ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Work package K

Methodology and quality

Deliverable K10: 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

Final version, 20.11.2020

Prepared by:
Piet Daas, Marco Puts (CBS, NL)
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
Sónia Quaresma (INE, PT)
Tiziana Tuoto, Loredana Di Consiglio, Giovanna Brancato, Paolo Righi (ISTAT, IT)
Magdalena Six, Marlene Weinauer, Alexander Kowarik (STAT, AT)

Work package leader:

Alexander Kowarik (STAT, AT)

alexander.kowarik@statistik.gv.at

telephone : +43 1 71128 7513

Contents

1. Introduction

More and more National Statistical Institutes are investigating the potential of using Big data. This has resulted in a number of statistics created that use Big data at various stages of production (Del K9). An overview of the 13 examples found is shown in Table 1. Not all processes have already completely matured and only two of them are in production. These numbers will undoubtable increase in the near future.

Table 1. Overview of Big data based official statistics and their status.

No.	Statistic	Data sources used	Countries	Status	Ref.
1	Consumer Price Index	Scanner data, web scraped prices	Various countries	In production	[1]
2	Traffic intensities	Road sensors	NL	In production	[2]
3	Online job vacancies	Scraped job vacancies	ESSnet	Towards implementation	Del. WP1
4	Enterprise characteristics	Scraped websites	ESSnet	Towards implementation	Del. WP2
5	Electricity/energy consumption	Smart meter data	ESSnet	Towards implementation	Del. WP3
6	Maritime and Inland waterway statistics	Automatic Identification System data	ESSnet	Towards implementation	Del. WP4
7	Financial transaction based statistic	Bank transaction data	ESSnet	Exploratory	Del. WPG
8	Earth observation derived statistics	Satellite and Arial pictures	ESSnet	Towards implementation	Del. WP7 WPH
9	Mobile network derived statistics	Mobile Network Operations data	ESSnet	Towards implementation	Del. WPI
10	Innovative tourism statistics	Various sources (e.g. webpages)	ESSnet	Exploratory	Del. WPJ
11	Innovative companies (small)	Scraped websites	NL	Towards implementation	[3]
12	Social mood on economy index	Social media messages	IT	Experimental	[4]
13	Mobile phone derived outbound tourism	Mobile Network Operations data	FI, EE	Experimental	[5,6,7]

In this report the processes developed for the examples listed combined with the personal experiences of the authors are used to derive a generic process by which Big data can be used in official statistics production. Be aware that, as an important starting point, the focus of this process is on using Big data as the main data source and not when used as an additional data source; see Del. WP8.1 and [8] for more info on this very important difference.

1.1 Generic process on using Big data

The basis for the workflow is the one described for the production of Traffic Intensity statistics based on road sensors [2,9]. In this process, first a select and reduce step is applied to the raw data (Step 1).

This seriously reduces the amount of relevant data remaining; e.g. lots of unneeded data, such as variables and units, are removed. Next information is extracted from the data remaining (Step 2). This, combined with the frame (Step A), is used in the last inference step (Step 3) to produce a statistic. In Figure 1 an overview is given of this generic process.

