ESSnet Big Data

Specific Grant Agreement No 2 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

Work Package 8

Methodology

Deliverable 8.4

Report describing the methodology of using Big Data for official statistics and the most important questions for future studies

Version 31-05-2018

Prepared by: ESSnet Big Data WP8 members

Anke Consten, Valentin Chavdarov, Piet Daas, Vesna Horvat, Jacek Maślankowski, Sónia Quaresma, Magdalena Six, Tiziana Tuoto

ESSnet co-ordinator:
Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone: +31 45 570 7441
mobile phone: +31 6 5248 7775

Contents

Executive summary 4
1. Introduction 6
1.1. General introduction 6
1.2. Introduction to the report on methodology 7
1.3 References 8
2. Methodological issues 9
2.1. What should our final product look like? 9
2.2.1 Introduction 9
2.2.2 Findings across WPs 10
2.2.3 Discussion 11
2.2.4 References 12
2.2. Data process architecture 12
2.2.1 Introduction 12
2.2.2 Findings across WPs 12
2.2.3 Discussion 13
2.2.4 References 14
2.3. Changes in data sources 14
2.3.1 Introduction 14
2.3.2 Findings across WPs 15
2.3.3 Discussion 15
2.3.4 References 16
2.4. Deal with spatial dimension 16
2.4.1 Introduction 16
2.4.2 Findings across WPs 16
2.4.3 Discussion 17
2.4.4 References 18
2.5. Unit identification problem 18
2.5.1 Introduction 18
2.5.2 Findings across WPs 19
2.5.3 Discussion 20
2.5.4 References 20
2.6. Sampling 21
2.6.1 Introduction 21

2.6.2 Findings across WPs 22
2.6.3 Discussion 23
2.6.4 References 24
2.7. Data linkage 24
2.7.1 Introduction 24
2.7.2 Findings across WPs 25
2.7.3 Discussion 26
2.7.4 References 26
2.8. Secure multi-party computation 27
2.8.1 Introduction 27
2.8.2 Findings across WPs 27
2.8.3 Discussion 27
2.8.4 References 27
2.9. Machine learning in official statistics 28
2.9.1 Introduction 28
2.9.2 Findings across WPs 28
2.9.3 Discussion 29
2.9.4 References 30
2.10 Assessing Accuracy 30
2.10.1 Introduction 30
2.10.2 Findings across WPs 31
2.10.3 Discussion 33
2.10.4 References 34
2.11. Inference 34
2.11.1 Introduction 34
2.11.2 Findings across WPs 34
2.11.3 Discussion 35
2.11.4 References 35
3. Conclusions 37
4. Abbreviations and acronyms 38
5. List of figures and tables 38

Executive summary

This report of Workpackage (WP) 8 focuses on methodology. At the start of the WP the most important topics in this area when developing Big Data based statistics were identified. These topics are: What should our final product look like?, Data process architecture, Changes in data sources, Deal with spatial dimension, Unit identification problem, Sampling, Data linkage, Secure multi-party computation, Machine learning in official statistics, Assessing accuracy and Inference,. This divers set of topics either aims to create the best results achievable from the data available (Deal with spatial dimension, Unit identification problem, Sampling Data linkage, Machine learning in official statistics, Assessing accuracy and Inference,), aim to deal with changes as good as possible (What should our final product look like?, Data process architecture, Changes in data sources) or want to exchange data in the most secure way (Secure multi-party computation). For each of these tasks, methods need to be available or developed to achieve them. The overview of the methodological work performed in each WP of the ESSnet revealed a number of examples belonging to each topic with the exception of multi-party computation. The latter topic has not been touched upon in WP1-7. The examples of methods for the other topics are listed in table 1 to which other relevant applications in other NSI related Big Data research are added. As such, this table provides an overview on familiar and new methodology in this exiting area of statistics.

Data process architecture is not included in table 1 because it lays the foundation of the comparison of the other topics and relates Big Data processes to that of other processes in an NSI. Be aware that this process view is mainly used to aid the reader here. Therefore, to help the reader to get more grips on the methodological topics identified, the topics are related to the various steps in a Big Data based statistical process. For this process, the Generic Statistical Business Process Model (GSBPM) is used. According to this model, a Big Data process is composed of 4 steps: Collect, Process, Analyse and Disseminate. The Collect step is all about obtaining data, the Process step is about processing data and quality checking, the Analyse step is about estimation and the Disseminate step is about producing output. Based on the experiences in WP1-7 of the ESSnet, the topics included in table 1 are assigned to each of these four steps. Some are even assigned to two steps, such as Machine Learning (Process and Analyse) and Changes in data sources (Collect and Process). The assigned topic of each step is indicated in table 1 between brackets. Below each topic, different kind of uses are listed, for which various methods are needed. It can be expected that many of these methods are *new* for official statistics. However, beware that *new* in this context may simply indicate methods *not yet familiar to official statistics* but available (and developed) in other areas of research.

Table 1. Overview of the methodological topics identified and there application in WP1-7 and other Big Data areas. Secure multi-party computation and Data process architecture are not included here. After each topic, in between brackets, the corresponding GSBPM step(s) is assigned. Subsequently, examples of applications are listed, indicating the different methods used in the various WP's and any other relevant Big Data projects.

What should our final product look like? (*Disseminate*)
Especially important in data-driven studies
- Relevant for all WP's (such as WP6, WP7, …)

Changes in data sources (*Collect, Process*)
- Especially important for relative new data sources (such as social media)
- Less relevant for data sources with (high quality) international standards (e.g. CDR, AIS, …)

Deal with spatial dimension (*Process*)
- Used to identify population (WP3, WP7, social media NL)
- Used to derive routs of ships in WP4
- Basis of satellite study in WP7
- Work of FlowMinder

Unit identification problem (Process)
- Identifying which part of the target population is included (WP1, social media NL)
- To distinguish mixed populations (WP3 business's and persons, social media NL)
- Deriving background characteristics of units (WP7 social media, social media NL)

Sampling (*Collect*)
- Draw samples of Big Data in exploratory studies
- Considering Big Data as a non-probability sample (PhD NL)
- To compare Big Data variable values with those in target population (WP1)
- To compare population composition in Big Data and target population (WP5)

Data linkage (*Process*)

Combining at three different levels: Location/area, unit and period
- Location/area Geolocation data, address and buildings (WP3, FlowMinder)
- Unit Companies + URL's (WP2), Companies + job adds.(WP1)
- Period GDP and traffic intensity (WP6), Consumer Conf. + Social media sent. NL)

Machine learning in official statistics (*Process, Analyse*)

Can be applied for processing and for estimation
- Processing (social media NL)
- Estimation (WP1, WP2, WP3, WP6, WP7)

Assessing accuracy (*Analyse*)
- Deal with bias (WP2, WP4, …)
- Deal with variance (WP6, …)

Inference (*Analyse*)

Can be Survey based, BD census like (complete coverage), or partly complete BD population
- Survey based (~WP2, WP7 Social media, Consumer Conf. + Social media sent. NL)
- Census like (WP4, road sensors NL)
- Partly complete (WP1, WP3, WP5, …)

Next, based on all work described in this report and in any of the other reports of WP1-8 of the ESSnet and in any other related Big Data work of NSI's, an overview is made of the general approach followed by an NSI that wants to include Big Data in official statistics. From this, the following step-wise methodology is obtained:

1) Get access to Big Data (BD)
2) Perform an BD exploratory data analysis study (including a privacy assessment)
3) Study the objects (units/events) in BD and check if events need to be converted to units for the foreseen application
4) Compare the coverage of the objects in BD to those of the target population of the NSI
5) Study the variable(s) of interest in BD and compare these with those needed by the NSI (variables may be combined and/or processed here; e.g. creating features)
6) Compare the development over time and/or per area of the variable(s) of interest in BD with similar variables included in any other survey or register based results (if available)
7) Check the performance of various models and/or machine learning based applications on improving the relation described in the previous step
8) Determine the effect of any assumptions, short-cuts made, and/or quality issues and corrections on the comparison described in the previous step (may need to restart at step 3-6)
9) In case of any positive findings, check the reproducibility and stability over time of those results
10) Produce a first (beta-)product

In principle, this is an overview of (data-driven) Big Data methodology. After each step a go/no go discission can be made to proceed to the next step. In step 6 data is compared which, especially for completely new output, may be challenging or even impossible to find. When that is the case, it is suggested to consult experts in the field. The list is a starting point and will undoubtable form the foundation of new and exciting future developments in the area of Big Data research.

1. Introduction

1.1. General introduction

The overall objective of the ESSnet on Big Data is to prepare the ESS for integration of Big Data sources into the production of official statistics. The award criteria mentioned that the project has to focus on running pilot projects exploring the potential of selected Big Data sources for producing or contributing to the production of official statistics. Aim of these pilots is to undertake concrete action in the domain of Big Data and obtain hands-on experience in the use of Big Data for official statistics.

A consortium of 22 partners, consisting of 20 National Statistical Institutes and 2 Statistical Authorities has been formed in September 2015 to meet the objectives of the project. According to the Framework Partnership Agreement (FPA) between the consortium and Eurostat, the project runs between February 2016 and May 2018. To concentrate the work as much as possible on the pilots, the consortium has organised its work around the pilots. More specifically, the consortium has subdivided its work into work packages (WP's). The work packages including their ultimate aims by the end of the project are listed in Table 2.

Table 2: description of the work packages in the ESSnet Big Data Programme

Work Package	Description
WP1 Webscraping / Job Vacancies	This WP wants to demonstrate by concrete estimates which approaches (techniques, methodology etc.) are most suitable to produce statistical estimates in the domain of job vacancies and under which conditions these approaches can be used in the ESS. The intention is to explore a mix of sources including job portals, job adverts on enterprise websites, and job vacancy data from third party sources.
WP2 Webscraping / Enterprice Characteristics	This WP investigates which webscraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises.
WP3 Smart Meters	This WP wants to demonstrate by concrete estimates whether buildings equipped with smart meters (= electricity meters which can be read from a distance and measure electricity consumption at a high frequency) can be used to produce energy statistics but can also be relevant as a supplement for other statistics e.g. census housing statistics, household costs, impact on environment, statistics about energy production.
WP4 AIS Data	The aim of this WP is to investigate whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used 1) to improve the quality and internal comparability of existing statistics and 2) for new statistical products relevant for the ESS.
WP5 Mobile Phone Data	The aim of this WP is to investigate how NSIs may obtain more or less 'stable' and continuous access to the data of mobile phone operators.
WP6 Early Estimates	The aim of this WP is to investigate how a combination of (early available) multiple Big Data sources and existing official statistical data can be used in order to create existing or new early estimates for statistics.
WP7 Multi Domains	The aim of this WP is to investigate how a combination of Big Data sources and

	existing official statistical data can be used to improve current statistics and create new statistics in various statistical domains.
WP8 Methodology	The aim of this work package is laying down a generally applicable foundation in areas such as methodology, quality and IT infrastructure for future use of the selected Big Data sources from the pilots within the European Statistical System.

As can be seen from table 2, the work packages 1 to 7 each deal with one pilot and a concrete output. The aim of WP 8 is to generalise the findings of the pilots in order to relate them to the conditions for future use of Big Data sources within the ESS.

Seven National Statistical Institutes participate in WP8: the national statistical institutes of Austria, Bulgaria, Italy, Poland, Portugal, Slovenia and The Netherlands (Work package leader).

WP8 results in four deliverables, of which this deliverable is the second one:

8.1. Literature overview (delivered on 31 December 2017, but it is living document)

8.2. Report describing the methodology of using Big Data for official statistics (the current deliverable, planned for 31 May 2018)

8.3. Report describing the quality aspects of Big Data (planned for 31 May 2018)

8.4. Report describing the IT-infrastructure used and the accompanying processes developed and skills needed to study or produce Big Data based official statistics (delivered 5 March 2018).

