Challenges to support Official Statistics production with new data sources in Emergency Situations:

Contribution to the panel discussion.

Piet J.H. Daas

Statistics Netherlands, Eindhoven University of Technology

**Keywords:** Big Data, methodological challenges, future research topics, reliable use.

# Introduction

New data sources (i.e. Big Data) are potentially very interesting for official statistics. When information is in high demand, the biggest advantages of these kind of sources are their rapid availability and large volume [1]. The latter may enable the production of very detailed statistics, for instance for small areas, in near real-time and also have the potential to make production processes more efficient. However, a number of challenges still need to be solved to enable the reliable use of Big data for statistical purposes [2,3]. These challenges have methodological, technical and cultural aspects [1]. In this abstract the methodological challenges regarding the use of Big data in official statistics are identified and discussed.

# Methodological challenges

The methodological challenges were identified via a literature study combined with the findings of the ESSnet Big Data I & II [4]. A total of 8 challenges have been identified [3]. However, the reader should be aware that Big data of poor quality [5] will never be successfully used even when all challenges listed below have been solved.

## Combining Big data with other sources

It is challenging to link Big data sources to other data sources. The main reason for this challenge is the fact that during linking -traditionally- a unit-oriented view is used and that limited data is available on the units included in many of the Big data sources used (section 4.1 in [6]). Although alternative approaches have been proposed, such as linking at the area or period level [2,4], developing methods to enabling the linking of Big data at the unit level with other sources will greatly improve the use of Big data in many areas. For Big data source that contain events, these first need to be converted to a unit relevant for official statistics.

## Reliable inference from Big data

This topic has both a population and a conceptual challenge.

### Population level

When a Big data source does not include the whole target population, one needs to -somehow- correct for the differences between the population included in Big data and the target population. This is especially challenging when Big data is used as the main source (section 6.4 in [6]). When Big data is used as an additional source, the source to which the data is added can be used to correct for that. In addition, coping with the dynamics of the population in Big data make this topic even more challenging.

### Conceptual level

A Big data source may not contain the variable of interest at the exact definition needed. In such cases, since the data is given, the conceptual measure of interest needs to be (attempted to be) derived -in some way or another- from the data available (section 6.3 in [6]). This is usually referred to as harmonisation and can be challenging. Some of the examples in [6] indicate that it is certainly possible to extract concepts for official statistics in a reliable way. Another important aspect is the stability of the derived variable over time (see 2.5).

## Application of data science methods in official statistics

More and more examples emerge that demonstrate the successful application of data science methods, such as Machine Learning, Deep Learning and Artificial Intelligence methods, in official statistics (section 2.9 in [7]). This topic has validation, transparency and optimization aspects. The points are strongly connected but discussed separately below.

a) Model validity: Deep Learning has, for example, been successfully applied to identify particular crops on satellite pictures (section 2.22 in [6]). However, the model developed for this task is composed of a neural network from which the features used cannot be directly inspected (see the next point). It is challenging is to find ways to validate the findings of these data science models.

b) Transparency: Some of the data science methods used are not very transparent. The example provided under the previous point, for instance, demonstrates this. In this case, the features used by the Deep Learning model cannot be directly inspected. This clearly reduces the transparency of the method used. The latter is one of the fundamental principles of official statistics [8] and is mentioned in a considerable number of the Deliverables of the ESSnet BD I & II [4]. The challenge here is to find ways to get insight into the way such data science models work.

c) Optimization: Ideally someone wants to develop a model that is highly accurate on the test set, has a low variance, low bias, and generalises well to unseen data. Creating a model that includes the best possible combination of these properties is challenging not only because it requires a lot of computational power and takes considerable time but also because there is a trade-off between some of these demands. Finding the optimal algorithm and best set of hyperparameters is a major challenge.

## Correlations and Big data

When studying large amounts of data, it is not unexpected to find a correlation between a Big data derived series and an (already existing) official statistic. This could simply be due to a spurious correlation but may also indicate a -very interesting- (new) finding for which the Big data source could be applied. Discerning between both cases is challenging. Studying how the relation develops over time and looking at it from a causal perspective may provide clues on the finding. Co-integration is another interesting way of looking at it.