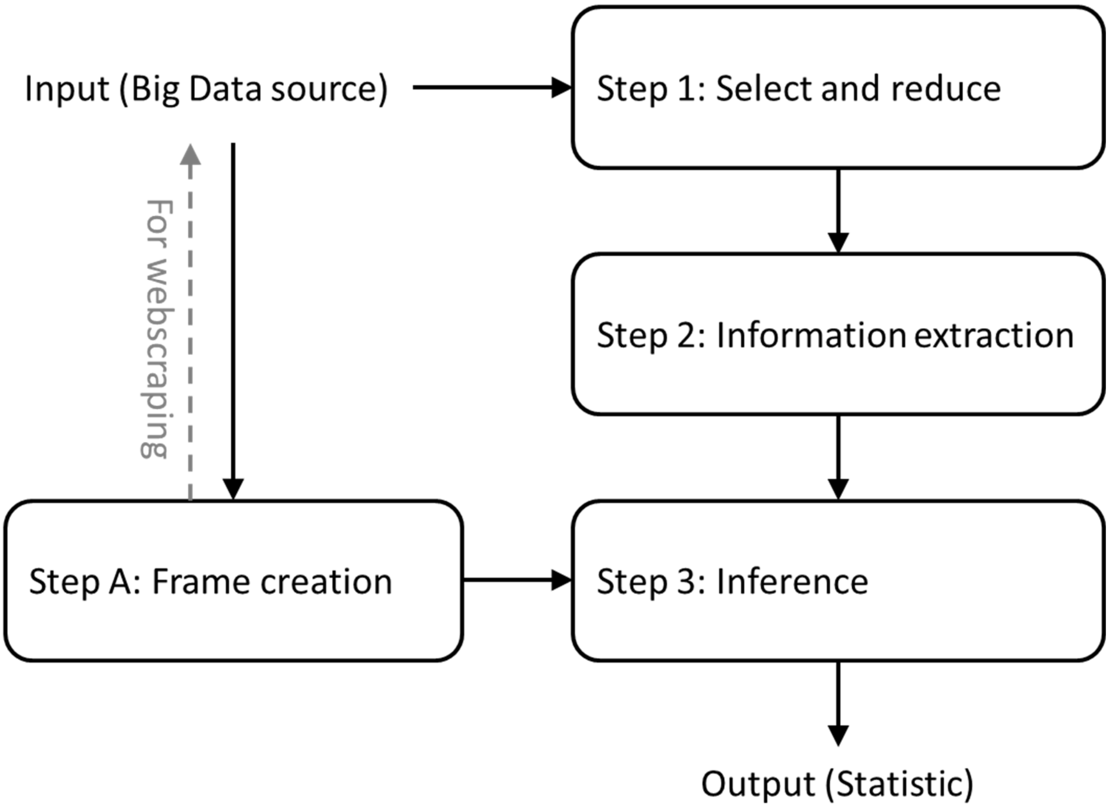

Figure 1. Overview of the 4 steps required in Big Data based statistics production.

Each of these steps can be related to the Business Functions (BFs) in the BREAL architecture. The latter architecture is specifically developed for big data (Del F1, WPF). In BREAL, step 1 corresponds to the 'Data Wrangling' BF, whereas step 2 relates to the 'Modelling and Interpretation' BF. Step A corresponds to the 'Enrich Statistical Register' and 'Support Statistical Production' BFs. The last step, step 3, is a 'Modelling and Interpretation' BF and also includes a 'Shape Output' BF. However, because step 2 and 3 point to a similar BF in the BREAL architecture but clearly differ in their overall purpose, we have chosen to keep using our own initially proposed names in this document. It is suggested here, to update BREAL with an 'Information extraction' BF.

2. Big Data as the main source

While studying the list of Big data based statistics included in Table 1, it became clear that the process originally developed for the processing of road sensor data [9] -in principle- contained the (general) building blocks observed in many of the other Big data based statistical processes. These steps are described in more detail below.

2.1 Step 1: Select and reduce

Usually, the first step in the process involves a selection at the unit and variable level. As a result the amount of data is often seriously reduced [2]. Goal of this step is either predominantly focussed on the removal of unwanted data, the selection of relevant records or a combination of both. It may require the need for specific quality indicators used as selection criteria. Examples of this step are: i) the selection of products from scanner data [1], ii) the selection of road sensor data to only include data of high quality of a specific set of variables needed [2], iii) the selection of AIS-messages that contain info on the position of ship (Del. WP4), iv) the splitting and selection of areas on satellite pictures for the particular region or country studied (Del. WP7), and v) the selection of social media messages containing specific words [4].

For each of the 13 examples listed in Table 1 this step has been reviewed. The findings are described in table 2.1 below.

Table 2.1. Overview of Big data based official statistics in relation to step 1.

No.	Statistic	Step 1	Ref.
1	Consumer Price Index	Select products needed	[1]
2	Traffic intensities	Select high quality road sensors and variables needed	[2]
3	Online job vacancies	Obtain job vacancies from web sites, remove duplicates and ghost vacancies	Del. WP1
4	Enterprise characteristics	Scrape data from enterprise websites and select relevant parts	Del. WP2
5	Electricity/energy consumption	Collect data from smart meters, select high quality data	Del. WP3
6	Maritime and Inland waterway statistics	Select required type of AIS messages and remove clearly erroneous ones	Del. WP4
7	Financial transaction based statistic	Select transactions relevant for the topic studied	Del. WPG
8	Earth observation derived statistics	Select pictures that include the area studied, split them in equally sized parts	Del. WP7 WPH
9	Mobile network derived statistics	Select mobile phones in MNO data for area and period studied	Del. WPI
10	Innovative tourism statistics	Select data relevant for area and period studied from data sources available	Del. WPJ
11	Innovative companies (small)	Scrape data from enterprise websites and select relevant parts	[3]
12	Social mood on economy index	Obtain messages for area and period studied, select relevant messages (or remove irrelevant ones)	[4]

13	Mobile phone derived outbound tourism	Select mobile phones in MNO data of providers for area and period studied	[5,6,7]

2.2 Step 2: Information extraction

In the next step, the data remaining is processed in such a way that the information required for the specific statistical need is obtained. This may simply involve extracting a value from a dataset, may require applying a filter or a transformation of the data, may involve a model-based approach or a combination of these steps. In the model-based approach, usually an underlying, not-directly measurable variable is estimated. Examples of information extraction are: i) determine the elementary aggregate of similar products sold [1,10], ii) correct for the missing data of road sensors [2], iii) remove outliers in the location data of ships (Del. WP4), iv) determine the type of crop grown an a patch of land from a picture (Del. WP7), v) determine if a company is innovative or not from web site texts [3], vi) determine the sentiment of social media messages [4], and vii) derive if a particular social media message indicates a person with a particular disease [11]. From these examples it is clear that very diverse methods can be applied in this step.

For each of the 13 examples listed in Table 1 this step has been reviewed. The findings are described in table 2.2 below. In the table the step where a statistical model is created when needed to extract information from the data source has not been included. Here, it is assumed that such a model is already available.

Table 2.2. Overview of Big data based official statistics in relation to step 2.

