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

Quality

Deliverable 8.2

Report describing the quality aspects of Big Data for Official Statistics

Version: 7 May 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

Content

Executive summary

In this report quality is discussed in the context of Big Data. Topics included were identified in a WP8 workshop during which seven quality aspects were listed as the most important ones when using Big Data for official statistics (in the context of WP 1-7 of the ESSnet on Big Data). The seven quality aspects identified are: Coverage, Comparability over time, Processing errors, Process chain control, Linkability, Measurement errors and Model errors and precision. What these aspects have in common is a clear relation with either one or more causes of error (Coverage, processing errors, Linkability, Measurement errors and Model errors and precision) or the need to detect and deal with changes in the composition of the source (Comparability over time and Process chain control). In the various WP's in the ESSnet Big Data many causes of error were found. Some of them are unique, such as the effects of the scrambling of the Automated Identification Signal of ships in WP4 and the coverage issues for job portal vacancies in WP1, clearly indicating the need for new (Big Data specific) checks and correction methods. These findings also indicate the need to develop or update a quality framework for Big Data sources. When all the work in the ESSnet Big Data is finished one should have enough information to start to construct such a framework. In addition, Big Data sources may also change over time for which a number of causes were identified; mainly related to changes in the composition of the data source. These causes affect the comparability over time and require the need to track those changes. Causes identified were technological changes, changes in the policy of the data holder and changes in the population composition and/or amount included. The need to check and control these causes and their effect on the entire chain is part of Process chain control. Because of the large volumes of data involved, it is important to use efficiently implemented quality indicators or predictors. For a data driven process this introduces the need to not only focus on the quality of the output but also on the quality of each individual step in the process chain. Overall it can be concluded that some familiar and some new quality aspects have to be considered for Big Data sources which urges the need for the development of an extended quality framework. However, the experimental findings of the ESSnet also suggest that it will be challenging to apply standardized quality measures to the range of Big Data sources used. There is a definite need to extend on the work described in this report.

1. Introduction

The Introduction consists of three parts. The first part, a general introduction for the WP8 of the ESSnet Big Data, is the same for the Report on Quality, as well as for the Report on IT Infrastructure and the Report on Methodology. The second part of the Introduction is specific for this Report on Quality and deals with the following description of quality aspects in the ESSnet Big Data. The third part tries to contextualize the seven quality aspects with UNECE's quality framework for Big Data.

1.1. General introduction to the ESSnet Big Data and the WP 8

The overall objective of the ESSnet Big Data is to prepare the European Statistical System (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 (NSI) and 2 other Statistical Authorities (ONA) 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 1.

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.

Table 1: Description of the work packages in the ESSnet Big Data Programme

As can be seen from Table 1, 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 is living document)

8.2 Report describing the quality aspects of Big Data (the current deliverable)

8.3 Report describing the IT-infrastructure used and the accompanying processes developed and skills needed to study or produce Big Data based official statistics (delivered on 31 January 2018)

8.4 Report describing the methodology of using Big Data for official statistics (planned for 31 May 2018)

1.2. Introduction to the report on quality

The quality aspects (chapters) described in this report are based on the results of the WP 8 workshop held on the 25th and 26th of April 2017 at Statistics Netherlands. During these two days, a group of 18 experts (representatives from a large number of the partners involved in the ESSnet Big Data, from as many WP's as possible, and from Eurostat) identified the main topics in the areas of Methodology, Quality and IT when using Big Data for official statistics in the context of WP 1-7 of the ESSnet on Big Data.

It is important to note that the members of the workshop in April 2017 focussed on quality aspects especially relevant in the WP1-WP7. Therefore, the seven quality aspects, which were identified as the most important ones in the context of the pilots from WP1 to WP7, make no claims of being the most important quality aspects when working generally with Big Data.

The seven quality aspects as identified in the workshop, sorted by the importance assigned in the workshop, are:

1. Coverage
2. Comparability over time
3. Processing errors
4. Process chain control
5. Linkability
6. Measurement errors
7. Model errors and precision

It is worth noting that our discussion of quality aspects is based on the experiences of pilots, not on products already declared as Official Statistics. As a result, the approach might differ a bit here: with official statistical products, it is clear from the beginning, which quality dimensions have to be met. Working with pilots involving Big Data sources, the approach differs, here a more data driven approach is followed. The work packages explored new data sources as well as possibilities to use them. The consideration of quality aspects (at the output side) was not the main focus from the beginning, but happened more in the course of exploring and using the new data source, or happened even only retrospectively.