1.2. Introduction to the report on methodology

A good part of statistical methodology is built around using sample surveys. Samples are deemed to be a great invention in social sciences. Probabilistic samples along with administrative data are the main data sources used by NSIs so far. Now Big Data gains importance. The use of Big Data is driven by technological progress. This and the fact that the data in Big Data sources differ from those in survey and administrative data requires the need for the development of Big Data methodology (Daas and Puts, 2014). This is the focus of this report.

By many Big Data researchers, the change in methodology is considered so fundamental that it is sometimes referred to as a paradigm shift (Kuhn 1970); i.e. a radical and fundamental change in the basic concepts and experimental practices of a particular scientific discipline. As is to be expected not everyone agrees, see for instance Couper (2013). It is, however, clear that the rise of Big Data has opened the eyes of many statisticians in ways to improve and enhance many official statistics (Glasson et al., 2013).

The Heerlen workshop in April 2017 defined eleven Big Data methodological issues:

1. What should our final product look like?
2. Data process architecture
3. Changes in data sources
4. Deal with spatial dimension
5. Unit identification problem
6. Sampling
7. Data linkage

8. Secure multi-party computation
9. Machine learning in official statistics
10. Assessing Accuracy
11. Inference

The above methodological aspects differ in terms of scope and complexity. For instance, accuracy assessment addresses almost the whole statistical production process: from collecting data through data processing to data analysis. Almost all stages of statistical production process have something to do with accuracy. Some issues are however more Big Data specific. Dealing with changes in data sources is a good example. Being technology driven, Big Data sources change more rapidly over time (certainly compared to administrative data) and NSIs need to adapt accordingly. If NSIs do not adapt to such changes, variation over time could be attributed not only in variation in the object of interest, but also in the variation in data sources as well. Such a component in overall variation is missing in survey-based statistics.

Literature is abundant with Big Data good practices. However, most of those good practices are one-time-only research work. For instance, a group of expert take a Big Data source, process data and come up with inference about the company's business model gap, or gaps in city management, etc. The researchers do not care about data source stability or the ability to link the data with other data sources in the long run. Production of Big Data based statistics on a regular base is quite different. For this, NSIs -for instance- need to deal with changing data sources and with a lot of restriction on the access to that data. WP5 clearly demonstrated that Big Data sources are not readily available. Despite the fact that no WP has reached a full statistical production cycle yet, there are already many insights on the use of Big Data for official statistics. The accumulated experience obtained so far is used to derive steps on how to embed Big Data in official statistics (see 3. Conclusions).

Looking back we can make an association with sample surveys. Despite being the foundation of nearly all statistical production process nowadays, sample surveys have not been accepted quickly by social sciences. In the beginning doing a survey has been ridiculously compared with a "cartographer to map only one square in ten on his grid."(Ayrton, 2017).

In the remainder of this report, the above mentioned 11 methodological issues are described one by one, followed by conclusions.

1.3 References

Ayrton, R. (2017) Time for a revival? A historical review of the social survey in Great Britain and the United States NSRM Link: http://eprints.ncrm.ac.uk/3999/1/Time%20for%20a%20revival%20-%20A%20historical%20review%20of%20the%20social%20survey%20in%20Great%20Britain%20and%20the%20United%20States.pdf

Couper, M.P. (2013) Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods* 7(3), pp. 145-156.

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

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., Khan, A. (2013) What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Kuhn, T. (1970) *The Structure of Scientific Revolutions*. University of Chicago Press.

2. Methodological issues

Big Data can be used in several ways when producing official statistics. It can be used as the main source of information, as an additional data source to aid estimates primarily based on survey or administrative data or merely to improve the weights in model-based estimation or to impute missing values. In all cases the methods used may differ. As a consequence, each of the issues described below could be interpreted from these different points of view. It was therefore decided to consider the experiences in WPs 1 to 7 of the ESSnet as the starting point for the focus of each section. Hence, the focus is on methodology when using Big Data in practice. The experience gained here, forms the basis of a -to be developed- general applicable Big Data methodology.

2.1. What should our final product look like?

2.2.1 Introduction

Highly detailed data could provide information we are not aware of in advance. It is good to think on the potential products. A review of publications in top economic journals from 1963 to 2011 clearly shows a huge evolution after the mid-1980s. Until the mid-1980s, the majority of papers were mainly theoretical (Einav and Levin, 2014). After that, the majority of papers were mainly empirical. Instead of explaining how economy works by means of theoretical models, authors extensively used statistical data and data from controlled experiments. The more extensive use of data reflects both (an increase in) the availability of more data sources and a tendency toward emphasis of empirical work in economics. Most likely Big Data will boost that tendency further.

Big Data have a huge potential to provide details about economic/social life unobserved until now. Many economists stress that the granularity is the most important feature of Big Data approach. Some of them go as far as making an analogy with the invention of the microscope in the 17th century: “[With a microscope] you can basically look at one organism at a whole new level of detail. And I like that analogy for Big Data as well. That in a lot of ways we’re looking at questions that people have looked at before, but you’re just turning up the microscope. I think that’s a pretty apt description when it comes to consumer spending, labour markets, crowd funding, there are so many examples I can think of where the questions are old but they will need to look at them with this new level of analysis that just, sort of explodes the number of policy implications and things like that you can get from them.” (Taylor et al., 2014).

Granular data provides a unique opportunity to device statistical products, which suggest a new reading of lasting economic questions such as labour market dynamic, education – earning relation, how online market works, price dynamic, consumers’ sentiment, business sentiments, income and assets inequality, etc. In addition to that, a Big Data approach is useful when trust in official statistics is low. The Billion Prices Project (BPP) is a good example. The very motive for the BPP was a deep distrust of Argentines in the inflation statistic published by government. Cavallo and Rigobon (2016) showed that official inflation published by Argentine statistics was about 8 percent while inflation based on online prices was over 20 percent from 2007 to 2011. Online prices based inflation was consistent with households’ inflation expectations and some economists’ calculations.

By 2010 BPP were collecting 5 million prices daily from over 300 retailers from 50 countries. For comparison US Bureau of Labour statistics collected about 80.000 prices on bi-monthly base.

However, by 2010 it becomes clear BPP could not go on relying only on grants. In 2011, A. Cavallo and R. Rigobon established a company PriceStats and produces daily inflation indexes in 20 countries using 15 million online prices from over 900 retailors. The financial industry is a customer of daily price indexes. BPP evolution demonstrates that a Big Data approach in producing statistics could be financially viable without relying on grants and subsidies. Technology opens potentials for Big Data based statistics to successfully compete with official statistics. NSIs are under competitive pressure to build capacity in Big Data approach and provide reliable and cost-effective data to government, researcher, business and households.

2.2.2 Findings across WPs

WP1 deals with job vacancies ads published online. Job ads contain more information than survey based statistics: job position, remuneration, location, requirement to candidates' qualification etc. Having a long list of characteristics, we can produce distributions of job vacancies across job position, remuneration, location, qualification. In addition to that, we could access average time of availability of job ads across same job ads characteristics.

Recently Facebook developed an application for hiring. It is easy for both small and medium businesses and job seekers to post job ads and job alerts. As the application is useful for small businesses, it is reasonable to expect a better coverage of small enterprises' job ads. Scrapping through social media is an immediate challenge not only to have a better coverage on job vacancies but to provide more data on SMEs as well. More data about SMEs are always welcome.

Summing up, online job vacancies being granular and timelier than survey-based statistics open a lot more opportunities for labour market analysis compared with survey based job vacancy statistics.

WP2 assess possibility to produce statistics on company internet activities. Companies tend to be very active on the internet. Activities range from selling products, looking for employees, sharing info about their new products, financing small-scale investments via crowd funding, tweeting, etc. When gather all the company information dispersed in the internet and structure it in nice tables we could gain insight on internet economy, produce timely statistics about changes in ownership structure, outputs, type of economic activity, etc. Moreover, this is much more cost-effective to evaluate companies' internet activity via scrapping through companies' sites and social media than asking companies whether they have a site, sell/buy online, hire online, etc.

When combined with survey based statistics web scrapping could provide a fuller picture about company behaviour and the way internet changes companies' business models.

WP3 assess employment of electricity smart meter's data in the production of electricity consumption. As the penetration of smart meters tend to increase it is quite possible to have comparable electricity consumption statistics in the EU in the foreseeable future.

Linking smart meters reading with business register could provide data on both final and intermediate electricity consumption timely. It enriches analysis of electricity intensity of economy and link electricity intensity with climate changes. Moreover, smart meters statistics could provide useful data about vacant dwelling and the proportion of vacant dwelling to overall dwelling stock. This could help better understanding real estate market and price trend.

When combining households' electricity consumption with natural gas and water consumption NSIs could provide interesting information about housing related expenditures across different income and assets owned groups of households.

WP4 searches the possibility to improve quality of existing statistics via employing AIS (Automatic Identification System) data. AIS provides data about ship position, speed, ship ID, headings etc. These data are very useful for port authorities and coast guard. The data should also be organized in a way to become informative about economic processes. The first step the WP4 team made towards this is making a reference framework of ships in European waters. Substantial deviation from reference framework could be a signal for change of trade pattern. Next step is to link reference framework with other indicators about international trade; ports' activity; indicators about demand for vessels like Baltic Dry Index, etc.

WP5 deals with the possibility to employ mobile phone data for statistical purposes. Mobile phone data could provide useful information about peoples' travelling; density of population, tourist flows and even leisure. Leisure is one of the factors for human wellbeing. Mobile data, like most of data owned by private companies, are proprietary. It is not easy to ensure a long-term access to them and produce statistics on regular base. Protecting the privacy of mobile companies' clients is also important.

WP6 searches the possibility to organize survey-based statistics, administrative data and Big Data in a model to forecast GDP and other summary indicators. GDP is lagging well behind economic processes and a forecast based on early available statistics coming from reliable sources could broaden the analytical tools of both researchers and decision makers. Consolidation of large number of indicators in a so-called composite indicator to measure business cycle has a long history. William Jevons observed in the 19th century that sunspots occurs periodically in 10 to 12 years and causes draught and flooding (Eurostat, 2017). For an agricultural society, sunspots are a good leading indicator about crop and overall economic activity. Nowadays economists developed much more sophisticated composite indicators like Human Development Index. It is interesting whether Big Data could improve quality of composite indicators via capturing more economic/social details that remain unobserved in other sources.

WP7 deals with three domains: assessing every day peoples' satisfaction via social media analysis; combining satellites images of land lots with administrative data about agriculture land and crop and measuring auto traffic between Poland and neighbours. Analysis of posts in the social media analysis provides information about moods of people active in social media. Social media analysis could provide data not only on the mood, but also on what people think about environment, immigrants, inequality, and other issues important for the EU future. Satellites images of agriculture land lots allow tracking crop from seed to final stage. Tracking crop progress may improve forecast about agriculture output and price. Commodity speculators will be happy to have good crop forecast. Traffic measurement by means of sensors could provide data about both tourist and goods' flows between Schengen countries on a much timelier base than surveys.

2.2.3 Discussion

Being technology driven Big Date has the potential to produce enormous quantities of data about economic and social processes. However, there is a risk when one dives (to) deep in the data sets, one might produce quite meaningless relations. This should be avoided.

The main reason to discuss spurious correlations here is to emphasize the need for a theoretical framework while studying Big Data. It is important to have an understanding of what variables are important and how causality runs across variables. Hence, the need for Big Data theory, which not only refers to methods for capturing, processing, and producing statistics from loosely structured data, but also includes theory about the object we accumulate information about. The domain knowledge of economists and social scientists of NSI's should therefore be part of Big Data team. However, new data also provides new opportunities. The development of new products or indicators should not be blocked by the knowledge of experts solely used to working with data produced in traditional ways. Room should be available to experiment in a "trial and error" way, to prevent that ideas are already blocked at the embryonic stage. In all cases, common sense and domain expert checks should be performed to prevent "spurious" results being published. One way of checking this is looking at the stability of these findings over time.