No.	Statistic	Step 2	Ref.
1	Consumer Price Index	Obtain prices and/or amounts of products sold	[1]
2	Traffic intensities	Correct for missing data	[2]
3	Online job vacancies	Extract job and company info	Del. WP1
4	Enterprise characteristics	Extract characteristics needed (direct or indirectly)	Del. WP2
5	Electricity/energy consumption	Determine characteristics of data to obtain info on units from which data is obtained (household vs company)	Del. WP3
6	Maritime and Inland waterway statistics	Filter location data to remove outliers (erroneous location data), remove erroneous records	Del. WP4
7	Financial transaction based statistic	Extract money earned or spend from the records selected	Del. WPG
8	Earth observation derived statistics	Extract required information from picture (e.g. solar panels, crops grown, urbanization level, …)	Del. WP7 WPH
9	Mobile network derived statistics	Extract information from MNO data available from mobile phones studied (e.g. location, foreign phones)	Del. WPI
10	Innovative tourism statistics	Extract information from combined data sources available (for area and period studied)	Del. WPJ
11	Innovative companies (small)	Extract and process texts from websites and determine if a company is innovative or not	[3]
12	Social mood on economy index	Determine social mood (sentiment) of messages selected	[4]

13	Mobile phone derived outbound tourism	Extract information from MNO/Mobile phone data of providers to detect which mobile phones have left the country (and where to)	[5,6,7]

2.3 Step A: Frame creation

In addition to the steps listed above, somewhere in the process a link has to be made between the population included in the Big data source and the target population of the statistics. Here, two situations may occur. In the first situation, an existing population frame is used to select the data is collected later on in the process. This happens, for instance, when data is collected via web scraping. Here, typically, the Business Register is used to select the business websites that will be scraped and studied (indicated by the dashed line in Fig. 1). Establishing the correct relation between a (statistical) business unit and a website is important here. For this, various approaches have been developed, such as URL-finding [12] or use the URL provided by the Chamber of Commerce [3]. By comparing some of the data on the website, such as the name, address and Chamber of Commerce number, with those in the business register one is able to check if the correct link has been made.

The second situation occurs when Big data that is available (i.e. has been collected) is compared with a population frame. This typically requires getting information from the units included in the Big data source. This can be challenging and may sometimes seem impossible. A very challenging example is the population active on social media [4]. Here, accounts of persons and companies are included. Many of these accounts do not provide enough information, such as background characteristics, to determine the exact person or company to which the profile belongs. When that occurs, one can try to derive any information relevant for that (indirectly) from the data that is available. An example of this is deriving the gender (and age) from the profile of a Twitter account [13].

Examples of studies in which the target population frame has been successfully obtained from the Big data source alone are: i) traffic intensity statistics based on road sensor data [2] and iii) AIS-based maritime travel data (Del. WP4). In all these cases, the population units are (nearly) completely included in the data used. To obtain the target population, it was essential to link the sensors to the roads location data [2] and remove the ghost ships and non-maritime ships from the AIS-data (Del. WP4), respectively. In cases where the population is not fully included, one may need to use non-probability based approaches, apply capture-recapture methods or decide to use the Big data extracted information as a proxy. Examples of the latter are online job vacancies as a proxy for vacancy statistics (Del. WP1) or social media sentiment as a proxy for economic mood [4].

For each of the 13 examples listed in Table 1 this step has been reviewed. The findings are described in table 2.3.

2.4 Step 3: Inference

In the end of the process the frame and the information extracted meet. In this step, nearly always a model-based approach is used in which one aims to infer statistics for the entire population (Del. WP8, Del. WPK) and corrects for any biases [3]. In a number of cases this is the second step of the process where models are used; e.g. no. 2 and 11 in Table 1. Examples of step 3 are i) creating a CPI-index based on the products sold [1], ii) estimating the traffic intensity for road segments [2], iii) estimating the number of small and large innovative companies in a country [3], and iv) estimating the number of susceptible with COVID19-symptoms on social media [11].

Table 2.3. Overview of Big data based official statistics in relation to step A.

No.	Statistic	Step A	Ref.
1	Consumer Price Index	Determine products for which data needs to be collected (content of the 'shopping basket')	[1]
2	Traffic intensities	Link road sensors to roads and road segments	[2]
3	Online job vacancies	Determine population (companies) for which job data can be obtained (compare with the target population)	Del. WP1
4	Enterprise characteristics	Link URLs of websites to enterprises in Business Register	Del. WP2
5	Electricity/energy consumption	Link smart meter accounts to target population (persons or companies)	Del. WP3
6	Maritime and Inland waterway statistics	Determine population of ships from AIS data available (remove erroneous ids; i.e. 'ghost ships')	Del. WP4
7	Financial transaction based statistic	Determine for which part of the population data is available in source and compare with target population	Del. WPG
8	Earth observation derived statistics	Determine data availability and coverage for the area and period studied	Del. WP7 WPH
9	Mobile network derived statistics	Determine part of the population for which MNO data is available	Del. WPI
10	Innovative tourism statistics	Determine population included in data sources available and compare with target population	Del. WPJ
11	Innovative companies (small)	Link URLs of websites to enterprises in Business Register	[3]
12	Social mood on economy index	Determine accounts active in area during the period studied, select accounts of population studied (persons or enterprises)	[4]
13	Mobile phone derived outbound tourism	Determine from MNO data/Mobile phone data of providers which part of the phones belong to target population studied, for the area and period studied. In particular which part has left the country (and returned)	[5,6,7]

For each of the 13 examples listed in Table 1 this step has been reviewed. The findings are described in table 2.4.