We would like to emphasize that we have deliberately chosen the term "quality aspects", because we wanted to avoid the term "quality dimensions". The reason behind is, that the term "quality dimension" is mostly used in quality frameworks, which cover the measurement of quality in a systematic and exhaustive way, whereas our approach is without any claim of completeness.

Nevertheless the quality aspects listed in this report are in some way related to the quality dimensions as defined in Article 12 of EU-Statistics Regulation 223[1]. Regarding quality frameworks it should be mentioned that the Code of Practice for European Statistics (CoP) is currently reviewed to be fitter for modern production of statistics. One of the main aspects of the revision is the potential use of so called new data sources including Big Data.

1.3 The 7 quality aspects and UNECE's quality framework for Big Data

Contrary to the case of survey data as data source – or to a certain degree also the case of administrative data - there exists no well-established strong quality framework for statistics based on Big Data. There are a variety of reasons for this, the most important ones are: Statistics based on Big Data sources is still a young field in Official Statistics, and the adaptation (or creation of a new) quality framework needs time. Further, Big Data sources are so diverse, that it is hard to cover all quality aspects in one framework.

The most well-known quality framework so far, focussing specifically on the quality of statistics based on Big Data sources, is the Big Data Framework by the UNECE from December 2014[2]. As noted in this framework, NSOs need a series of quality principles that apply across the business process when working with Big Data rather than focusing only on quality of statistical outputs. This is the reason

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32009R0223&from=EN, accessed March, 27th 2018

[2] https://statswiki.unece.org/display/bigdata/2014+Project?preview=%2F108102944%2F108298642%2FBig+Data+Quality+Framework+-+final-+Jan08-2015.pdf, accessed March 3rd, 2018.

why the Big Data quality framework developed by them provides a structured view of quality at the three phases of the business process:

- Input, comprising acquisition, or pre-acquisition of the data
- Throughput, comprising the transformation, manipulation and analysis of the data
- Output, comprising the reporting of quality with statistical outputs basing on Big Data sources

The report further identifies three hyperdimensions: the source, the metadata and the data. The hyperdimension *source* refers to factors associated with the type of data, the characteristics of the data owner and the governance under which the data is administered and regulated. The hyperdimension *metadata* relates to information available to describe the concepts, the contents of the file data set, and the processes applied to it. The hyperdimension *data* is about the quality of the data itself. All of the hyperdimensions have corresponding quality dimensions. It is important to note that all quality (hyper)dimensions can play a role in all three phases of the business process.

In Table 2, we try to contextualize the quality aspects of this report in relation to the structure (the three phases as well as the three hyperdimensions) of the UNECE's quality report:

- We assigned the quality aspects of this report to one (or more) phases of the business process.
- We illustrate with the help of colours to which hyperdimension the considered quality aspect (mostly) corresponds to.

Please note, that "Coverage" as well as "Comparability over time" are listed both in the Input as well as in the Output phase. Further, "Comparability over time" affects all three hyperdimensions, because the availability of the source, the metadata as well as the data itself can change over time.

Input	Throughput	Output
Coverage	Processing errors	Comparability over time
Measurement error	Process chain control	Coverage
	Model errors and precision	
Comparability over time		
Linkablity		

Hyperdimension: Source
Hyperdimension: Data
Hyperdimension: Metadata

Table 2: The seven quality aspects in the business process and with relation to the hyperdimensions "Source", "Data" and "Metadata"

It is worth noting that the UNECE's quality framework does not provide a taxonomy of the processes involved in the Throughput-phase, nor do they list a specific set of quality indicators for this phase. Instead, they describe some general aspects. Similar considerations can be found in our quality aspects "Processing errors" and "Process chain control".

The order in which the quality aspects are listed in the following does not represent the assigned importance (as described above), but follows the phase to which they (mostly) correspond to.

2. List of Quality Aspects

2.1. Coverage, Accuracy and Selectivity

2.1.1. Introduction

Information on the population included in a Big Data source is vital for reliable statistics. Important for this issue are the lack of information on the units included, their duplication and their selectivity.

