2010; Brackstone 1999; OECD, 2011), comparability (over time and space, Eurostat, 2009) volatility and stability (Couper, 2013)

Methodological framework for quality

Lack of quality in input or throughput can be dealt with suitable methodological framework for improving the quality of the output statistics

Methodological framework basically makes assumptions on the mechanism generating the data:

- Selectivity (of BD source);
- Missing Data (and outliers);
- Sampling of BD;
- Relationship between target variable and BD data (superpopulation model).

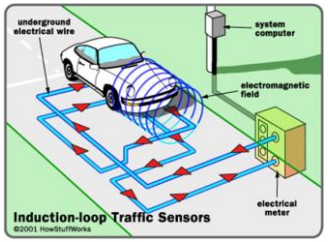
Exploiting auxiliary information (source: survey data, census, administrative data, etc.)

Identification of a model where Ignorability conditions or MAR mechanism hold is crucial to make valid inference

Tam and Clarke (2015b), Lavallée (2015) give model-assisted /model based (bayesian, frequentist) approach to the inference

Quality framework in practice: Selectivity

Traffic loop detection data (Daas et al. 2015): consist of measurements of traffic intensity. Each **loop counts the number of vehicles** per minute that pass at that location, and measures speed and vehicle length one.



The set of locations in the highways with traffic loop is not a random sample

Methodological Framework:
predictive modeling approach (observed data + predicted data by estimated model) – MAR fulfilled

Selectivity: a subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective (Buelens et al., 2014)

Quality framework in practice: Selectivity

GPS data (Pratesi et. al., 2015): **inhabitants' mobility using data of private vehicles tracked with a GPS device.** The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points. Vehicle traces were then mapped on the road network.

Mobility collected on a self-selected sample of car journeys

Methodological Framework:

- reconcile data from the two independent sources (Big Data and sample surveys) ;
- if no common variables are available, known correlated data could be used;
- use Big Data directly as a covariate under a model-based approach.

Quality framework in practice: Missing data

Remote sensing for Agricultural Statistics (Tam and Clarke, 2015b): investigate the use of satellite sensor data for the production of agricultural statistics such as land use, crop type and crop yield. **The coverage of satellite data is the same as the coverage of land parcels**.

Missing data due to persistent cloud cover.

Methodological Framework:

- Set up a model where ignorability conditions hold: Calibration or imputation techniques;
- Model with Ignorability conditions not found for certain areas (Weather may affect the type of crops being grown, or yields, missing data related to the target variable).

→

Use traditional data collections for these areas

Quality framework in pract.: Privacy and linkability

GPS Data (Pratesi et. al., 2015): the final aim is to study the possible agreement between the **level of poverty and the diversity of its inhabitants' mobility in the (small) areas** under study using BD as a covariate (EU-SILC survey collects the target variables- income)

Due to technical problems and **legal restrictions, it is unfeasible to have unit-level data that can be linked with administrative archives**, census or survey data

Methodological framework:

- Use of Fay-Herriot area-level model instead of unit-level model;
- The modified version proposed by Ybarra and Lohr (2008) allows for measurement error in the auxiliary variables(due to self-selection)

Privacy, use of the data given legal limitations, organisational restrictions, and confidentiality and privacy concerns.

Linkability, is the ease with which the data can be linked or merged with other relevant datasets.

Quality framework in practice: Validity

Use of Mobile Positioning Data for Tourism Statistics (Eurostat et al., 2014): The feasibility Study investigates opportunities of using mobile phone data CDR (*subscriber_id*; *country_code*; *event_time*; *cell_id)* for tourism

Tourism Definition: the activity of visitors who are taking a trip to a main destination which is outside the usual environment, which lasts less than a year, and which is for any main purpose, including business, leisure or other personal purpose, other than being employed by a resident entity at the location that has been visited (Eurostat 2013).

Mobile positioning data does not provide information on the purpose of the visit neither does it allow work-related visits to be identified in which the employer resides in the country visited.

Meth. Framework: assessment that can be carried out relies on the correlations with other measures of the same concept that are measured at the same time (traditional survey or administrative data).

Validity: How much the theoretical definition matches empirical evidence (quantitative measurements and calculations based on the CDR). Invalid measurement leads to systematic error or bias in estimates (McCall, 2001).

Quality framework in practice: Usability

Internet as Data source (Barcaroli et. al., 2015): Aim: **produce information on the use of Internet** and other networks by the enterprises for various purposes (e-commerce, e-skills, e-business, social media, e-government, etc.) currently estimated by the ICT survey, **using web scraping techniques**, associated, to text and data mining algorithms.