2.5 Discussion

From the above, it is clear that the process shown in Figure 1 contains the steps relevant for the production of statistics based when Big data is used as the main source. In the next chapter alternative ways to use Big data are discussed. These findings will be linked to the process described in this chapter.

Table 2.4. Overview of Big data based official statistics in relation to step 3.

No.	Statistic	Step 3	Ref.
1	Consumer Price Index	Determine index from data available	[1]
2	Traffic intensities	Determine traffic intensity per road segment, correct estimate for poorly covered segments	[2]
3	Online job vacancies	Determine (proxy of) job vacancy statistics from data available	Del. WP1
4	Enterprise characteristics	Update and add data to Business Register for enterprises studied	Del. WP2
5	Electricity/energy consumption	Determine energy/electricity consumption from data available, correct for missing part of the population	Del. WP3
6	Maritime and Inland waterway statistics	Determine waterway statistics for topic, for country and period studied	Del. WP4
7	Financial transaction based statistic	Create statistics for the topic, area and period studied, correct for coverage issues/bias	Del. WPG
8	Earth observation derived statistics	Create statistics for topic, for period and area studied	Del. WP7 WPH
9	Mobile network derived statistics	Create statistics for topic, for area and period studied, correct for missing part of the population	Del. WPI
10	Innovative tourism statistics	Create statistics for the topic, area and period studied, correct for coverage issues/bias	Del. WPJ
11	Innovative companies (small)	Create statistics for small and large innovative companies, correct for bias in data	[3]
12	Social mood on economy index	Determine overall mood for area during period studied, correct for coverage errors and bias	[4]
13	Mobile phone derived outbound tourism	Create statistics regarding the number of mobile phones for the country studied have visited other areas during the period studied, correct for bias/coverage	[5,6,7]

3 Big data as auxiliary data

Apart from using Big data as the main source of information (as discussed above), it can also be used as auxiliary data. Here, the information provided by Big data is added to information obtained from another source, such as survey or administrative data. There are a number of ways in which the information provided by Big Data can be included. These are: i) as an additional source, ii) to replace survey questions, and iii) to calibrate a model-based estimate. Each of these uses is discussed below

3.1 Big data as an additional source

In this case, Big data is used to increase the quality of the findings based on the 'other' data source [8]. It has the additional advantage that, when the 'other' data source contains a known part of the population (one that can be made representative of the target population), the selectivity issue of Big data is no longer a major concern.

A considerable number of studies have been published were Big data is used as an additional source. The most important examples are listed below.

1. Several studies describe how survey data can be combined with Big data to make detailed regional predictions for wellbeing and poverty. For example, Marchetti et al. [14] describe how mobility data recorded from GPS is used as a covariate in an area level model to predict poverty on a low regional level in Italy. Schmid et al. [15] use mobile phone data to estimate literacy rates in Senegal.
2. In a number of studies machine learning algorithms are used to determine the relation between survey data and sensor or mobile phone data. Here, the latter data set is subsequently used to make detailed regional predictions. For example, Noor et al. [16] analysed the correlation between night-time light intensity from satellite images and survey sample data on household income in Africa. In another study, day time satellite images were used to predict well-being with deep learning [17], while others combined mobile phone data with survey data on poverty and used this to predict poverty and well-being on small regional level in Rwanda [18].
3. Another example is the use of AIS data for Inland waterway statistics of the Netherlands [Del. E1, WPE]. Here, plans are to include the data of the journey of ships to correct for undercoverage.
4. Another way of including Big data is in a multivariate structural time series model. In such models, auxiliary series derived from Big data sources are combined with time series obtained from repeated sample surveys and/or series derived from registers. An example is the study of Harvey and Chung [19] in which a time series model is proposed for the Labour Force Survey in the UK extended with a series of claimant counts. Other examples are studies in which social media or Google trends are used to increase the prediction speed of survey based estimates, such as Consumer confidence [20] and Unemployment estimates [21]. Here, the Big data series are available at a higher frequency than the series obtained with repeated surveys. This enables the use of a time series modelling approach to make predictions for the survey outcomes in real time at the moment that (the outcomes for) the Big data series are available.

3.2 Big data to replace survey questions

Here, Big data is used as a source of information to replace survey items. Replacing survey items with estimates from a Big data source is beneficial in several ways, e.g. Big data may allow for more timely collection, reduce response burden, may reduce cost and time in the production of results (in long term), allow for increased sample sizes and may extend the population surveyed. Benefits strongly depend on the source used.

However, the replacement may also introduce additional obstacles. The main problems identified in replacing survey questions with Big data are coverage and conceptual issues. Therefore, two important questions need to be considered:

1. Is it possible to establish the population of interest?
2. Are definitions in the Big data source coherent with definitions of the survey? (A negation may be overcome by changes of survey definitions by e.g. responsible Working Groups and Task Forces)

Therefore, the same data source may lead to different obstacles in different use cases. This is outlined with three use cases. In the first two web data is used and in the third case sensor data of farm equipment is used to replace survey questions.

