Big Data: from noise to evidence

Patricia O'Hara[1] and Pieter Everaers[2],

[1] *National Statistics Board, Ireland;poharaca@gmail.com*
[2] *European Commission, Luxembourg; Pieter.Everaers@ec.europa.eu*

Abstract: Big Data are becoming more prominent on the agendas of statisticians. The availability of a large variety of Big Data information is expected to substantially influence the process of producing statistics and generating statistical information. Research methodology is heavily affected and requires a rethinking of the basic concept of validity. The traditional survey based approach involving operationalisation from theory and hypothesis, via theoretical concepts, to measurement and data collection will be replaced by an approach where more importance is given to an inductive fact and relations-finding approach. The availability of manifold data, like in Big Data, that are not collected based on a conceptual model that has been developed from a theoretical model for a specific research question, but purely as the by-product of a digital development, also creates a different data environment for the concept of validity. Moreover, this raises governance issues, as the dependencies from data providers and statistical decision making process will be affected. All this will necessarily lead to a different production and dissemination process for statistics and therefore will affect the quality framework for official statistics. Specifically, at the governance level, the European Statistics Code of Practice will be affected by this change in the nature and process of translating policy and research questions into the production and dissemination of statistics. As there will be additional stakeholders in this Big Data world, the mandate and tasks of ESGAB will also be affected.

Keywords: European Statistics, ESGAB, Quality, Measurement, Validity, Big Data,

1. Introduction

The digital revolution has caused a large volume of data to become available for statistics and research. Narratives like Data Revolution and Big Data (Kitchin, 2014) describe this emergence of new information sources. These new data sources are not developed for statistical or research purposes, but represent a diverse range of public and privately generated fine-scale data about citizens and places in real time which are the 'noise' or by-product of the

production of consumer services and products. Examples of big data generators include financial institutions and retail chains, emergency services, transport providers, travel and accommodation websites, social media websites and mobile phone operators.

While not generated for statistical purposes, the increased availability and use by statisticians and researchers of data about persons, households and business from these new sources raises important issues for the production of official statistics as it requires changes in the way such statistics are produced. The changes needed may relate to measurement techniques for collecting the data, analysing and transferring them into statistics and information that describe societal phenomena, but also relate to the methodologies to be applied. With these new data sources, the traditional research process of selecting certain measurement techniques and defining variables to be measured, will reverse into a process of reasoning about why to use certain existing variables and justifying their suitability for measuring societal phenomena. The meaning of the concept of *validity*[1] of a variable measuring correctly a concept, changes significantly in this context.

Currently many high quality official statistics are based on well-established data sources, statistical methodologies and processing techniques, that have proven to deliver results of a sufficiently high quality, and to be fit for the purpose in delivering the 'evidence' for research and policy purposes. The quality of official statistics based on the new data sources is, as yet, largely unproven to be of the same high level. The challenge of using the new data sources – to transfer this data from 'noise' into 'evidence' that fits for the purpose of decision making – is to ensure the high quality of statistical information based on such data.

Quality in official statistics has several dimensions. In this paper the impact of the new data sources - mainly those that can be described as Big Data - on these quality dimensions is

[1] Data shall be obtained of such a kind and in such a way that legitimate inferences can be made from the manifest to the latent (Galtung, 1967, p29).

illustrated by looking especially at the process of operationalising[2] the theoretical concepts to be analysed into measurable variables and in the general characteristics of the data sources such as ownership, confidentiality issues, and the provenance and lineage of datasets, as far as they impact on, or have a perceived impact on, quality.

These new data sources – as a by-product – as they are not collected via the well-known data collection machinery of statistical organisations, are owned and controlled by a range of 'stakeholders' both public and private. Thus, their use for official statistics purposes requires the creation of new forms of cooperation and relationships with the data providers. The traditional institutional quality requirements from, for example, the European Statistics Code of Practice (ES-CoP)[3] would be difficult to directly apply to these organisations. This implies that quality control will have to be exercised and guaranteed indirectly via, for example, inclusion in the Code of Practice of principles for handling this type of information to be used by national statistical organisations. Thus, the ES-CoP will be affected by this change in nature of the process of translating policy and research questions into the production and dissemination of official statistics. Moreover, the responsibilities of the European Statistics Governance Advisory Board (ESGAB)[4] using the ES-CoP as their main reference document for describing compliance to the highest statistical governance standard, will have to be adapted to allow for the inclusion in statistical production of these sources and the characteristics of the stakeholders involved.