Once **completed the web scraping** activities the set of words related to the target variable is too numerous to be managed in the modelling phase (**document/term matrix too big**) – Spurious correlation

Methodological Framework: selecting only the words that showed a significant influence on the target variables by implementing a:

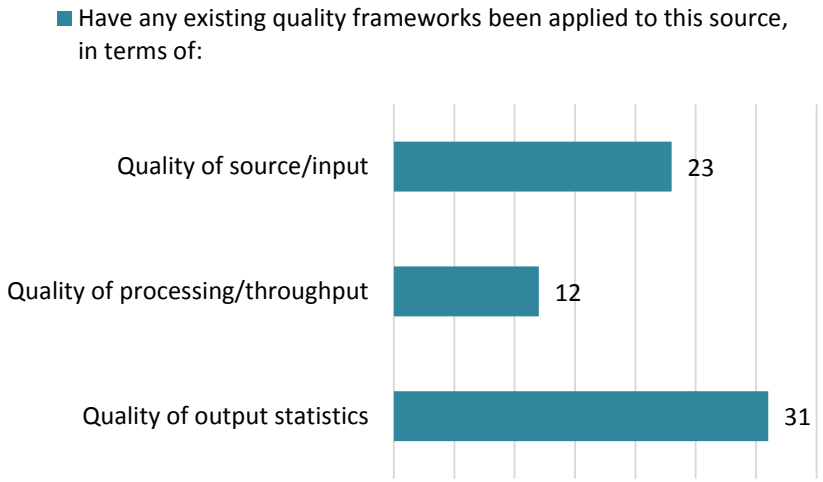
- correspondence analysis between a given target variable and the words contained in the scraped texts ;
- Chi Square analysis : target variable by presence/absence word.

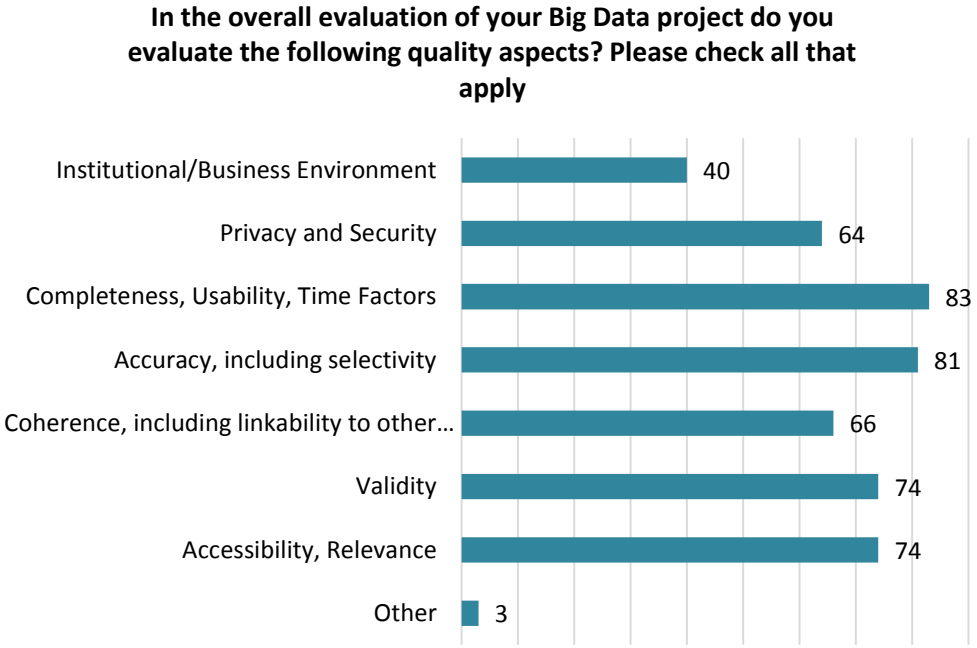
Usability: is the extent to which the NSO will be able to work with and use the data without the employment of specialised resources or place significant burden on existing resources.

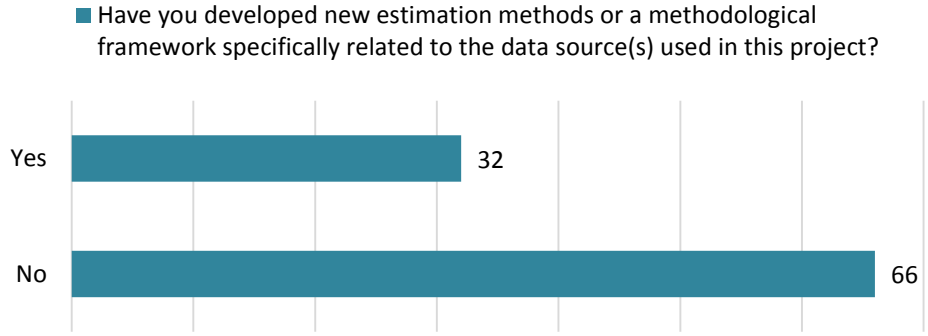
Global survey: focus on quality and methods

a) 115 Projects using Big Data.