Use case 1: ICT survey

Web Data is used to replace survey questions from the annual Eurostat ICT enterprise surveys. These include whether the enterprise has a website, a link to social media on the website and an online shop. (WPC). This case has the advantage and disadvantage typical for webscraped data; e.g. comparability over time because of continuous updates, etc. Looking at the two key questions reveals that this use case looks very promising.

1. Yes, it is possible to establish the population of interest. There is a coverage problem in scraping businesses in general: not all businesses have a web site and some types of businesses have a higher change of having a web site, leading to an over-coverage of large enterprises. While this is problematic in drawing other than internet-applied enterprise characteristics (e.g. whether the enterprise is innovative) from web data, it is no problem for the ICT target variables: Existence of Website, Social Media and Webshop.
2. Yes, the definitions are mostly coherent. Patience needs to be paid to the definition of enterprise groups. A more concrete definition of social media by the ICT working group could even increase coherence between both sources.

Use Case 2: Job vacancy survey

Here, web data is used to replace survey questions from the Job Vacancy Survey, e.g. the number of job advertisements in different sectors (WPB). Again, this case has the advantage and disadvantage typical for webscraped data. However, compared to the first case this use case looks less promising.

1. No, it is not possible to establish the population of interest. Not all job vacancies are advertised online. Some types of jobs might more likely be advertised online than others.

This means that online job advertisement (OJAs) data might not only miss many jobs, but might also not be representative for the overall job market.

2. No, important variables for comparing online job ads and JVS are not available.

Use Case 3: Crop yield survey

Here, sensor data of farm equipment is studied to see if it can be used to replace survey questions for agricultural statistics [22]. Sensor data collected by farm equipment was studied via the John Deere portal. Potentially, this looks very promising but getting stable access to the data will be challenging.

1. Yes, it is possible to establish the population of interest. Data is available at the farmer level.
2. Yes, an almost 100% overlap was found with the field operations data collected by the sensors. Other available data of potential interest are with machine data, agronomic service providers activity, soil and environmental conditions.

By now, no survey items of Eurostat questionnaires are productively replaced with Big data. But diverse promising pilots are running.

3.3 Big data to calibrate a model-based estimate

Last but not least, Big data can provide potential to reduce the bias in surveys: Therefore, auxiliary information from Big data is used for calibration in survey weighting or more general in model-assisted survey estimation. Two examples for which this is the case are:

Mobile Phone data:

Surveys with target variables on travel habits benefit from additional information from mobile phone data. In Austria, since 2018 the traditional calibration with age, sex and region is extended with information of daily stays from mobile phones. Here, mobile data's information of active SIM cards abroad is used [5]. Similar approaches are conducted in e.g. Estonia and Finland [6,7].

Web data:

Although the conduction of the ICT survey in enterprises is obligatory on European level, the participation of businesses may be voluntary on national level. If so, often a bias is observed that 'digital-developed' enterprises are more likely to participate in the survey. Web-scraped information on web activity, which is highly correlated to the extent of digital development of a business, may help to reduce this bias.

As for administrative data, also for Big data the calibration is only as useful as the quality of the data used for calibration itself. In most setups the auxiliary information is defined as known error free information and it is an additional challenge to include an estimated uncertainty from the Big data source in the survey error estimation. Therefore – as for all use cases of Big data, but particularly for the usage of big data for calibration – high quality Big data must be ensured.

Another example is a study presented at the BigSurv20 conference were the use of TV-tuning (Big) data was discussed to calibrate panel data collected by a commercial company [23,24].

3.4 Discussion

When Big data is used as auxiliary data, it is clear that somewhere in the process the information extracted from Big data has to be combined with that of the other data set. The examples mentioned above, clearly indicate that this occurs *after* the information has been extracted step for the Big data source; i.e. prior to the inference stage. Comparing these findings with the process provided in Figure 1, makes it obvious that -when Big data is used as an auxiliary data source- steps 1 and 2 are both included. This suggests that the process in Figure 1 very likely includes all steps relevant for the creation of Big data aided statistics.

Apart from the uses mentioned above, it can be envisioned that in the future Big data could also function as a sampling frame to start collecting data via a survey (or another form of data collection). For instance, by inviting users of a particular social media platform, for instance those that post messages on a specific topic, to participate in a related survey.

4. Research questions and guidelines

In this chapter both the important research questions regarding the use of Big data in official statistics and any guidelines available are discussed. We prefer to designate the research questions as challenges here, as this better reflects what is presented in the chapter. The guidelines are based on a literature study and the findings of the ESSnet Big Data I & II. When guidelines derived from examples of using Big data as an additional data source are included this will be indicated explicitly.

From the work performed in the ESSnet Big Data I & II and the overview of the methodological (Del K9) and quality based issues (Del. K11) obtained, it has become clear that a considerable number of important challenges have been raised. Investigating these challenges with the aim to solve them, will greatly stimulate the use of a whole range of (new) Big data sources in official statistics and improve the quality of the statistics produced (Del. K 11). Many of these challenges are related. However, the reader should be aware that Big data of poor quality will never be successfully used even when all challenges listed below have been solved.

4.1 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 (Del. K9, section 4.1). Although alternative approaches have been proposed, such as linking at the area or period level (see above and Del. K9, chapter 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.