---

[2] The process in research of the translation of an image of reality into an empirical measurable characteristic (Korteweg en Van Weesep, 1983, p73).

[3] The European Statistics Code of Practice was established in 2005 by the Statistical Programme Committee. It is presented in the Recommendation of the Commission of 25 May 2005 [COM(2005) 0217 final – Not published in the Official Journal]. The Code defines standards on the independence, integrity and accountability of the national and Community statistical authorities. It therefore contributes to the improvement of good governance, the quality of statistical data and user confidence in the authorities concerned. http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15.

[4] See paragraph 3

2. The quality of official statistics

Quality in official statistics is multi-dimensional. It includes the dimension of basic mathematical and methodological issues; the scientific and social rationale for choice of concepts for describing phenomena; the institutional dimension, that is the environment in which the statistics are produced and disseminated; and, last but not least, the place of official statistics in the context of the norms and values of the society in which they are produced, disseminated and used in decision making, monitoring and evaluating. In simple terms these elements can thought of as levels of increasing quality of official statistics *(see Figure 1)*. Each level requires a specific assessment by a different stakeholder. Highest quality official statistics naturally have to score high on all these elements.

Level 1: Mathematical quality: this relates to the correct choice and application of certain measures and using (descriptive and inductive) statistics to apply numerical values to phenomena. Based on the measurement level of the available information, a researcher has to make choices between available statistical (in a narrow sense) techniques. Professional statisticians are trained in this competence.

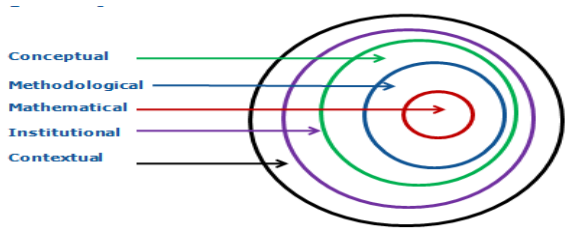
Level 2: Methodological quality: this relates to the use of the correct (widely accepted and legitimate) methods to collect and analyse the data and includes issues like the definition of the research population, the measurement units, and the use of data collection techniques; in short, the combination of type of research strategy and data sources. These choices and application of methodology are part of the training of researchers in social sciences and economics.

Level 3: Conceptual quality refers to the use of the correct 'theoretical model'[5]; correct from the point of view of the society's common understanding (and the current paradigm in science), its norms and values. Such a theoretical model describes the evolution and/or the

[5] The description in theory of the relationship between characteristics (Galtung, 1967).

functioning, via assumed (causal) relations between the 'theoretical concepts', of a specific issue in society. For example the concept of 'poverty' (a multidimensional theoretical construct) can be described via the combined effect of several theoretical concepts income, access to services, etc. The operationalisation of these theoretical concepts via indicators (e.g. household income in the case of poverty) into measurable variables (net yearly disposable household income in Euros) is a crucial part of the research model. Different theories lead to different theoretical models, use of different concepts and different choices of variables. The acceptance in a certain societal environment of this translation of the relationship between theoretical concepts into measurable variables is often considered as an element of the quality of statistics. By using the results, science and society (policymakers) are the main assessors of this quality element.

Figure 1: Quality levels



Level 4: Institutional quality in official European statistics is reflected by the level of compliance with the principles of the E-S CoP of the National Statistical Institute (NSI) and the Other National Authorities (ONA) involved in providing information to official European statistics as disseminated by Eurostat. The institutional quality in European statistics in the European Statistical system is assessed via a system of regular peer reviews (Karkkainen and Del Bario, 2016) and improved/maintained via agreed improvement actions. The principles of the code of practice cover the correct application and the responsibilities of quality on levels 1

to 3. The European Statistics Governance Advisory Board (ESGAB) plays an important role in monitoring this quality of European Statistics (see paragraph 3).

Level 5: Contextual quality describes the level of formal acceptance by the highest authority, be it the national government or prime minster, or in the case of Eurostat, the College of Commissioners, of the institutional quality: the correct implementation of the Code of Practice in producing and disseminating European Statistics. ESGAB also assesses the state of affairs on this quality level.

3. The role of the European Statistics Governance Advisory Board [6]

ESGAB was established by the European Parliament and the Council in 2008 to provide an independent overview of the ESS with particular regard to implementing ES CoP insofar as it relates to Eurostat, and to assessing the implementation of the Code in the European Statistical System as a whole. ESGAB's aim is to enhance professional independence, integrity and accountability – key elements of the Code of Practice – in the European Statistical System, as well as the quality of European statistics. It comprises six members and a chairperson.