2.2.4 References

Cavallo, A., Rigobon, R., (2016) The Billion Prices Project: Using Online Prices for Measurement and Research, Journal of Economic Perspectives, Volume 30.

Einav, L., Levin, J., (2014) Economics in the age of Big Data, Science 346, Available at: https://web.stanford.edu/~leinav/pubs/Science2014.pdf.

Taylor, L., (2014) Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?, *Big Data & Society*.

Eurostat (2017) Handbook on Cyclical and Composite Indicators. Available at: http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-17-003

2.2. Data process architecture

2.2.1 Introduction

Big Data processes are composed of 4 phases. These are Collect, Process, Analyse and Disseminate (GSBPM 2013). During the Big Data life cycle, the data is checked, transformed, aggregated etc. and -in the end- turned into information. This process is not linear and contains many cycles as findings at the end of the chain may initiate the need to improve some or a number of the previous steps (see below). Big Data can be processed in a batch but also in a streaming way. Each requires different methods.

2.2.2 Findings across WPs

No WP in the ESSnet Big Data has reached the stage that a complete Big Data process has been setup, however, but both WP2 and 4 are almost there (WP2 del 2.2; WP4 del 4.3). In WP5 the complete setup of producing statistics from raw mobile phone data is described and discussed (WP5 Del 5.3). The only example of a complete Big Data based statistical process known to date is the Traffic Intensity statistics process of Statistics Netherlands (Stat. Neth., 2015; Puts et al., 2016). The general steps described below apply to each of these examples.

1) Collect

In this step, it is all about getting access to the data and making sure it can be processed. This can be challenging as many Big Data sources are collected and maintained by private organizations. New for official statisticians is that many Big Data sources are produced by private organizations. This may require new ways of getting access and may cause these data sources to be much more dynamic than what NSI's are currently used to.

2) Process

Step two is all about processing and preparing the data prior to analysis. Quality checking and ultimately quality improvement are the major concerns here. Because of the large amounts of data that need to be checked, visualization based methods are being developed for this (Tennekes et al., 2013). Any insights obtained may result in a need to adjust the previous step by, for instance, adding additional data or including an additional data source.

Fluctuations in the quality of the data, because of the causes mentioned above, may influence the way the data needs to be processed and checked at this stage. This suggests the need for flexible and rapidly applicable methods. In the rapidly developing area of streaming data analysis, new methods are developed to quickly process micro-batches of data in (nearly) real-time (Ellis 2014). These approaches can also be applied to extract insights form large data sets. Data cleaning is also an area where there is also a need for new methods. Checking and cleaning large amounts of data requires the need for fully automated approaches (Puts et al., 2017). Manual editing and checking should not be done as this hugely increases the time needed to process all data.

3) Analyse

Since Big Data is not perfect, modelling is the way to go during analysis (section 2.9). In addition, the increased focus on prediction when using Big Data may require the need for a whole new range of analysis methods not (yet) common to official statistics. Examples of the latter are many of the Machine Learning based approaches (section 2.6). Here, the focus is on predictive accuracy and less on the construction of correct stochastic data models. This represents a considerable paradigm shift in the statistical community. Findings observed at this stage may require the need to adjust some of the earlier steps to improve the quality of the data and its findings. Ultimate goal of this stage is to produce new insights, i.e. extract knowledge.

4) Disseminate

In this stage, output is produced. Usually considerable amounts of data have been used, making it likely that a visualization step is included here. This enables a better understanding of the data at, for instance, a detailed regional level. A whole range of visualizations specific for huge amounts of data has and are currently being developed. Other products are reports and publications. The output of this step can also be used as input for other processes.

2.2.3 Discussion

For Big Data based statistics the need emerges for methods (and an IT-infrastructure) that are able to deal with large amounts of data that fluctuate in quality and composition. Depending on the application and time available, efficiency may be preferred over accuracy here. Since adjustments made in one of the steps earlier in the process may affect the findings in any of the subsequent steps, an iterative approach is usually applied; hence the Big Data processing life cycle (see IT-report). This indicates the need for an evaluation at the end of each step. The main cause of this is the fact that a Big Data process is constructed in a data driven way, i.e. from input to output. This in contrast to the development of traditional statistical processes that tend to follow the reversed order; from output to input. Here, the concept to be published and measured is proposed first, followed by an approach to collect data on it. In both cases, the quality of the final product should be high as always for official statistics.

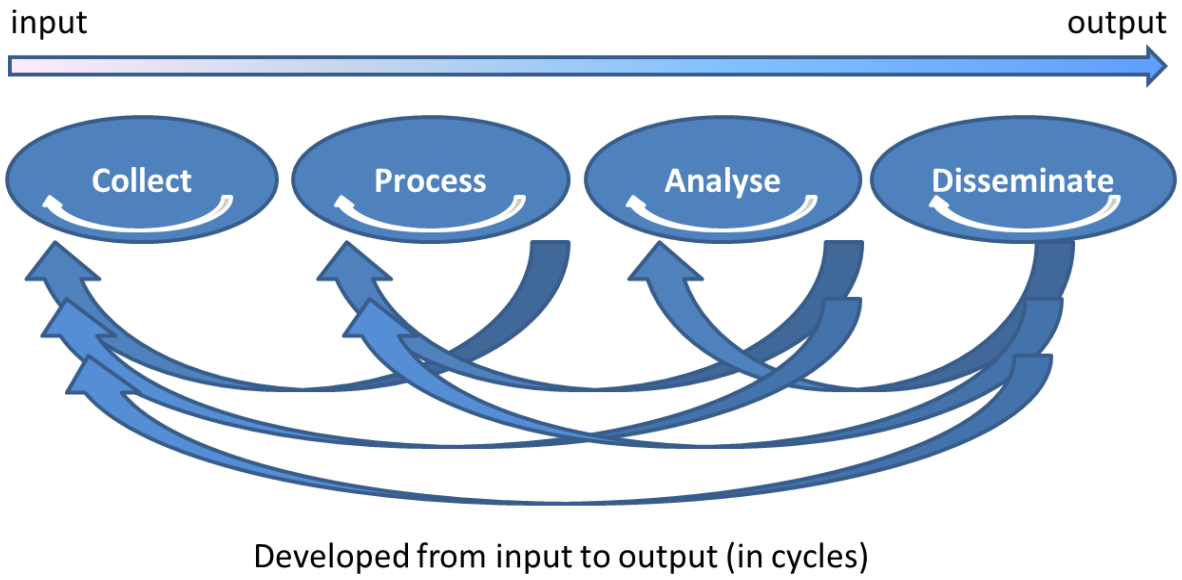

Figure 1: Illustration of the Big Data process life cycle. Arrows indicate the adjustment of previous steps after new findings and quality checks in the subsequent steps. The process starts at the left (input) and ends at the right (output).

2.2.4 References

GSPBM (2013) Generic Statistical Business Processing Model version 5.0. Located at: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

Ellis, B. (2014) *Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data*. Wiley.

Puts, M., Tennekes, M., Daas, P.J.H., de Blois, C. (2016) Using huge amounts of road sensor data for official statistics. Paper for the European Conference on Quality in Official Statistics 2016, Madrid, Spain

Puts, M., Daas, P., de Waal, T. (2017) Finding Errors in Big Data. In: *The Best Writing on Mathematics 2016*, Princeton, USA. (Pitici, M., ed), pp. 291-299, Princeton University Press, USA.

Statistics Netherlands (2015) A first for Statistics Netherlands: launching statistics based on Big Data. Located at: https://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A753C017165F2/0/afirstforlaunchingstatisticsbasedonbigdata.pdf.

Tennekes, M., de Jonge, E., Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science* 11(1), pp. 43-58.

WP2 Del 2.2 (2017) Methodological and IT issues and Solutions. Chapter 3 "Description of a Reference Framework for Web Scraping of Enterprises Web Sites". Located at:
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_3 1.pdf

WP4 Del 4.3 (2018) Report about sea traffic analysis using AIS-data. Located at;
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_2 1_v1.0.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

2.3. Changes in data sources

2.3.1 Introduction

The internet economy is gaining importance because technological progress opens opportunities to supply more internet-based services. The boom in information technology makes it possible to connect a lot of "things" (cell phones, automobiles, vessels and almost everything that come to your mind) to the internet and/or to each other and all these devices generate and exchange data. One can imagine the gigantic volume of data produced by millions of such devices. The Internet of Things

(IoT) will not only change the industry in a profound way by creating opportunities for better management of resources, it will also change the way the economy and society works and how data is being produced and disseminated. The explosion of data may help NSIs to not only speed-up statistics production but also create statistics on new economic and social phenomena.

2.3.2 Findings across WPs

The enormous sources of data can be classification in different ways indicating their potential application or may just indicate an overall population of sources. An UNECE task team on Big Data (2013) suggested the following classification:

- Social networks (human-sourced information): Facebook Twitter, Blogs and comments, Personal documents, Pictures, Videos, Internet searches, etc. These data are loosely structured;
- Traditional Business systems (process-mediated data): these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. These sources produces well-structured information;
- Internet of Things (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured.

The IoT class has two subclasses: data from sensors and data from computer systems.

It is possible to make another Big Data classification for instance one that's based on the type of data included, i.e. numbers, texts, pictures, movies or sound. Other classifications may refer to specific needs. Existing statistical classification like ISIC could be of help for a particular need. In all cases, however, the enormous growth in number of potential data sources and the differentiation of these sources clearly indicates that researchers have to learn how to extract information from all kind of sources of which some have never been used for official statistics. For the use of texts and pictures/movies, it is also clear that a whole range of new methods is needed. Advantage is that these may be borrowed or based on methods from areas of science that already have been studying these types of sources for a considerable time.

In all WP's new data sources were studied (see section 1.1).

2.3.3 Discussion

Devices will become more sophisticated over time, companies' technology and business model change, and in some cases, the substance of information might change. It is good to make difference between changes caused by the increasing sophistication of sources and those affecting the information content. These can have different implications on Big Data methodology. New sources of data will also make the use of Big Data easier. Such sources can improve the quality of statistics by, for instance, improving their accuracy and likability.

Changes in the information content affect Big Data based statistics. Changes could happen for many reasons: changes in business model, changes in policy, changes in the underlying process, etc. Nowadays, millions of devices transfer and store enormous quantity of data. A large part of this information may be of use. For example Facebook, being the third most visited site worldwide is important source of information for social interaction (Blazquez and Domenech, 2018). A

considerable part of the data is new for official statisticians. What kind of information do pictures from nice beaches, posts with lots of abbreviated text and emoticons, etc. contain? Still, many companies and Facebook themselves recognize the ability of such data to determine consumer's profiles and use it for marketing purposes. They are ready to spent resources to dig into the ocean of Facebook data to extract some useful information for their business. In the process of digging, they become capable to pick up relevant data and skip worthless data. Statisticians should adopt this attitude and mentality more. An example of this is the work on social media sentiment performed at Statistics Netherlands (Daas and Puts 2014 and van den Brakel et al., 2017) and that included in WP7 (Del 7.7, 2018). In both cases, indirectly derived information, i.e. sentiment in social media messages, was used to create interesting social statistical information.

2.3.4 References

Blazquez, D., Domenech, J. (2018) Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change* 130, pp. 99-113. Available at: http://dx.doi.org/10.1016/j.techfore.2017.07.027.

UNECE task team BD (2013) Classification of Types of Big Data, Available at: https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data

Daas P.J.H., Puts, M.J.H. (2014) Social Media Sentiment and Consumer Confidence. Paper for the Workshop on using Big Data for Forecasting and Statistics, Frankfurt, Germany.