## Dealing with a changing world

Access to and the composition of Big data may change considerably over time (section 2.3 in [7]). For instance, some data sources may become completely unavailable because the company that produces them stops their activity on that particular topic, decides to make the data no longer publically available or blocks access because of a regulatory (law) change. Other changes may have a less radical effect, such as a change in the composition of the data source by adjusting, removing or including one or more variables. In each of these cases, the organization that has setup a process by which Big data is used needs to be able to adjust the existing process to these (unforeseen) changes. This suggests the need for the development of so-called fall-back scenarios (when a data source or variable is no longer available), procedures to correct for any variables changed and/or make arrangements with Big data providers to assure availability of the data. Another very important point in this context is the observation of so-called ‘concept drift’. Here, models developed at a specific point in time start to deteriorate and, as a result, decrease their performance. This is caused by a changing world in which the relation initially observed and included into the model (gradually) changes. Finding ways to deal with changes in (Big) data sources and/or changes in the world around us is a challenge with a considerable number of methodological aspects.

## Dimensionality reduction of Big data

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space. Preferably the transformation is done in such a way that the low-dimensional representation retains the most meaningful features of the raw, original, data. The advantage of this reduction is that working in high-dimensional spaces is often undesirable because the raw (high dimensional) data tends to be sparse and analysing it may be computationally intractable. At various stages this problem has been ran into in the ESSnet and is mentioned, for instance, in Del 2.2 of WP2, Del 6.3 of WP6, Del 7.3 of WP7, Del C2 of WPC. Developing methods that enable the extraction of the best set of reduced features, which are preferably interpretable, for a whole range of Big data sources is a very interesting challenge and this work would greatly support the research in many of the topics mentioned above.

## Privacy protection and Big data

Traditionally, statistical disclosure control by national statistical institutes has focused on tables and microdata collected and produced by the institutes themselves. However, the increasing use of Big data makes it possible to create large data collections with very detailed information on units obtained from many sources. This abundance of data poses new problems for statistical disclosure control [1] as well as methodological challenges, which need to be addressed. For official statistics the most important methodological challenge is to find ways to decrease the change of (re)identification of persons or businesses. This topic could also include studies in the area of secure sharing of data (section 2.8 in [7]).

# Conclusions

In this abstract the most important methodological challenges when using Big data in official statistics are discussed. Guidelines are proposed in [3]. They are certainly not finalized and are based on the current state-of-art. For the reliable use of Big data, it is essential that progress needs to be made in, certainly, a number of the topics identified. Considering the rising need for, the effort in and the increasing interest in the use of Big data such progress will certainly be made in the coming years. In addition, it is also important to tackle (a number of) both technical and cultural barriers [1] as they will certainly hinder the future application of Big data in official statistics.

# References

[1] De Broe, S. et al. Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics. Statistical Journal of IAOS (2020), *accepted for publication*.

[2] Kitchen, R. Big Data, New Epistemologies and Paradigm Shifts. Big Data & Society, April–June (2014), 1-12.

[3] Daas, P., et al. Report describing the methodological steps of using big data in official statistics with a section on the most important research questions for the future including guidelines. Deliverable K10, WPK, ESSnet Big Data II (2020).

[4]  ESSnet Big Data, Main page of the project website, link: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main\_Page (2020).

[5] Quaresma, S. et al. Revised version of the Quality Guidelines for the Acquisition and Usage of Big Data. Deliverable K3, WPK, ESSnet Big Data II (2020).

[6] Daas, P., et al. Revised version of the Methodological report. Deliverable K9, ESSnet Big Data II (2020).

[7] Consten et al. Report describing the methodology of using Big Data for Official Statistics and the most important questions for the future studies. Deliverable 8.4, WP8, ESSnet Big Data I (2018).

[8] United Nations. Fundamental Principles of Official Statistics. A/RES/68/261, 3 March. Link: https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf, (2014).