b) Respondents stressed the importance of quality and methodological issues

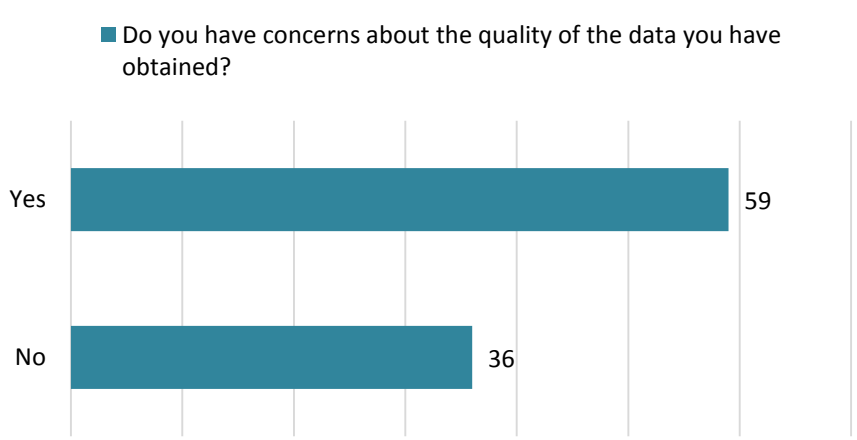


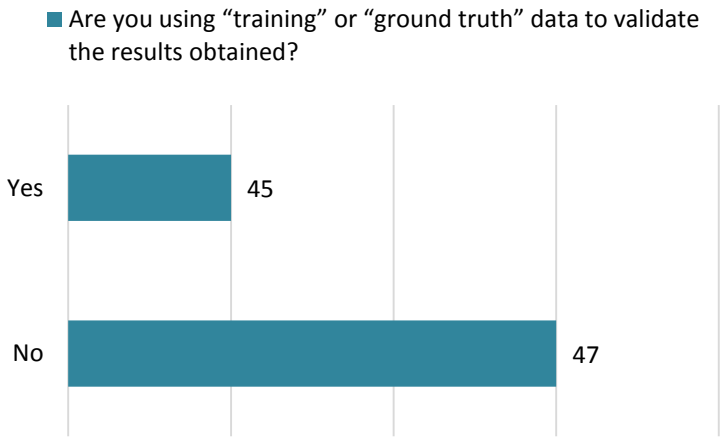




Global survey: focus on quality and methods

Quality and ground truth data





Conditional analysis (data: respondents both questions – 85)

Prob (concerns=yes| ground truth data=no) = 50%

Prob (concerns=yes| ground truth data=yes)= 66%

Conclusion

NSOs have started to focus on Quality of Big Data sources relatively recently

A) The responses to the general questions on the Global Survey highlight that **quality** and **methodology have a prominent position in the list of concerns when using Big Data sources**

Conclusion

B) **Use of Traditional data** (Survey, Census, Administrative data) seems to be **fundamental for setting up the Quality and Methodological framework** (in theory and practice)

Scope:

- Evaluating the quality of Big Data;
- Improving the quality of the official statistics (Blended data approach)

C) The examples show **not unique Methodological framework** for all the Big Data sources

D) Challenges and opportunities that offer the Big Data lead to give a **different vision of the quality issues** (cost-benefit)

Conclusion

Quality is not an absolute**. It must be **evaluated relative to the stated aims** of the survey and the purpose to which is put, **and the investment (time and money) in obtaining the data (Couper, 2013).

***Statistical agencies need**, above all, **sources of data** that cover a known population **with error properties that are reasonably well understood** and that are not likely to change under their feet* (Citro, 2014).

*One important caveat […], is the need to **create transparent methodological documentation (metadata) that describes the ways in which Big Data are used in** the construction of any kind of **estimate*** (Horrigan, 2013).

References

- ABS (2010). The ABS Data Quality Framework. Available at: https://www.nss.gov.au/ dataquality/aboutqualityframework.jsp. Accessed November 2014.
- Barcaroli G. et al. (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. *Austrian Journal of Statistics*, 44, pp. 31-43.
- Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25, pp. 139–149.
- Buelens, B.et al. (2014). Selectivity of Big Data. *Discussion paper*, 2014/11, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, pp. 137-161.
- 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
- Daas, P.J.H., Puts, M.J., Buelens B., van den Hurk, P.A.M.(2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, 31, pp. 249–262.
- Eurostat et al. (2014). Feasibility of Use: Methodological Issues. Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics Report 3a.
- Eurostat (2013). Methodological manual for tourism statistics, Version 2.1. Cat. No: KS-GQ-13-007-EN-N.
- Horrigan, M.W. (2013). Big Data: A Perspective From the BLS. Amstat News January, pp. 25–27. Available at: http://magazine.amstat.org/blog/2013/01/01/sci-policyjan2013/ (accessed July 2015).
- 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.
- McCall, R. B. (2001). *Fundamental Statistics for Behavioral Sciences*, 8th edn, Wadsworth, Belmont.
- 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.
- Pratesi M. et al. (2015). Small Area Model-Based Estimators Using Big Data Sources. *Journal of Official Statistics*, 31, pp. 263–281.
- 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 Paper* – ABS, Catalogue no. 1351.0.55.054, March 2015.
- UNECE (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team.
- Ybarra, L.M.R., Lohr, S.L. (2008). Small Area Estimation When Auxiliary Information is Measured With Error. *Biometrika* , 95, pp. 919–931.