Van den Brakel, J., Söhler, 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.

WP7 Del. 7.1 (2018) The General Report for each Case Study/Domain including recommendation on legal aspects, availability, sustainability, methodology, quality and technical requirements. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/69/Deliverable_7_7.docx

2.4. Deal with spatial dimension

2.4.1 Introduction

Big Data sources that contain a spatial component are very interesting for official statistics. The data may either directly refer to a specific location on earth, e.g. by geo coordinates, or indirectly, such as the name of a city, a building or an id of a telephone mast. In the first case the location data points to a very specific location while in the second case this points to a particular area, such as the place of a building or the coverage area of a telephone mast. In the second case, additional information, such as the geolocation of a city or building, is needed to assure the correct geolocation of the data in the source. On the internet, various sources that can provide such information are available, for example LatLongnet (2018) to obtain the latitude and longitude of a city. The availability of spatial data may enable the assignment of particular objects or events to specific areas. It may also provide a way in which data in multiple sources can be combined (see section 2.8).

2.4.2 Findings across WPs

WP1 mentions the possibility of relating the demand for specific jobs to specific areas. The location of companies advertising jobs can be used for that. It was found that different on-line sources showed different patterns, underlying the importance of understanding the provenance and processing undertaken on each sources. A specific issue was that some vacancies advertise a choice of locations, creating some conceptual difficulties for how to account for vacancies spatially (WP 1 Del 2.2).

This would enable geographic information about job vacancies to be published. Interesting information could be provided when compare company residence and job vacancy location. They could differ. The topic has, however, not been studied in WP1.

In WP2 geographic information, such as addresses, were collected from enterprise websites. This could be used to update this information in the Business Register.

In WP3 geocoded data has been used to link electricity consumption derived from smart meters to geographic areas. The geolocation of a metering point data needed to be derived via the address associated with it. In Estonia, this was done at a massive scale via the Estonian Land Board's web based Massgeocoding service (WP3 Del 3.1). About 90% of the metering point addresses could be geo-coded automatically.

In WP4 the major component of the data studied is geolocation data. A subset of the Automatic Information System (AIS) messages includes data on the geolocations of a ship (WP4 Del 4.1). However, because these messages are radio signals there are typical ways in which the signal can be disturbed. This causes errors in the geolocations transmitted, for example, resulting in ships in the Sahara. A cleaning procedure was developed enabling the construction of the journeys of ship and their location in harbours (WP4 Del 4.2, Del 4.3).

In WP5 mobile phone data is studied. Here the locations of mobile phone masts provide important geo-information on the mobile phones connected. The area covered by a base station, the overlap between such stations and the accuracy by which a phone is connected to a particular masts are important considerations (WP5 Del 5.3 and more). Such data can, for example, be used to study Day Time Population (Tennekes and Offermans, 2014) or for population density (De Meersmann et al., 2016).

In WP6 road sensors are studied in Slovenia. These have a geolocation assigned indicating their position in the country (WP6 Del 6.1). These have not been linked to other data sources in the WP.

In WP7 several data sources are studied with geolocation information. A part of the social media messages studied contains a geolocation (WP 7 Del 7.1). Other messages may be assigned by the location of the user (location field) or from the message content (because of the object they describe). In this way, topics discussed or sentiment can be assigned to specific areas. When Tourism/border crossings are studied, various sources could be used that provide geolocation data, such as mobile phones, web cams and road sensors (WP7, Del 7.2). Their potential is discussed. For the agricultural studies, satellite data is used that is linked to various registers, for instance for the identify crops grown. Field interviewers (WP7 Del 7.3) checked the ground truth.

2.4.3 Discussion

A number of challenges lay ahead. The first is enriching non-geolocation containing data with high quality geospatial information. For example, data that contain specific locations, such as cities, should be enriched with the geolocation of the object mentioned. This requires the availability of a list of geocodes assigned to those objects. For some this can fairly easily be done, such as cities and soccer stadiums, but other might require more effort. Various publically available data sources that contain both, i.e. a description of the object and its geolocation, can be used for this; such as OpenStreetMap (www.openstreetmap.org). There might be a role for NSI's in this area when standardization is required. Once geolocated data is added to sources, combining them can be done directly when both use the same geocode system and refer to the same geographic areas; see WP3

Del 3.1. If this is not the case, a conversion is required. An example of this is illustrated for mobile phone location data in De Meersmann et al. (2016).

Advantages of the availability of geolocation information, is that a whole range of data sources may be linked to particular areas and that maps can be created in a fairly easy way. The work of Flowminder (2018), an organisation which mission it is to improve public health and welfare in low- and middle-income countries, provides examples of the possibilities geolocations offer. Enriching official statistical publications with geolocation data would enable the tailoring of such statistics to specific geolocation areas. This is a very interesting addition.

2.4.4 References

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F., Reuter, H.I. (2016) Assessing the Quality of Mobile Phone Data as a Source of Statistics. Paper for the European Conference on Quality in Official Statistics (Q2016), Madrid, 31 May-3 June.

Flowminder (2018) Research and innovation. Link: http://www.flowminder.org/work/research-innovation

LatLong.net (2018) Website that provides the latitude and longitude of various objects and addresses. Link: https://www.latlong.net/

Tennekes, M., Offermans, M. (2014) Daytime Population estimations based on Mobile Phone Metadata. Presentation for the Joint Statistical Meeting 2014, Boston, MS, USA.

WP1 Del 2.2 (2018) Final Technical Report (SGA-2). Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_draft_v0.2.docx

WP3 Del 3.1 (2016) Report on data access and data handling. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_1

WP4 Del 4.1 (2016) Creating a database with AIS data for official statistics: possibilities and pitfalls. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/18/WP4_Deliverable_4.1_2016_07_28.pdf

WP4 Del 4.2 (2017) Deriving port visits and linking data from Maritime statistics with AIS-data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/8/8d/WP4_Deliverable_4.2_2017_02_10.pdf

WP4 Del 4.3 (2017) Report about sea traffic analyses using AIS-data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_21_v1.0.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

WP6 Del 6.1 (2016) Potential Big Data and other sources with business cases for the aim of early estimates. Located at:

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/bc/WP6_Deliverable_6.1.pdf

WP7 Del 7.1-7.3 (2017) Multidomains: Report for Population domain, Report for Tourism/Border crossing and Report for Agriculture. Located at:

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/15/WP7_Deliverable_7.1_7.2_7.3_2017_02_01.pdf

2.5. Unit identification problem

2.5.1 Introduction

The data in Big Data sources is the result of events or produced by units (Daas and Puts, 2014). When events are detected, such as checking in or out of a railway station with a chip/smart card, these may need -depending on the foreseen use- to be converted to units first. In the example given, this means deriving the trip of a person, identified by a particular card ID, from the check-in and check-out events at various locations. After this has been done, the task remains of dealing with the units in a Big Data source.

Next is the challenge of identifying the unit in the Big Data source. Downside of Big Data is that the source may not contain much data on the units generating it. This is a pity as getting grip on the

composition of the population of units generating the data is very important for official statistics production. As a result, considerable effort needs to be put into identifying the units in such a source. Be aware that the units might be various kinds of objects, such as a person (WP5), a company (WP2), a job (WP1), a dwelling (WP3) or a vessel (WP4), to name a few. There is also the difference between an administrative unit and a statistical unit that needs to be considered here (Wallgren and Wallgren, 2017).

In general, statistical offices make a clear distinction between administrative and statistical units. Statistical units are the units over which statistics are produced. Administrative units are the units on which data is available in an administrative source. These may not be identical. In particular for companies this is a well-known issue (Wallgren and Wallgren, 2017). A similar situation occurs in a lot of Big Data sources. Obtaining more information on the (observed) unit generating the data is the first thing that needs to be determined. All other considerations, such as linking it to a statistical unit, follow from that.

2.5.2 Findings across WPs

In WP1 the relation between job vacancies and job advertisements is discussed (WP1, Del 1.1). Not every vacancy may result in an online advertisement, some advertisements are placed on multiple web sites, some advertisements may stand for in multiple vacancies and it has even been found that some advertisements are not directly linked to a job. It was found that the data needed to discern between these cases is not always available. It can also be problematic to identify the enterprise unit that advertises the jobs needed. Company name matching is used for that. A study performed in Germany on these topics revealed that the distribution of vacancies compared with advertisements disaggregated by NACE differs, suggesting a sectoral bias in terms of proportion of vacancies having multiple advertisements.

In WP2 considerable effort was put into retrieving the correct website of a company. This is done in an automated way but requires some manual checking (WP2, Del 2.2). The name of the company and address and location data or important input for the approach developed.

In WP3 units in the smart meter dataset refer to so-called metering points. However, how these units relate to an address, a dwelling, a household or a company is not always clear. The electricity consumption data of a metering point was used as a feature to discern between households and companies (WP3 Del 3.2). Some metering points were found to correspond to multiple households or companies.

WP4 discussed the causes of errors in AIS data. Two were related to the units included. The first was the fact that non-maritime ships, such as fishing ships and yachts, also transmit AIS data. The second were caused by invalid, i.e. erroneous, messages. By removing non-maritime ships and non-existing vessels, a population frame was constructed which was used in all subsequent studies (WP4 Del 4.2).

In WP5 the whole conversion chain from raw network data, mobile devices, the individual carrying them and territorial cells is discussed (WP5 Del 5.3). Potential issues are: some individuals may carry multiple phones, some switch them off and some might even let their phone (temporarily) use by others (WP5 Del 5.2).

Since WP6 predominantly focusses on the production of early estimates of official statistics, the focus is less on the units.

The work in WP7 on social media discusses the population active and touches upon the relation between units and social media messages sent (WP7 Del 7.1). Not every user creates messages and some create much more messages then others. In addition, the medium is preferably used by man, by people living in (large) cities and by students. Since not all the information needed to determine this is directly provided by users, considerable work needs to be done to derive information from the people active on social media.

2.5.3 Discussion

Many Big Data sources contain data produced by units. Identifying which units produced the data is challenging. This is caused by several reasons. The most important one is the fact that in many Big Data sources the data available on the units included is limited. By deriving features from the data available or by collecting additional data it may become possible to identify those units are, at least, obtain some background characteristics (Zheng and Casari, 2018). The latter is, for example, often the case in social media (WP7 Del 7.1). In many of the WP's the relation between the unit providing the data and the unit of interest for statistics was an important topic. For example, in WP3 the first is a metering point and the second is a business or a person. Address, geolocation and dwelling data including the electricity profile of the meter (a feature) are used to determine which metering point corresponds to which area and type of unit. A similar problem occurs in WP1 where the unit of measurement are job advertisements and the unit of interest are jobs at companies. By combining data from various job portals and extracting additional data from advertisements, the relation was derived in the best possible way. In WP2 the relation between companies and websites was studied. For companies with an unknown web site, an URL finding approach was developed, that is currently being applied in a number of countries. WP4 had to tackle the unit problem at the highest level possible: by creating a population frame of (real) maritime ships active during the period studied. Without this frame, it was impossible to use the data in any of the subsequent studies. In WP5 there is a whole chain of conversions; from raw network data to mobile devices, from mobile devices to individuals and from individuals to territorial cells.

2.5.4 References

Buelens, B., Daas, P.J.H., Burger, J., Puts, M., van den Brakel, J. (2014) Selectivity of Big Data. *Discussion paper* 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

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.

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

Wallgren, A., Wallgren, B. (2017) *Register-based Statistics: Statistical methods for Administrative Data*. Wiley.