ESGAB prepares an annual report to the European Parliament and the Council. It can also advise Eurostat on appropriate measures to facilitate implementation of the Code; on how to communicate the Code to users and data providers; on updating the Code; and on questions related to user confidence in European statistics, if considered necessary[7].

Based on this mandate, the focus of ESGAB is in general on the institutional and contextual level of quality. However, in specific cases, ESGAB might want to relate its comments to the other levels, as in the situation when ESGAB assesses Eurostat's compliance to the ES-CoP on applying the correct methodology or using the most current theories. The emergence and use

[6] See also Annex 1.

[7] See www.ec.europa.eu/esgab

of Big Data sources is a development that will substantially change the process of producing and disseminating official statistics and involve new stakeholders and relationships. In assessing the functioning of the ESS and Eurostat therefore, ESGAB will have to adjust its assessment to include these new situations.

4. The traditional process of transforming data into statistics

Data on persons, households and businesses are fundamental to social and economic statistics and research. In traditional (survey) statistics the collection of data is based on a well established process of sample-based surveys and questionnaires. For decades in many European countries these data have been complemented by information from (governmental) administrative sources and registers. In the sample-based statistics process an important step in measurement is the translation of a research question into items (represented in traditional questionnaires, or in electronic queries) that allow allocation of a numerical value to phenomena. The quality of this translation process, operationalising the research question, via a conceptual model that describes the assumed relations between a set of assumed relevant theoretical constructs via concepts into variables and the range of possible scores on this variables, is considered to reflect the validity of the measurement. The concept of validity captures the commonly shared understanding in a society (based on accepted axioms) that what is measured is that which is indeed intended to be measured. Validity contributes highly to the methodological and conceptual quality of official statistics (level 2 and 3, see paragraph 2).

The process flow of traditional research, based on a researcher collecting his/her data, starts with reflections on the correct theoretical model, followed by the selection of concepts (based on axioms and theories) and operationalised into variables that can be measured (categories and values). This leads, via the data collection phase, to a set of data that, based on using statistical techniques and measures, is translated into results that are assumed to reflect the concepts in reality. In empirical research this implies that the assumption on relations between concepts that are proposed at the start of the research are proven or disproven (the hypotheses to be rejected or not).

With administrative sources and registers (often containing relatively basic information on persons, households and business and kept by ministries or other governmental organisations with a public function) the initiative for the values and categories has shifted to the keepers of those sources. The amended regulation on European statistics[8] and consequently also the national practices facilitate an early involvement of official statistics in the development and review of these sources and consequently a better use of these data for statistics.

5. The translation of noise into valid statistics

The new data sources are distinctly different and outside the traditional processes of turning data into statistics; whether produced in the public or private sphere they are the (unplanned) by-product of an activity or process. Their production, while perhaps subject to in-house quality procedures of the producing institution, has not been required to comply with a code of good practice (like the ES-CoP) that could guarantee the quality of the data for use in official statistics. The data are available, often without an ex ante reflection on what (concept or theory) they could represent. Moreover these data may be collected without any methodological plan. Finally the mathematical characteristics of these data are often ex ante not defined. However, these sources are increasingly being regarded by researchers and statisticians as potentially valuable data that can be used for research and statistical purposes. The use of these data can not only be seen as an option, but moreover as an opportunity to increase the use of statistics in all types of decision-making. Their use also brings the potential benefit of reducing the administrative burden on society from specific surveys.

Earlier in this paper the multi-dimensional quality of official statistics has been described as having various quality levels that all together contribute to the overall quality. To achieve trust in the quality of the results, based on new 'big data' sources, researchers and statisticians will be obliged to demonstrate how the data and analysis fits into these quality assessments. This will require the correct application of descriptive statistics (level 1); demonstrating how the

[8] *ttp://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Ami0058*

distributional characteristics allow generalisation of the results to a certain research population (level 2); an analysis of how they reflect theoretical concepts (level 3), and showing to what extent the information produced and disseminated based is compliant with the principles of the Code of Practice (level 4).

Compliance with the first two levels of quality has to be argued from a mathematical point of view. This paper does not deal with these issues, and this is not so much a concern for ESGAB. The changes in assessing the validity, however, will influence ESGABs work and assessment. The choice of concepts and how they are validated – the conceptual quality - is based on norms and values and accepted mainstream theories of the society concerned. The adequacy of the measurement of societal issues for the purpose of policy-making needs a reflection on the appropriateness and correctness of the choice of certain concepts and conceptual models. So the assessment of quality needs to include the ex-post validation of the data collection to the existing theories.