Guidelines:
Several ways to combine Big data with other sources have been applied within the list of examples included in Table 1. In some cases individual units in Big data source can be linked at the most detailed level with other data via geolocation coordinates [2, WP7, WPH, WP4], product codes [1] or a combination of other characteristics [WP 1-3, 3]. When the data is available at the country or area level, linking at a higher level of aggregation has been applied [4,5,6]. Another way of combining Big data with other sources is in a multivariate structural time series model. This requires the need for a Big data derived time series and repeated measurements over time in the other data source used [19-21].

4.2 Reliable inference from Big data

This topic has both a population and a conceptual challenge.

4.2.1 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 (Del K9, section 6.4). When Big data is used as an additional source, the source to which the data is added can be used to correct for that (see section 2.5). In addition, coping with the dynamics of the population in Big data make this topic even more challenging.

Guidelines:
Here, reliable estimates can be obtained when either the whole target population is included in the Big data source [1-3, WP4] or when Big data can be linked to a data source that can be made representative for the target population [19-21]. This may require bias correction [3]. In all other cases, at best, a proxy for the population estimate is obtained [5-7, WP1,WP3]. When the Big data is rapidly available the estimate can be used as an early indicator to detect changes (WP6).

4.2.2 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 (Del. K9, section 6.3). This is usually referred to as harmonisation [25] and can be challenging. Some of the examples in Table 1 indicate that it is certainly possible to extract concepts in a reliable way. Another important aspect is the stability of the derived variable over time.

Guidelines:
Some Big data sources contain the variable of interest at the level of definition needed [1,2,WP3,WP4,WPG,5-7]. For the others the use is more derived. It is essential to keep checking the stability of this relation over time. See also the guidelines for topic 5.

4.3 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 (Del. WP8.1, section 2.9). This topic has validation, transparency and optimization aspects. The points are strongly connected but discussed separately for completion here.

a) *Model validity:* Deep Learning has, for example, been successfully applied to identify particular crops on satellite pictures (Del K9, section 2.22). 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 [26] as is mentioned in a considerable number of the Deliverables of the ESSnet BD I & II. 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.

Guidelines:
Many data science methods have been used when extracting information from Big data; this is especially the case for text and image based sources (Del K10, WPK). The downsides of these methods have been described above. We will focus mainly on the validity and transparency aspects here as these are essential when producing official statistics. One way to check the validity of data science based findings is to compare the estimate with the official number, if available. However, this is only possible when i) the data science based estimate is derived from a source that can be made representative for the target population [3] or ii) when a series of estimates is available and its development can be compared with a series of official numbers [27]. In all other cases, the result may differ which does not provide a clue on the model's validity. This difference could be due to population composition difference, differences between de concept measured and the one needed and/or model misspecification. Regarding the transparency of the methods applied, it is obvious that ways need to be found to obtain information on the way such models work. Research in the area of fair, transparent and responsible AI will certainly assist here [28].

4.4 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 [29]. Discerning between both cases is challenging. Studying how the relation develops over time [30] and looking at it from a causal perspective [31] may provide clues on the finding. Co-integration is another interesting way of looking at it [19,32].

Guidelines:
In a number of cases, interesting correlations observed between a Big data derived series and an official statistics greatly stimulated a continuation of the initial exploratory study [2,4,WP6, WP1,33]. However, often its challenging to proof this correlation was based on an actual association and not merely spurious. Alternative methods to study the relation between data series, such as co-integration [32], have been successfully applied to gather supporting evidence of an association between series. Studying how the relation develops over time [30] and looking at it from a causal perspective [32] may also provide clues on the finding. In one case, determining the effect of the Big data derived series on the estimate of a survey based time series model unequivocally demonstrated that both series were associated by an underlying cause [20].

4.5 Dealing with a changing world

Access to and the composition of Big data may change considerably over time (Del. WP8.1 section 2.3). For instance, some data sources may become completely unavailable because the company that produces them i) stops their activity on the topic included, ii) decides to make the data no longer publically available, iii) change the conditions under which the data is generated without communication this or iv) 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) [34], 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' [3,35]. Here, models developed at a specific point in time start to deteriorate and, as a result, decrease their performance [36]. 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.

Guidelines:
In our experience, many Big data sources change over time. In the most fortunate case, additional variable(s) become included which is not a major concern. A number of examples that negatively affect Big data based findings are listed below and the ways to deal with them are described.

The first is when the definition of a variable (gradually) changes over time. This results in so-called concept drift; i.e. the original concept measured can no longer be determined exactly [35]. There are other changes that have a similar effect, such as changes in the behaviour of people/companies and changes in the content of the variables used [36]. All in all, methods need to be developed that i) are able to detect such changes and ii) can correct for them. Detection can, for instance, be done by developing indicators for specific quality aspects of the data. Examples of this are indicators for the quality of road sensor data [2] or the findings on a standardized data set [36]. Correction for model degradation has been successfully resolved by updating the model with large amounts of new cases [3,36].