WP1 Del 1.1 (2016) Inventory and qualitative assessment of job portals. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/49/Deliverable_1_1_final.docx

WP2 Del 2.1 (2017) Methodological and IT Issues and Solutions. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf

WP3 Del 3.2 (2017) Report on production of statistics: methodology. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_2

WP4 Del 4.2 (2017) Deriving port visits and linking data from Maritime statistics with AIS-data. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/8/8d/WP4_Deliverable_4.2_2017_02_10.pdf

WP5 Del 5.2 (2017) Guidelines for the access to mobile phone data within the ESS. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/ac/WP5_Deliverable_5.2.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

WP7 Del 7.1 (2017) Multi domains report: Report on Population domain area containing basic information on the data access, quality issues, methodology (focus on combining data) and technical aspects. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/15/WP7_Deliverable_7.1_7.2_7.3_2017_02_01.pdf

Zhang, A., Casari, A. (2018) *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly, USA.

2.6. Sampling

2.6.1 Introduction

Probabilistic samples are deemed to be a great invention in social science in the beginning of 20 century. Many important statistics as households' income and expenditures, consumers' price index, poverty are based on sample surveys. These statistics are important inputs for national account. For example, households' budget survey lays foundation for assessing private consumption and CPI is an important component of overall GDP deflator. However, decreasing responses and budget cuts suggest that the golden age of sample surveys is declining. Nowadays in many countries respond rates tend to decrease, respondents skip questions, and accuracy of responses get lower. "Sample error" is one component of total survey error. Errors related to refuse to respond, skip some of the questions, or respond in a biased way make normal distribution assumption and significance tests highly suspicious.

To cope with lower quality of survey and hardening budget constraints, NSIs resort to other sources of data: administrative data and Big Data. There is a huge difference between sample surveys and Big Data in terms of concepts and instruments. Statisticians draw samples from a population frame. The sample should be representative for the population of interest. In Big Data in contrast, we have to deal with the data that is available; produced by the included Big Data population. For example, consumer's sentiments assessment is based on analysis of messages of the population active in social media and job vacancies assessment deals with online job ads. The important question when using Big Data is whether the available Big Data population is a good proxy for the target population of the NSI and whether what is detected is measured properly. Sampling may play a role here.

However, it is important to make clear that using sampling and the word 'representativity or representativeness' in the context of Big Data is an extremely slippery business. In the methodological document of WP5 (Del 5.3) a whole section is devoted to this topic, named 'Sampling design methodology and the curse of representativity' (pp. 58-64). We will not repeat the whole text here but refrain to the conclusion:

> *"In conclusion, sampling design-based inference is a robust methodology providing firm scientific grounds for the production of official statistics but it is not a panacea for all potential situations we face when producing these statistics. An abuse or misuse of the notion of* representativity *should not be resorted to as an argument to defend this methodology against other alternatives. We believe that the key idea for a high-level quality estimation is not only to use low mean square error estimators, but also to show their robustness against misspecifications of any factor of variability (either the sampling designs or the underlying statistical models or whatever)."*

It should be clear that for Big Data care must be taken when considering sampling based approaches as a means to extract information from this source (see section 2.11). Combining Big Data with that of a survey sample is a constructive way to make advantage of the strengths of both sources (see also section 2.7).

2.6.2 Findings across WPs

WP1 searches for possibility for producing job vacancy statistics based on searching through job portals. WP1 deals with the population of companies that publish job ads online. Do these companies represent the whole business population? A survey in Germany suggests that big companies employ online job ads much more often than small and medium (WP1, Del. 1.2). We can reasonably assume online job vacancies survey under covers SMEs and covers big companies well. One way to cope with SME under coverage is to make a job vacancy survey on SME. Combining online and sample survey could improve job vacancy statistics. Another option is to use online job vacancies statistics as they are since the big companies are included.

Internet technology develops and it is reasonable to expect more SMEs to hire online. Recently Facebook developed an application to facilitate companies to hire and job seekers. Within Facebook companies can post job ads and job seekers could post job alerts for positions they are interested in. Companies could track application and communicate directly with the applicants. The application is useful for small companies. Web scrapping should spread to social media to grasp job ads posted in the newly developed application.

WP2 aims at finding URL addresses of enterprises via web scrapping techniques. The population subject to web scrapping is all companies that have a web site. A URL identification method may be more efficient for companies that have an e-commerce activity. Possibly, URL identification is biased to sites that provide online trade as well. There are at least two different ways to deal with this potential bias component. A first one is, in a full model based approach, to adopt an estimator that makes use of weights obtained by pseudo-calibrating units for which it is possible to scrap their websites with known totals pertaining to the whole population of interest. Another possible solution is, whenever also survey data are available, to adopt a combined estimator that makes use of predicted values for units belonging to the sub-population of units with successfully scraped websites, and of observed values for the sub-population of units with unassessed websites; both components of the estimator can make use, as in the previous case, of weights obtained by calibrating with respect to the two sub-populations.

WP3 is a pilot study to produce data about electricity consumption of buildings equipped with smart meters. The population of households and companies is limited to those living in buildings with smart meters installed. Given the prospects that more smart meters will be installed in the future, it can be stated that the data will become less selective over time. Sampling might play a role here, as not all information needed is covered by smart meters readings. The electricity produced and consumed by households/companies is not included. Capturing the latter is an important challenge as own production tends to increase and smart meters do not measure it.

The aim of WP4 is to employ data from Automated Identification System about vessels' geographical position, type, heading, etc. and thus to enrich maritime statistics. AIS system tracks all ships and data covers the whole population of ships in the European waters. Differences in coverage are observed between various sources of AIS data. Apart from combining the data in all sources of AIS data available, sampling could be a way to obtain more information on this issue.

WP5 searches for possibility to use mobile phone data for statistical purposes. The population are the clients of mobile services provider who granted the NSI access to data. The population included in the data of a single mobile operator depends, among others, on its market share, the proportion between natural/corporate persons and the people that provided access to data. There may also be an issue caused by differences between people that have a mobile phone subscription and those that use a prepaid phone, as information of the latter group may be less well known. Studying small samples of data might be a way to obtain insight on the effects of differences in inclusion of these groups. For correction, other approaches are suggested (WP5 Del 5.3).

WP6 investigates how to combine early available survey statistics and Big Data from different sources to produce early estimates about economic performance. WP6 deals with a model that organizes many individual data rather than with producing data. Sampling addresses, the data inputs of the model and is not an issue for the model itself.

WP7 investigates how to combine Big Data, administrative data and official statistics to design new statistical products and improve current statistics. Three domains of interest are subject to detailed investigation: sentiment of people revealed in social media, agriculture and tourism/border traffic. Sampling issues differ across the three domains in WP7. Sentiment assessment is based on analyses of posts in social media. The important question is whether sentiments of the people active in social media represent the sentiments of society as a whole? When people share an opinion, they may shape their sentiment as well. Sentiments that are formed in the social media are not kept within the media but spread across people outside media. Summing up social media is a powerful instrument not only to share opinions, but to shape opinions as well. 5 Stars Movement, the most voted party in election in Italy on March 4 came to existence via intensive employment of social media to reach potential supporters and gain importance over time. This demonstrates how important social media can be.

Agriculture domain combines satellite images of land and administrative data about farmers and land. Both sources of data cover the whole land. Sampling was used to check satellite image based findings for particular areas; i.e. determine the ground truth.

Tourism/border crossing domain measures intensity of border traffic between Poland and its Schengen neighbours. Data came from different sources, such as: traffic sensors, survey based statistics on tourism and social media. Different data sources refer to differences in population coverage. Survey based data are derived from random samples of the target population, sensors reading depend on their location and the way in which they are able to measure all border crossings and social media data is produced by the part of the population active on it. It is interesting to see whether the combination of surveys and Big Data could improve the quality of tourism statistics.

2.6.3 Discussion

Sampling is a great way to quickly gain insight in the content and quality of a Big Data source during an exploratory study. However, this is merely a first step in studies that aim to determine if a particular Big Data source can be used for official statistics. In the subsequent steps, the way sampling is looked upon differs; certainly across WPs 1 to 7. In the case of traffic loop data, AIS data and agricultural land satellite images, the measurements cover the whole population of interest and sampling is not an issue to worry about. Any errors in the sources are predominantly the result of issues with the instruments that collect the data and are not the result of a (large) part of the population missing.

For smart meters and online job vacancies, it is clear that they do not include the whole target population of the NSI. Here, sampling may play a role in assessing and correcting for the remaining part of the population. Assessment of sentiments revealed in social media imposes the question whether the sentiment of people active on social media represents the sentiments of all people in a particular country. It is hard to say yes or no. We need time to evaluate whether sentiments revealed in social media are a good proxy for sentiments of the whole society. However, a Statistics Netherlands study suggests that this is the case and that sentiment can be used (Daas and Puts 2014). The subsequent study of Van den Brakel et al. (2017) revealed that the sentiment in public Facebook and Twitter messages improved survey-based consumer confidence estimates but that the sentiment in these messages is not exactly identical to consumer confidence. The overall sentiment is more positive compared to consumer confidence (Daas and Puts 2014).

2.6.4 References

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

Van den Brakel, J., Söhler, 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.

WP1 (2016) Del 1.2 Interim Technical Report. Located at:
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

2.7. Data linkage

2.7.1 Introduction

Combining multiple sources of data is becoming increasingly common in the production of official statistics, as it presents great opportunities for generating statistics of greater scope and richer details, with less burden on the respondents (Di Zio et al., 2017). For survey and administrative data, this can be fairly easily achieved when administrative data is used as the sampling frame for the survey. However, not every NSI has the availability of administrative data sources of high quality and not every country uses a unique identification system across these sources (UNECE 2007). However, even when unique identifiers are absent, probability based methods can be used to combine administrative and survey data provided auxiliary variables of sufficient quality are present (Lohr and Raghunathan 2017). Combining Big Data with other data sources is the next challenge.

The first observation for Big Data is that -in the case of linking- there are several differences compared to the common linking activities applied by NSIs. When linking Big Data with survey data and/or administrative data, the scenario most similar to the approach commonly used by NSIs is when a unique ID is available. This is the case, for instance, when data on enterprises are scraped from the web; when a Chamber of Commerce number is available on the website. However, in some Big data sources, sometimes hardly any information is (directly) available on the units included. In such cases, the metadata available on the units producing the data is often very limited; if present at all (Daas et al., 2016). This prevents the direct linking of the units included with those in other sources. When that is the case, other ways of combining sources need to be considered. These are discussed below.

2.7.2 Findings across WPs

Currently three ways of combining Big Data with other sources are described. These are: i) linking at the location/area level (Marchetti et al., 2015), ii) linking at the unit level (Daas et al., 2016) and iii) linking at the time series level (Van den Brakel et al., 2017).

i. Linking locations and areas

A considerable number of Big Data sources contain geolocation data (see also section 2.4). This opens the possibility to combine such data sources on the locations and/or areas included. In WP3 (Del 3.2) this has been done to obtain estimates of electricity use for businesses (Estonia) or households (Denmark). In both cases, address information was used to combine smart meter data to geolocations. This also enabled aggregating the data to larger geographical areas (more in section 2.4).
WP4 investigated whether it was possible to link AIS data from Dirkzwager to the survey data on maritime statistics from Poland based on the coordinates. It is important to realize that the survey data from Poland does not provide MMSI numbers, only IMO numbers. Therefore, a reference frame of ships for Poland was linked with data containing coordinates of ports. After that, the combined dataset was split into two groups: “ships in ports” and “ships not in ports”. This work verified that “ships in ports” based on the coordinates were indeed in Poland’s dataset from survey data. Therefore, it was possible to link European AIS data with survey data.