Empirical research will be required to prove the correlation between distributions based on existing statistics with those stemming from new data sources. A high correlation justifies the amendment of existing theories by adding relations between concepts based on new sources. This allows acceptance of new theories that can then, in turn, support the translation of theoretical concepts into (new) variables that are included in the new sources. This inductive research will lead to new and commonly accepted understandings, and new paradigms and axioms that form the basis for assessing the validity of certain measurements. With new data sources the concept of 'validity' maintains its value. However from being intrinsic to an ex ante assessment, it will develop into a concept used to justify, ex-post, the use of new sources in proving evidence. The ES-CoP and ESGAB's assessment of its implementation will have to take account of the decision-making processes involved in making these conceptual choices.

6. From valid statistics to evidence

Official statistics should comply with the requirements of the 4th and 5th level of quality. The principles of the ES-CoP describe the requirements for the institutional arrangements of the statistical offices (NSI's and ONAs). Only compliance with these principles assures that official statistics can be trusted as evidence in policy making and monitoring.

In general, organisations involved in producing new data sources like Big Data, because of their nature, are far from being compliant to these CoP requirements. Budget security, professional independence, the use of the most suitable methodologies, awareness of response burden, high quality statistical staff, release calendars, confidentiality protocols, to name a few elements, are not considered at all to be relevant for these organisations. It will also be impossible to request these organisations to become compliant to those principles. Nevertheless, the organisations producing official statistics, the potential users of these new data sources, should themselves be compliant. This is essential to protect their status as trustful and authoritative produces of high quality statistics and avoid the risk that the use of new data could negatively influence their reputation. New forms of quality assurance could include transparent declarations on the characteristics of the producers of the new data and the quality of the data themselves.

ESGAB in assessing the quality of official statistical organisations will have to assess the adequateness of the governance structures of the organisations to handle these new data and data providers. Issues like independence from these providers, the guarantee of confidentiality of the concerned citizens and businesses, the good description of the meta data are to be assessed. To guide the work by ESGAB and more important to function as a self-regulatory system for those who subscribe to the ES-CoP the Code will have to be expanded with principles that describe the most relevant issues in using these new sources.

A set of principles in the CoP that describe the characteristics of the cooperation and relationship with the owners and keepers will therefore have to be developed. This means that the user of the data should be assured sufficiently that the Code requires and ensures cooperation with the keepers and owners of these new data sources in a manner that guarantees compliance.

7. Conclusions

As a result of the upcoming use of new data sources, like Big Data, in official statistics, the European Statistics Code of Practice, as a self-regulatory system as referred to in Regulation 223/2009, will have to be updated with a principle(s) that describe the practice for official statistics in governance of these new data sources. This will include descriptions of the organisation of the cooperation with the owners of these data sources and describe the manner the official statisticians are using these data and agreements made with the owners of the sources. Beyond this, the principles will have to refer to a scientific approach justifying the concepts to be equally valid representations of theoretical constructs as well identified constructs from existing theory. The European Statistics Governance Advisory Board, using the ES-CoP as their main reference document for describing compliance to the highest statistical governance standard, will therefore, also have to adapt to the inclusion in statistical production of these sources and the characteristics of the stakeholders involved.

8. References

Galtung, J, 1967, Theory and Methods of Social Research, London, George Allen & Unwin

Karkkainen, K. and L. Del Bario, European Statistical System peer reviews: an efficient means to implement the European Statistics Code of Practice; European Conference in Quality in Statistics, Madrid 2016

Korteweg, P and J. Van Weesep (eds), 1983, Ruimtelijk onderzoek, Leidraad voor opzet, uitvoering en verwerking, Bussum, Unieboek.

Kitchin, R., 2014, The Data Revolution London, Sage.

Annex 1

The European Statistical Governance Advisory Board is an independent body established to oversee how the European Statistics Code of Practice is implemented in the European Statistical System.

The Advisory Board carries out its mission:

- through an annual report for the Parliament and Council on the implementation of the Code of Practice by the Statistical Office of the European Communities (Eurostat);
- through an assessment of the implementation of the Code in the European Statistical System as a whole, included in the annual report;
- by advising on the implementation of the Code by Eurostat and the European Statistical System as a whole;
- by advising on communicating the Code to users and data providers;
- by advising on the updating of the Code.

The Advisory Board may advise the Commission to build user confidence in European statistics.