The second example is when a variable becomes no longer available. This, for instance, occurred for the social media sentiment based Consumer Confidence indicator of the Netherlands [20,27] when Facebook messages became no longer available in the dataset collected by a commercial company [P. Daas, personal communication]. Fortunately, this indicator was not yet officially used. The best solution to deal with this issue was to create a fully Twitter based sentiment indicator. This indicator would have to be developed on messages of a much lower quality [27]. In general, being overwhelmed by the loss of important data should be prevented as much as possible. This could, for instance, be done by making arrangements on the stability of content of the data source with the provider of the data [37].

The third example is a complete loss of a Big data source. This is the worst case possible. Here again, it would be good to prevent this from happening by making arrangements on the stability of content of the data source and the delivery dates with the provider of the data [37]. This arrangement should also include a minimal period between the statement of the data provider that they will stop the delivery of data and the actual date on which the delivery ends. The latter provides time to search for alternative data sources and determine their quality. This situation actually occurred for a number of administrative sources [34].

4.6 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 [38]. 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.

Guidelines:
There are a number of cases where the dimensionality of Big data is or can become really high; this is especially the case for text and image containing sources [3,WPH]. Finding ways to reduce this will not only speed up computational time, it will also enable the use of even more data for official statistics, benefit many Big data exploratory studies and may even reduce the complexity of machine learning models [39]. However, applying many of such methods, for instance Principal Component Analysis, results in the loss of the interpretability of the reduced features [40]. This will seriously hamper their application for official statistics (see section 3.3).

4.7 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 [41,42,43]. This abundance of data poses new problems for statistical disclosure control as well as methodological challenges, which need to be addressed [44]. 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 (WP8 Del 8.3, section 2.8).

Guidelines:
Inclusion or the availability of Big data may seriously affect that way data has been protected so far. There are no official statistics examples known, which is a good thing. However, a number of examples in the outside world reveal the importance to prevent this. One is the reidentification of publically released “de-identified” Australian Medicare Benefits Scheme data [45]. A weakness in the encryption technique used enabled it to be reversed. Another is the study performed by researchers at the MIT’s Whitehead Institute were 50 individuals were re-identified that had submitted personal DNA information in genomic studies such as the 1000 Genomes Project [46]. Remarkably, the only used a computer, an Internet connection, and publicly accessible online resources. It is important to update

current best practices in this official statistics area with methods capable to withstand the information in Big data.

5. Conclusions

In this paper, we propose a general process for using Big data as the main data source in official statistics production. This process is composed of two data processing steps, one inference step and one population oriented step. We have meticulously compared this process with those documented for the 13 Big data based statistics examples listed in Table 1. It was found that all examples at least described the first two data processing steps. For three cases all steps have been implemented and estimates for the target population have been produced; i.e. no. 1, 2 and 11 in Table 1. For some cases the results are used as a proxy for the target population; i.e. no 3-5, 12 and 13 in Table 1. Here, linking the Big data units to the target population and correcting for the missing part of the population remain the biggest challenges. Some sources enable an expansion of regular statistics, e.g. no. 6, 8 and 10 in Table 1. All others have not reached these stages yet, but are well on their way to produce Big data based statistics.

References

[1] Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data. Working paper, September.

[2] Puts, M.J.H., Daas, P.J.H., Tennekes, M., de Blois, C. (2019). Using huge amounts of road sensor data for official statistics. *AIMS Mathematics* 4, pp. 12-25.

[3] Daas, P.J.H., van der Doef, S. (2020). Detecting Innovative Companies via their Website. Statistical Journal of IAOS (2020), *in press*. doi/10.3233/SJI-200627.

[4] ISTAT. (2020). Social Mood on Economy Index. Methodological note, Webpage.

[5] Wurian, R., Laimer P. (2020). Urlaubs- und Geschäftsreisen - Kalenderjahr 2019: Ergebnisse aus den vierteljährlichen Befragungen, Snellbericht 3.4, Statistics Austria. (German only).

[6] Nurmi, O., Piela, P. (2019). The Use of Mobile Phone Data in Tourism Statistics. Paper for the 62nd ISI World Statistics Congress, Kuala Lumpur, Malaysia.

[7] Ahas, R., Tiru, M., Saluveer, E., Demunter, C. (2011). Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. Paper for the NTTS 2011, Brussels, Belgium.

[8] De Broe, S., Struijs, P., Daas, P., van Delden, A., Burger, J., van den Brakel, J., ten Bosch, O., Zeelenberg, K, Ypma, W. (2020). Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics. Statistical Journal of IAOS, *accepted for publication*.

[9] Puts, M. (2016). Advanced big data sources - Mobile phone and other sensors, ESTP training course, slide 5.

[10] Willenborg, L. (2018). Transitivity of price indices, CBS discussion paper, (2018), May.

[11] Puts, M., Daas, P. (2020). Applications of big data to official statistics: Social Media. Webinar 3 of ESTP Big Data training course, slides 27-33.

[12] Van Delden, A., Windmeijer, D., ten Bosch, O. (2019). Searching for business websites. Discussion paper, Statistics Netherlands, The Hague, The Netherlands.

[13] Daas, P.J.H., Burger, J., Quan, L., ten Bosch, O., Puts, M. (2016). Profiling of Twitter Users: a big data selectivity study. Discussion paper 201606, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

[14] Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, S., Pappalardo, L., Gabrielli, L. (2015). Small area model-based estimators using big data sources. Journal of Official Statistics 31, pp. 263–281.