The study of Marchetti et al (2015) in which Big Data was combined with more traditional data to measure social well-being in particular areas in the Tuscan region in Italy is another successful example of linking data at the area level. For some mobile phone data applications, such as the Day Time Population (Tennekes and Offermans, 2014) or for population density (De Meersmann et al., 2016), geolocation data is also used and combined. This is also the case for WP5 (Del 5.3).
The company Flowminder (2018) has other examples of combing all kinds of data sources at the location and area level available on their website.

ii. Linking units

Linking the units in a Big Data source with those in other more traditional sources is the most common way of combining data. However, this is not always trivial. For instance in WP2, were company web sites are studied, in those cases were the official web site is unknown, the website is searched for by applying URL finder methods (WP2, Del 2.2). Considerable progress has been made in this area during the ESSnet Big Data. Here, other data available, such as the name and address, are used to find the URL of the website for the most similar company on-line. However, identifying units in other areas may be more challenging when direct identifiers are lacking (see also section 2.11). A way to solve this is by deriving so-called features (see 2.5) from the data that is available for those units; this is illustrated for gender in Daas et al. (2016). When unique identifiers are lacking, probabilistic linking approaches need to be applied (Di Zio, 2017) using as much information as is available.

iii. Linking time series

When the data in a Big Data source enables one to produce a time series of a particular variable or a combination of variables, such a series can be compared with one or more official statistics. When the Big Data series demonstrates a development similar to that of an official statistics, it may be

worthwhile to investigate this 'association' further. For instance, in a model-based approach. In WP1 time series analysis has been applied to on-line job vacancy data to produce flash estimates of job vacancies (WP1 Del 2.2). In WP6 the relation between traffic intensity and GDP in Slovenia was studied (WP 6 Del 6.1). It might even be possible to create a model in which the information provided by both series is combined. Van den Brakel et al. (2017) developed such an approach for the official Dutch Consumer Confidence Index and social media message sentiment in the Netherlands. Here, it was demonstrated that the monthly aggregated sentiment in public Facebook and Twitter messages improved the precision of the survey based Consumer Confidence estimates. Import things to consider when comparing time-series are correlation, cointegration, long-run stability and any causal relations. The reader is referred to an econometrics book, such as that of Wooldridge, (2102), for an introduction to these topics.

2.7.3 Discussion

Combining Big Data with other sources is possible but challenging because of the different structure of many Big Data sourced compared to the more traditional survey and administrative sources. Major concern here is the presence, or better the absence, of data that can be used to directly identify units in Big Data sources. Features can be derived to provide some insights (see section 2.5) which opens up the possibility of probabilistic linking approaches. On the other hand, alternative ways of combining Big Data with other sources are available. The two exampled mentioned are geolocation and time-series based. In the first approach, geolocation data is used to link data to specific areas or addresses. In the second case, aggregates over specific time period and relates the series derived from the sources in a model based approach. Both alternatives demonstrate how Big Data can be successfully included in official statistics.

2.7.4 References

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.

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F., Reuter, H.I. (2016) Assessing the Quality of Mobile Phone Data as a Source of Statistics. Paper for the European Conference on Quality in Official Statistics (Q2016), Madrid, 31 May-3 June.

Di Zio, M., Zhang, L-C., de Waal, A.G (2017) Statistical methods for combining multiple sources of administrative and survey data. The Survey Statistician 76, pp.17-26.

FlowMinder (2018) What we do: Examples of our work. Available at: http://www.flowminder.org/

Lohr, S.L., Raghunathan, T.E. (2017) Combining Survey Data with Other Data Sources. Statistical Science 32 (2), pp. 293–312

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Gianotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., Gabrielli, L. (2015) Small Area Model-Based Estimators Using Big Data Sources. Journal of Official Statistics 31 (2), pp. 263-281.

Tennekes, M., Offermans, M. (2014) Daytime Population estimations based on Mobile Phone Metadata. Presentation for the Joint Statistical Meeting 2014, Boston, MS, USA.

UNECE (2007) Register-based Statistics in the Nordic Countries - Review of Best Practices with Focus on Population and Social Statistics. Located at: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

Van den Brakel, J., Söhler, 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.

Wooldridge, J.M. (2012) Introduction to Econometrics: A Modern Approach. Available at; http://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf

WP1 Del 2.2 (2018) Final Technical Report (SGA-2). Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_draft_v0.2.docx

WP2 Del 2.2 (2017) Methodological and IT Issues and Solutions. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf

WP3 Del 3.2 (2017) Report on production of statistics: methodology. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_2

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

WP6 Del 6.1 (2016) Potential of Big Data and other sources with business cases for the aim of early estimates. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/bc/WP6_Deliverable_6.1.pdf

2.8. Secure multi-party computation

2.8.1 Introduction

When the data available at two or more organizations needs to be combined in a privacy preserving manner without letting one of the partners having access to *all* the data, this requires a level of security provided by, for instance, secure multiparty computation (SMC). This technique is a form of cryptography and enables combining data in a fully secure manner without each of the partners involved having to reveal the microdata to one another. The result of this process is usually data in an aggregated form to prevent the de-identification of (some of) the data of each of the units involved. An example of applying SMC is combining health data of persons collected by one institute, such as a hospital, with socioeconomic data of those people provided by the National Statistical Institute. The outcome of this process is, for instance, a table in which the relation between health and living conditions is shown (at a certain aggregated level).

2.8.2 Findings across WPs

In the ESSnet Big Data no workpackage has applied SMC. However, at certain NSI's and in Eurostat its potential is investigated. The first practical implementation of SMC was in 2008 when 1200 Danish farmers used it to determine the market price of sugar beets contracts without having to reveal their (sensitive) selling and buying prices and without resorting to an externally trusted party (Bogetoft et al., 2008).

2.8.3 Discussion

In the ESSnet Big Data no workpackage has used SMC. However, several institutes and Eurostat are looking into the potential of SMC methods as it increases the re-use of data in a FAIR manor. Here, FAIR stands for Findable, Accessible, Interoperable, and Re-usable data use (Force 11, 2018). By sharing data, one contributes to the advancement of science and to knowledge discovery. Because of the vast amounts of data available at NSI's that cover long time periods, NSI-data are of considerable interest to others. SMC is a very important way of preserving the privacy of such data. If NSI want to apply this method, it is important to inform the public about it and explain it benefits. In 2017, Statistics Netherlands, the University of Maastricht and two Dutch research institutes started a project in which "the possibilities of applying SMC methods to enable the secure exchange of data" is studied (MU, 2017). The first results are expected to be available at the end of 2018.

2.8.4 References

Bogetoft, P., Christensen, D.L., Damgård, I., Geisler, M., Jakobsen, T., Krøigaard, M., Nielsen, J.D., Nielsen, J.B., Nielse, K., Pagter, J., Schwartzbach, M., Toft, T. (2008) Multiparty Computation Goes Live. Cryptology ePrint Archive. Located at: https://eprint.iacr.org/2008/068

Force 11 website (2018) The FAIR Data Principles. Located at: https://www.force11.org/group/fairgroup/fairprinciples.
MU (2018) Institute of Data Science lands two major research projects on FAIR data. Web article, Located at: https://www.maastrichtuniversity.nl/news/institute-data-science-lands-two-major-research-projects-fair-data.

2.9. Machine learning in official statistics

2.9.1 Introduction

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed (Samuel 1959). An ever-growing part of the models uses machine learning algorithms for data manipulation, data sources integration and forecasting/nowcasting estimation. With the advent of Big Data and smart sensors, i.e. for data-driven applications, machine learning methods see even more use as they facilitate the import of big volumes of information into statistical models. Depending on someone's educational background and years of experience of working in official statistics, there is a difference in what methods are considered as machine learning. Linear regression is an example of a method of which many statisticians agree that it is not machine learning. However, in the general discussion below all of these methods are included even if they are usually not referred to as machine learning.

Often the use of a model is impeded by the shape of data: missing data may invalidate the results of the perfect model. In such cases, machine learning algorithms can also be used for assessment of the missing data (imputations). Other reasons to use models are to compensate for data of low quality or for differences in the population composition of the Big Data source and the target population.

Two different types of machine learning algorithms exist. The first type takes a set of data and tries to find clusters of similar observations. These kinds of algorithms are called unsupervised learning algorithms (we cannot easily check if the groups are right). New observations can then be classified according to the algorithm into a class that best represents the observation's features. On the other hand supervised learning algorithms (easily checked by comparing the target variable and its estimation) exist that produce a model for target and independent variables. In practice, a supervised learning algorithm takes a set of observations with known-valued target variables and different logical and/or mathematical models are used on the variables until the algorithm finds the one that best describes the relationships between the target and predictors. Upon feeding the algorithm with new information, it is possible to calculate the estimations of the observed variables that best fit the fed unobserved variables according to the chosen model.

To reduce under fitting (high bias, low variance) and overfitting (low bias, high variance) due to extreme matching of the algorithm models to the learning set, different resampling techniques are used, such as bootstrapping or cross validation, in order to obtain a model that tends to give more general results.

2.9.2 Findings across WPs

WP1: machine learning is used for text analysis. Python's gensim model was used for text analysis and deduplication and similarity functions to determine professions. Classifying occupation and industry sector were studied with Python ScKit learn and some studies made use of Rtexttools.

WP2: enterprise website link retrieval was executed with use of logistic models, random forests and neural nets fitted according to known links. Some use Naïve Bayes classifiers on bag-of-words data for features classification. There are two goals of using machine learning. The one is to determine whether the enterprise uses e-commerce on their website. The second is to determine the type of social media presence of enterprise (e.g., marketing, commercial, enterprise image, etc. based on classification from ICT in enterprises survey).
WP3: ARIMA models are used for modelling smart meter data and random forest for estimation/forecasting of vacant living spaces. Cell-wise outlier detection methods were used for anomaly detection.
WP4: work is focused on using ships' AIS data as a source which they do not use in conjunction with any machine learning methods.
WP5: In this WP, machine learning techniques were not found appropriate for the task at hand. Therefore, other alternatives to traditional estimation methods were investigated; e.g Bayesian inference methods (see WP 5, Del 5.3 for more details).
WP6: machine learning is extensively used to integrate different sources of data or to quickly calculate estimates on large data in a timely fashion. A variety of methods are used, from the simplest linear regressions, through different optimization methods such as principal component analysis (PCA, a dimensionality reduction method), to ensemble methods. Every method in the R package Caret was used for testing purposes.
WP7: for satellite image classification K-means methods, decision trees, K-Nearest Neighbours and SVM were used while for population happiness a range of classification methods for text analysis were used. Naïve Bayes is used to determine the sentiment for Population use case (e.g., happy, sad, calm, angry, discouraged, and depressed).

2.9.3 Discussion

From the above examples, it is clear that in the various WP's of the ESSnet machine learning algorithms were used. In fact, they were used at two different steps in the GSBPM model: at the process and at the analyses phase. In the former, they were predominantly used as a form of pre-processing and to a lesser extent for quality checking, while in the latter stage they were predominantly used for estimation and classification. In this context, it is important to note that machine learning algorithms don't always return perfect results. One of the assumptions when using these algorithms is that the observations are randomly distributed. As such the larger the amount of data fed to an algorithm in the learning phase, the better (and more generalizable) are the results. This works in the favour of Big Data sources but also means, that before they can be used a large gathering of data needs to be implemented. Furthermore, the relationships between target variables and predictors are not always easily identified which hinders interpretations of models. Also, from a technical prospective, machine learning methods use a lot of resources like computer memory and execution time. All of this means that before using machine learning much consideration must go towards the assessment of their usefulness compared to their drawbacks.

An important point of discussion is what methods are actually included in the group of machine learning approaches. People from different backgrounds and of different age, have different views on what they consider typical machine learning methods. We had some discussions on the matter and following the reasoning in Statistical Modelling: The Two Cultures (Breiman L., 2001). We therefore have decided to not include regressions into the realm of machine learning. They do

however make part of statistical learning (Hastie et al., 2017) and are as such also relevant in the application of Big Data.

2.9.4 References

Breiman, L. (2001) Statistical Modelling: The Two Cultures, Statistical Science, Vol. 16, No. 3, 199-231. Located at: https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726 (on 3. 30. 2018)

Hastie, T., Tibshirani, R., Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Wiley. Available at: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Samuel, A. (1959) Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3 (3). doi:10.1147/rd.33.0210

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

2.10 Assessing Accuracy

2.10.1 Introduction

Accuracy is an aspect of the overall quality of a statistical product. The accuracy assessment should be analysed in a 'total error' like approach (Weisberg 2015). Zhang (2012) describes a so-called two-phase life cycle model for administrative data which can also be applied to Big Data. In working with Big Data sources, there might be accuracy issues due to:

- Absence of part of the target population in Big Data, i.e. coverage related (e.g. people active on social media, job advertisement for small and medium size enterprises);
- Difference in target and Big Data population composition, often described as 'lack of representativeness' of Big Data (e.g. job advertisements vs job vacancies, mobile phone users vs resident/present population, smart meters vs households);
- Errors in data (e.g. measurement error, validity errors like (temporary) failure of a sensor);
- Differences in definitions (e.g. variable measured doesn't have the same definition as the variable we want to estimate);
- Different sources of variation (e.g. internet prices for the same item vary from site to site, structural variations of job portals due to portal updates).

When any of the above-mentioned issues are identified as affecting the results coming from Big Data, some indicators should be evaluated in order to assess accuracy. For this topic, the reader is referred to the Deliverable 8.2 "Report describing the quality aspects identified in studies focussing on the use of Big Data for official statistics". Deliverable 8.2 provides measures and indicators related to lack of representativeness, measurement errors, processing errors, as they were dealt with in the studies and applications carried out in the WPs of the ESSnet on Big Data. When considering the statistical output obtained by Big Data sources, one should try to assess both the bias and the variability contribution.

Big Data can be used in various ways: as the main source or one of many sources; next to survey of administrative data; in a model to estimate a specific indicator. Therefore, the accuracy issue is often associated with accuracy of the model. For example, ME (mean error) indicates if the model is biased or not; MSE (mean square error) or RMSE (root mean square error), MAE (mean absolute

error), MaxAE (maximum absolute error), (adjusted) R squared give information on variation of the model.

2.10.2 Findings across WPs

The aim of WP1 is to demonstrate which approaches are most suitable to produce statistical estimates in the domain of job vacancies. In on-line job advertisements, there is over-coverage (some advertised vacancies are out of scope for purposes of official statistics) and under-coverage (not all job vacancies are advertised on-line), there can be duplicates (e.g. the same vacancy is advertised in two different job portals), some data can be missing (e.g. closing date) and there can be errors because of misclassification of job advertisements. Job vacancy survey data often differ a lot from on-line job vacancy data. In WP1 the issues affecting accuracy and selectivity for on-line job vacancy data were well understood but no solution has been identified so far. An internal report discussed many quality issues (WP1 2017).

A general approach that was explored for assessing selectivity (particularly by the UK) is the idea of linking reporting units in the job vacancy survey to the company names in the on-line data. In principal, if one can understand the differences between the survey at the level of the individual enterprise, then one should be able to better understand the biases in the on-line data. However, this linking process is itself difficult and prone to error. This is discussed in section 4.2.3 of the abovementioned internal document and, in a general framework, in section 2.8 of this report (WP 2017). Furthermore, even for the subset of reporting units where it is possible to achieve a good match, the patterns between the survey and the on-line data for individual enterprises are often not consistent. Some examples illustrating this problem are provided by WP1: the time series pattern for the survey data is often very different from on-line data and different sources of on-line data may vary between themselves. Basically, WP1 does not yet have a method for measuring selectivity due to the fact that there are many confounding factors with on-line job vacancy data.

WP2 performed web scraping experiments with the aim of deriving experimental statistics on enterprises from information found on the web, especially the websites of enterprises. There were four use cases, referring to URL Retrieval, E-commerce, Job vacancies ads on enterprises' websites and Social Media Presence. The assessment of accuracy can be carried out in a different way distinguishing the first use case (URL retrieval) from the others. For URLs retrieval, accuracy can be measured at the unit level in the subset of cases for which the URL is already available from other sources (survey or administrative) and can be evaluated against the retrieved ones: in this case, the usual accuracy indicators (precision, recall) can be calculated. In the other three use cases, where the final aim is to produce experimental statistics, together with the usual accuracy indicators at unit level (calculated in the subset of units responding to the ICT survey), also an accuracy indicator at aggregate level (Mean Square Error; MSE) can be produced. The MSE has two components: bias and variance. Variance can be calculated by using resampling methods, while for the bias one possible method is based on simulation, where the sources for bias (accuracy at unit level, and different values of the target parameter in the covered and non-covered sub-populations) can be taken under control.

WP3 aims to demonstrate the usage of electricity meters for energy statistics. There are accuracy issues due to data linkage: measured units don't correspond to statistical units, so address is used for linkage, under the assumptions that the address is usually correct (actually there are also problems

with quality) and that an entity is actually located there. However, results were not always satisfactory, as the address quality was not always good. Actually, it was possible to match only 31% of businesses and 44% of households by using address in Estonia. The quality of linking in Denmark was better which was mainly due to better address quality. In the case of businesses, WP3 measured the accuracy of the results by comparing it with survey data but there was not data available for households. Another issue related to accuracy in WP3 is the lack of data for self-production for companies that produce electricity by themselves (only the net flow is measured). Absolute differences and relative differences were used when comparing survey and smart meter data.

WP4 investigates quality aspects of AIS data in deliverable 4.3. In terms of target population, the problem was encountered that AIS data contains information on all ships possessing AIS (in European Waters). The focus of WP4 was solely on maritime ships, so they needed to filter out other types of ships. This was done by using the unique identification number maritime ships have (MMSI) to construct a reference frame of maritime ships. This frame was applied to further analyses.

WP4 checked the quality of different sources of AIS data: European AIS data, national AIS data from Denmark, Greece and Poland and satellite data. The different sources do not have the same coverage. National data usually contain data on more ships and more data points per ship compared to the tested European and satellite data.

The used AIS data contains many errors due to the nature of the signal. AIS is a radio signal, rendering it sensitive to meteorological and magnetic factors. The errors caused by these circumstances can be present in every aspect of the AIS messages: both static (e.g. ship's identity) and dynamic information (e.g. ship's location) can be distorted. WP4 dealt with the errors in the static part of the message by using the frame of ships that was constructed by only using ship's identities that occur most frequently. The errors in the dynamic part of the message were dealt with by applying a median filter. That is, consecutive data points were combined and the median filter for 10 minutes was selected.

Definition differences are only a small issue when using AIS data. When filtering out maritime ships to determine port visits, only ships carrying goods to load/unload at the port should be counted. However, the frame of maritime ships mentioned before also contains a small number of ships, such as tugs, that might or might not carry goods to the end of (un)loading. There is no way of telling this from AIS data itself, so some ships that do not carry goods may be included.

The aim of WP5 is to investigate whether Mobile Phone data can be used to produce population estimates and mobility. During the ESSnet, some limitations in the access to mobile phone data preclude a fully-fledged investigation about the methodological and accuracy issues in dealt with the data. The activities concentrated on processing aggregated mobile phone data to connect them with the target populations at stake (although some partners had a limited access to some form of microdata). The question of accuracy is one of the central issues in the WP5 deliverable on quality (WP5 Del 5.5). However, when aggregated mobile phone data are used in a Bayesian hierarchical model to make inferences from the data to the target population, the standard tools to assess accuracy in this context can be used (credible intervals, posterior variance, and coefficients of variation, as well as model assessment techniques). Particular for a Bayesian approach, it is of vital important to assess the goodness-of-fit of the model to the input data, to control the a priori hypotheses (WP5, Del 5.3).

The aim of WP6 is to investigate how a combination of Big Data sources and existing official statistical data can be used to create early estimates for statistics. One of the Big Data source used is traffic sensors data. In the beginning, there are errors (e.g. measurement error, temporary sensor failure), so the data are edited to reduce these errors. Variability in traffic sensors data is expected due to the nature of the process of travelling on roads, and due to errors in data. Errors are also present in statistical data. For early estimates, one or more models are used. ME, RMSE, MAE and MaxAE are usually taken into account for comparison of different models: ME should be close to 0; The model with the lowest value of RMSE could be considered as the best, but also MAE and MaxAE should be as low as possible.

In WP7 three topics were investigated: sentiments revealed in the social media; agriculture; tourism/border crossing. Sentiment analysis is based on the posts from social media. People that are active on social media may differ from the total population and their number and activity can change over time. So the representativeness is an issue when using this data source. This makes a challenge to compare the results with the current population. Also extracting background characteristics of the users and accurately determining the sentiment of the text are not easy tasks. Depending on the social media channel, in some countries Twitter is not so popular and used only by a selected group of people. The accuracy is measured with machine learning algorithms and varies between 60 and 90%, depending on the training dataset and country (pilot was conducted by Poland, Portugal and UK).

In the field of agriculture, WP7 combined satellite images with administrative data and survey data to train machine learning algorithms to recognize different crop types. Sources can be linked accurately, findings can be verified by manual inspection of land lots and the first results are very promising. The accuracy, measured by the number of fields with crop types identified correctly, varies from 75% to 85% depending on the crop type and machine learning algorithm used (K-Nearest Neighbors algorithm or KNN) and Support Vector Machines (SVM) are the most accurate. The pilot was conducted by Poland and Ireland, using different approaches.

In WP7 there is an accuracy issue with the road traffic as there are some gaps in the data and some data must be estimated. This leads to the possibility of providing the data under or overestimated.

2.10.3 Discussion

Accuracy of Big Data based estimates depends on differences in the population composition in the source and those in the target population, errors in the data used, definition difference of variables and differences in their variation. Comparing the findings obtained from a Big Data based approach with those provided by a more traditional data source, e.g. survey or administrative data, is a first step towards obtaining insights in the accuracy of the former. Both bias and variance need to be considered here. Resampling is suggested to calculate the variance, while for the study of the bias a method based on simulation has been proposed. The reader is referred to the Deliverable 8.2 “Report describing the quality aspects identified in studies focussing on the use of Big Data for official statistics” for more details on the measures and indicators related to coverage, lack of representativeness, measurement errors, processing errors, as they were dealt with in the studies and applications carried out in the WPs of the ESSnet Big Data.

2.10.4 References

Weisberg, H. F. (2005). The Total Survey Error Approach: A Guide to the New Science of Survey Research, University of Chicago Press

WP1 (2017) Internal report Session 6 – Data Quality: A quality framework for on-line job vacancy data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/7d/WP1_Quality_Framework_v1.1.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

WP5 Del 5.5 (2018) Some Quality Aspects and Future Prospects for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/3/30/WP5_Deliverable_5.5_Preliminary_draft_version.pdf

Zhang, L-C. (2012) Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66(1), pp. 41-63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

2.11. Inference

2.11.1 Introduction

Big Data can be used in different ways for official statistics. It can be used as the main source of information, used as an additional source next to survey or administrative data, or merely to aid in the estimation process. Since the methodology in the last two cases does not differ from what is traditionally used, these cases are not discussed here. The reader needs to realize, however, that in all cases mentioned information needs to be extracted from Big Data which may require specific Big Data methodology. This is especially the case for unstructured data sets.