[15] Schmid, T., Bruckschen, F., Salvati, N., Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society Series A 178, pp. 239–257.

[16] Noor, A., Angela, V., Gething, P., Tatem, A., Snow, R. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. Population and Health Metrics 6.,p. 5, doi 10.1186/1478-7954-6-5.

[17] Engstrom, R., Hersh, J., Newhouse, D. (2017). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. Technical report.

[18] Blumenstock, J., Cadamuro, G., On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science 350, pp. 1073–1076.

[19] Harvey, A.C., Chung, C. (2000). Estimating the underlying change in unemployment in the UK. Journal of the Royal Statistical Society Series A 163, pp. 303-339.

[20] Van den Brakel, J., Sohler, E., Daas, P., Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Survey Methodology 43 (2), pp. 183-210.

[21] Schiavoni, C., Palm, F., Smeekes, S., van den Brakel, J.A. (2020). A dynamic factor model approach to incorporate big data in state space models for official statistics. Discussion paper January 2019. Statistics Netherlands, Heerlen.

[22] Snijkers, G., Gómes Pérez, J. (2020). Exploring precision farming data: a valuable new source for official statistics? A pilot with System-to-System data communication applied to John Deere data. BigSurv20 presentation, 13 Nov., online.

[23] Christian, L., Ricci, K. (2020). Integrating organic data and designed data for higher quality measurement: Overcoming coverage limitations of big data, BigSurv20 presentation, 6 Nov., online.

[24] Ricci, K, Christian, L. (2020). Is bigger always better? Evaluating measurement error in organic TV tuning data. BigSurv20 presentation, 6 Nov., online.

[25] Griffith, L.E., van den Heuvel, E., Fortier, I., Sohel, N., Hofer, S.M., Payette, H., Wolfson, C., Belleville, S., Kenny, M., Doiron, D., Raina, P. (2015). Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. Journal of Clinical Epidemiology 68(2), pp. 154-162.

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

[27] Daas, P.J.H., Puts, M.J.H. (2014). Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany.

[28] Katz, Y. (2017). Manufacturing an Artificial Intelligence Revolution. SSRN, http://dx.doi.org/10.2139/ssrn.3078224.

[29] Daas, P.J.H., Puts, M.J.H. (2014). Big data as a Source of Statistical Information. The Survey Statistician 69, pp. 22-31.

[30] Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science 343(6176), pp. 1203-1205.

[31] Pearl, J., Mackenzie, D. (2019). The Book of Why: The New Science of Cause and Effect. Penguin, New York.

[32] Engle, R.F., Granger, C.W.J. (1987). Co-integration and error correction: Representation, estimation and testing. Econometrica 55 (2), pp. 251–276. doi:10.2307/1913236.

[33] Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2015). Big Data and Official Statistics. Journal of Official Statistics 31 (2), pp. 249-262.

[34] Daas, P., Arends, J. (2012). Secondary data collection. Statistical methods series 201206, Statistics Netherlands, Heerlen, the Netherlands.

[35] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), pp. 1-37, doi.org/10.1145/2523813.

[36] Daas, P., Jansen, J. (2020). Model degradation in web derived text-based models. Paper for the 3rd International Conference on Advanced Research Methods and Analytics (CARMA), pp 77-84, doi.org/10.4995/CARMA2020.2020.11560.

[37] Struijs, P., Braaksma, B., Daas, P. (2014). Official Statistics and Big Data. Big Data & Society, April–June, pp. 1–6.

[38] Leskovec, J., Rajaraman, A., Ullman, J.D. (2020). Mining of Massive Datasets, third edition. Chapter 11, Dimensionality Reduction. Cambridge University Press, UK.

[39] Marsland, S. (2015). Machine Learning: an Algorithmic Perspective, 2nd ed. Chapman & Hall/Crc Machine Learning & Pattern Recognition, CRC Press, London, Chapter 6: Dimensionality Reduction.

[40] Ma, Y., Zhu, L. (2013). A Review on Dimension Reduction. International Statistical Review 81(1), pp. 134–150, doi:10.1111/j.1751-5823.2012.00182.

[41] Rubinstein, I. (2013). Big Data: The End of Privacy or a New Beginning?, International Data Privacy Law 3(2), pp. 74-87, doi:10.1093/idpl/ips036.

[42] Jain, P., Gvanchandani, M., Khare, N. (2016). Big data privacy: a technological perspective and review. Journal of Big Data 3, 25, doi 10.1186/s40537-016-0059-y.

[43] De Montjoye, Y-A. et al. (2018). On the privacy-conscientious use of mobile phone data. Scientific Data volume 5, 180286, doi:10.1038/sdata.2018.286.

[44] Jensen, M. (2013). Challenges of Privacy Protection in Big Data Analytics. IEEE International Congres on Big Data.

[45] Office of the Australian Office Commissioner (2018). Publication of MBS/PBS data, Commissioner initiated investigation report. Australian Government, Located at: https://www.oaic.gov.au/assets/privacy/privacy-decisions/investigation-reports/publication-of-mbs-pbs-data.pdf.

[46] Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y. (2012). Identifying Personal Genomes by Surname Inference, Science 339(6117), pp. 321-324, doi: 10.1126/science.1229566.