When the data in a Big Data source provides the main input for the statistic produced, in essence two different ways of using the data can be discerned. The first is that one aims to include as much of the data generated in the final output, as this enables very detailed statistics on small areas or subgroups. This approach is typical for smart meter data (WP3 Del 3.1), AIS based transport statistics (WP4 Del 4.3), mobile phone data (WP5 Del 5.3) and Road sensor data (WP6 Del 6.1; Puts et al., 2016). In all these cases, in the end, a model is used to infer from the data. The model is used to compensate for missing data, data of low quality, and/or differences in the population composition of the Big Data source and the target population. This need becomes particularly obvious when very detailed, regionalized, statistics are being produced. Here, because of low coverage in some areas, the original abundance of data reduces to a small amount (for some areas) and hence a model must be applied to compensate for this.

Another approach is used when only a relative small part of the data in a Big Data source provides information on the topic of interest. Here, after an initial data selection step, also a model-based approach is followed to enable to infer the phenomenon studied from the relevant part of the data. More on these approaches can be found in a recent European Master of Official Statistics Webinar on Big Data method and techniques (Daas and Puts, 2018).

2.11.2 Findings across WPs

The approaches mentioned above by which Big Data can be used for official statistics are all observed in the WP's of the ESSnet Big Data. Usually the focus is (initially) on Big Data as the main source but sometimes Big Data is used merely as an additional data source.

From the description given in table 2 (in section 1.1) it's obvious that WP1, WP3 and WP4 focus on using their Big Data sources (job vacancies on web sites, smart meters and AIS-data) as the main source of input. In WP2 this seems not to be the case; improving general information on enterprises is mentioned in table 2. WP6 focuses on combining various sources while in WP7 some studies focus on Big Data as the main source of information (for more details, the reader is referred to the findings section 2.10.2). In WP5 a Bayesian approach is followed to infer (WP5 Del 5.3). Clearly, a whole range of approaches to infer from Big Data has been used in the ESSnet.

2.11.3 Discussion

When Big Data is used as the main source of information, extracting knowledge from such data does not mean that a total new way of drawing inference needs to be developed. However, because of potential selectivity (i.e. 'representativeness') issues (see 2.10), special attention needs to be paid to causes of bias. Variance is considered less of an issue here as large amounts of data are being used. Therefore, a researcher needs to pay attention to the effects of (low) data quality and any of the decisions made during the processing of Big Data on the final outcome. In a perfect world, where Big Data contains perfect quality data including all units, simply adding all values up should suffice to obtain the total of the population. In reality, however, missing data (for particular areas or units) or including data from units not belonging to the target population, need to be dealt with. Because of this, a seemingly precise Big Data based estimate may still be way off (Buelens et al., 2014; section 2.10).

The difficulty when huge amounts of data are used for statistical inference is that the data in the source needs to cover the entire target population to enable model-free inference. This is (likely) what the (in)famous paper on "The End of Theory" in Wired magazine (Anderson 2008) wanted to indicate. However, in practice, model-free inference is hardly possible as quality issues or simple data delivery issues may prevent the continuous inclusion of the entire population (Puts et al., 2017). Applying a model enables one to deal with the real-world fluctuations that occurs in a data-deluged world. From the above it is clear that Big Data models need to be developed that are able to:

i) compensate for the lack of insight in the inclusion probabilities of the units in Big Data (unknown design issue);
ii) identify the relevant units/events in huge data set with a high precision and preferably high recall (imbalance issue);
iii) be implemented efficiently. Inferring from huge data sets should not take up to much time (efficiency issue).

Current research focusses on these exiting areas. Studies on non-probability sampling touch the unknown designs issue (Buelens et al., 2018) as do catch-recatch studies (WP5 Del 5.3). Applying Bayesian inference methods is another suggested approach (WP5 Del 5.3). However, one should not assume that Big Data is simply a large sample (Doherty, 1994) and be careful when (mis)using the 'representativity' argument (see section 2.6).

2.11.4 References

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired magazine 16-07. Link: https://www.wired.com/2008/06/pb-theory/

Buelens, B., Burger, J., van de Brakel, J. (2018) Comparing Inference Methods for Non-probability Samples: Inference from Non-probability Samples. Int. Stat. Review., *in press*.

Buelens, B., Daas, P., Burger, J., Puts, M., van den Brakel, J. (2014) Selectivity of Big Data. Discussion paper 201411, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Daas, P., Puts, M. (2018) Big Data methods and techniques. Webinar for the European Master in Official Statistics (EMOS) 2018. Link: http://konference.ef.uni-lj.si/emos/big-data-methods-and-techniques/

Doherty, M. (1994) Probability versus Non-Probability Sampling in Sample Surveys, The New Zealand Statistics Review March 1994 issue, pp 21-28.

Puts, M., Daas, P., de Waal, T. (2017) Finding Errors in Big Data. In: The Best Writing on Mathematics 2016, Princeton, USA. (Pitici, M., ed), pp. 291-299, Princeton University Press, USA.

Puts, M., Tennekes, M., Daas, P.J.H., de Blois, C. (2016) Using huge amounts of road sensor data for official statistics. Paper and presentation for the European Conference on Quality in Official Statistics 2016, Madrid, Spain.

WP4 Del 4.3 (2018) Report about sea traffic analysis using AIS-data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_21_v1.0.pdf

WP5 Del 5.3 (2018) Proposed Elements for a Methodological framework for the Production of Official Statistics with Mobile Phone Data. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/45/WP5_Deliverable_5.3.pdf

WP6 Del 6.1 (2016) Potential of Big Data and other sources with business cases for the aim of early estimates. Located at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/bc/WP6_Deliverable_6.1.pdf

3. Conclusions

Work packages 1 to 7 demonstrate how Big Data can be used in different domains such as assessing online-posted job vacancies and the identification of vessels position in European waters. The variety of implementations studied in the ESSnet Big Data suggests and demonstrates the potential of Big Data to produce various economic and social statistics. Since the executive summary of this report (page 4-5) summarizes the findings of this report, this part will not be repeated here.

No WP reached the full cycle of Big Data statistical production yet, i.e. from data access through data processing to data dissemination. However, some WPs are close. WP2, WP3 and WP4 are examples were Big Data based methods emerge and start to become fully developed (see table 1 on page 4-5). However, these methods still have to be flexible enough though, since Big Data are a product of technological development. Technology development underlies the variety and dynamic of data sources. To be in line with technology NSIs using Big Data need to cope with many dynamic data sources. One example: recently Facebook developed an application for online hiring targeting SMEs. To grasp all online job ads in the new situation one now has to scrap not only job portals and companies' sites but social media as well.

It is important to realize that Big Data sources are scattered across many private and government agencies. To ensure access to data and produce Big Data based statistics NSIs have to build partnerships network with both public and privately owned organizations. It is important that NSIs become the preferred partner and not just one of many partners. May be statistical institutes have a chance to become such a partner because of a number of unique selling points. These are: they have developed methodology for producing statistics, they have expertise to work with many data sources, they have no interest in a particular business but have interest in producing data of high quality useful for society and they guaranty to guard the privacy of all individuals included in the data sources. It is essential to ensure stable, long lasting, access to data to enable the production of Big Data based statistics that also enables the development of generic Big Data methodology.

Based on all work described in this report and in any of the other reports of the ESSnet and other related Big Data work of NSI's, an overview is made of the general approach followed by an NSI that wants to include Big Data in official statistics. From this, the following step-wise methodology has been derived:

1) Get access to Big Data (BD)
2) Perform an BD exploratory data analysis study (including a privacy assessment)
3) Study the objects (units/events) in BD and check if events need to be converted to units for the foreseen application
4) Compare the coverage of the objects in BD to those of the target population of the NSI
5) Study the variable(s) of interest in BD and compare these with those needed by the NSI (variables may be combined and/or processed here; e.g. creating features)
6) Compare the development over time and/or per area of the variable(s) of interest in BD with similar variables included in any other survey or register based results (if available)
7) Check the performance of various models and/or machine learning based applications on improving the relation described in the previous step
8) Determine the effect of any assumptions, short-cuts made, and/or quality issues and corrections on the comparison described in the previous step (may need to restart at step 3-6)
9) In case of any positive findings, check the reproducibility and stability over time of those results
10) Produce a first (beta-)product

After each step a go/no go discission is made in which it is decided to proceed to the next step. In step 6 data is compared which, especially for completely new output, may be challenging or even impossible to find. When that is the case, it is suggested to consult experts in the field. The list is a starting point and will undoubtable form the foundation of new and exciting future developments in the area of Big Data research.

4. Abbreviations and acronyms

AIS – Automatic Identification System
BD – Big Data
ESS – European Statistical System
ESSnet – European Statistical System network
ESA – European System of National Accounts
EU – European Union
FPA – Framework Partnership Agreement
GDP – Gross Domestic Product
ICT – Information and communication technology
IoT - Internet of Things
KNN - K-Nearest Neighbors algorithm
MAE - Mean absolute error
MaxAE - Maximum absolute error
ME - Mean error
MP – Mobile Phone
MSE - Mean square error
MMSI – Maritime Mobile Service Identity
NSI – National Statistical Institute
PCA – Principal Component Analysis
SGA – Special Grant Agreement
SMC – Secure multiparty computation
SMEs – Small and medium-sized enterprises
SVM - Support Vector Machine
UNECE – United Nations Economic Commission for Europe
URL – Uniform Resource Locator
WP – Work Package

5. List of figures and tables

Table 1. Overview of the methodological topics and there application in WP1-7 and other Big Data areas. Secure multi-party computation and Data process architecture are not included here. For each of topics applications are listed indicating different methods used……4-5

Table 2: description of the work packages in the ESSnet Big Data Programme……..6-7

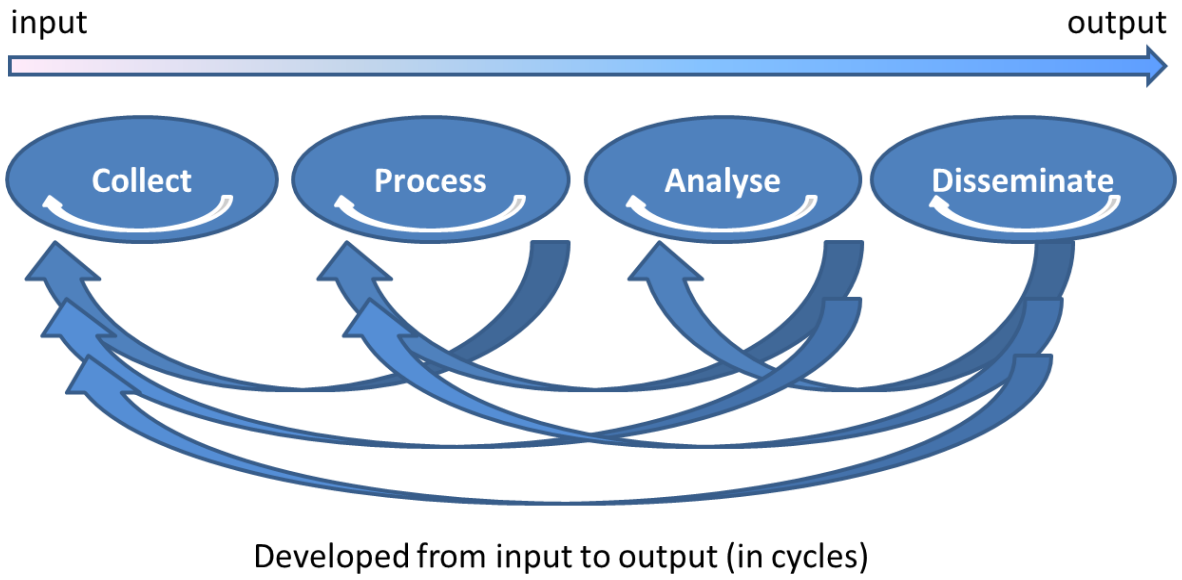

Figure 1: Illustration of the Big Data process life cycle. Arrows indicate the adjustment of previous steps after new findings and quality checks in the subsequent steps. The process starts at the left (input) and ends at the right (output). *14*