Coverage is one of the quality aspects (errors) that affect the **accuracy** of statistical information. Accuracy is considered as one of the quality dimensions, both for traditional data as survey data, administrative data and Big Data as well (Daas et al. 2009, Statistics Canada 2002, UNECE Big Data Quality Task Team 2014). Coverage is identified as a quality dimension related to the hyperdimension "Data" as identified in a quality framework for administrative data sources (Daas et al. 2009).
Generally speaking the accuracy is related to the degree to which the information correctly describes the phenomena of interest and it is usually characterized in terms of error in statistical estimates, traditionally decomposed into bias (systematic error) and variance (random error) components. The coverage errors, as potential source of inaccuracy, can be distinguished in **under-coverage** and **over-coverage**. The former occurs when units belonging to the population of interest are not included in the available data, whereas the latter refers to the situation in which out-of-scope units (including duplicates) are erroneously in the data. Generally speaking, under-coverage and over-coverage may become a key concern if they affect the **representativeness** of data. Buelens et al (2014) state: "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as **selective**." Since the selectivity is related to specific aspects (variables), it could be the case that a set of data that is highly selective may nonetheless be useable for some purposes but inadequate for others.
The coverage and representativeness of Big Data may be complicated by the fact that often Big Data refer to units only indirectly related to the statistical units of interest. The lack of direct connection (see Linkability in section 2.5) between the statistical target population and the available population coming from the Big Data source makes it often problematic to assess the coverage and representativeness of Big Data.
Selectivity (or representativeness), as well as other coverage errors affecting accuracy, should be analysed in a total error approach.

Coverage indicators developed for survey data and for administrative data can usually be used to measure the coverage of information available on Big Data. The same reasoning applies to selectivity indicators. The possible indicators proposed for Accuracy and Selectivity by the UNECE Big Data Quality Task Team (2014) are the following:

1. Measures of distance between Big Data population and the target population (e.g. Kolmogorov-Smirnov Index, Index of dissimilarity)
2. Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by Big Data source
3. Assessment of spatial distribution of measurement instrument and of periodicity of observations

The second point can be expanded as in survey data and administrative data, referring to the over-coverage rate and under-coverage rate. Linkability can be a prerequisite if one needs to assess coverage and selectivity on the basis of methods which assume comparison at unit level.

The representativeness issue can be somehow fixed if there is the possibility to calibrate the Big Data or to perform external validity checks using reference datasets. The relationship between the target statistical population and the Big Data source plays a crucial role in assessing the risk of selectivity. In the next figures, some examples are provided:

Scenario 1. Partial overlapping between target population and population covered by the Big Data, under and over-coverage occur

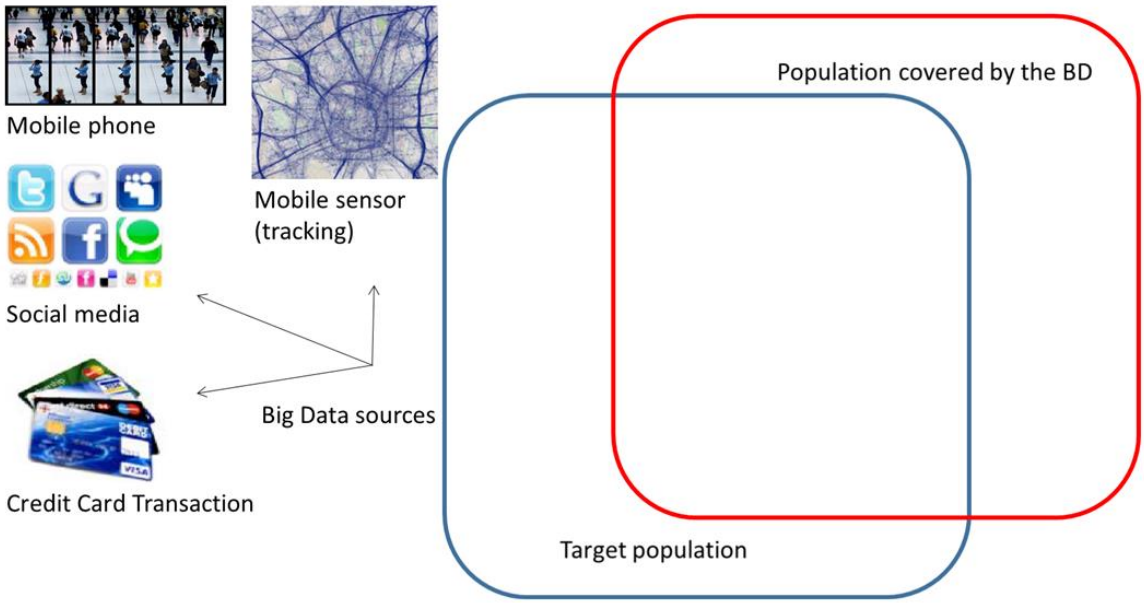

Figure 1 Illustration of Scenario1

The first scenario is also representative of the case of multiple populations, which are present in some Big Data sources. The images represent potential examples of Big Data sources belonging to the considered scenario. Social media, for instance, is composed of messages created by accounts of persons, companies and groups. The same holds for smart meter data, here electricity data from households and companies is mixed. It is important to exclude any of the units not belonging to the target population prior to analysis.

Scenario 2. The population covered by the Big Data is a subset of the target population

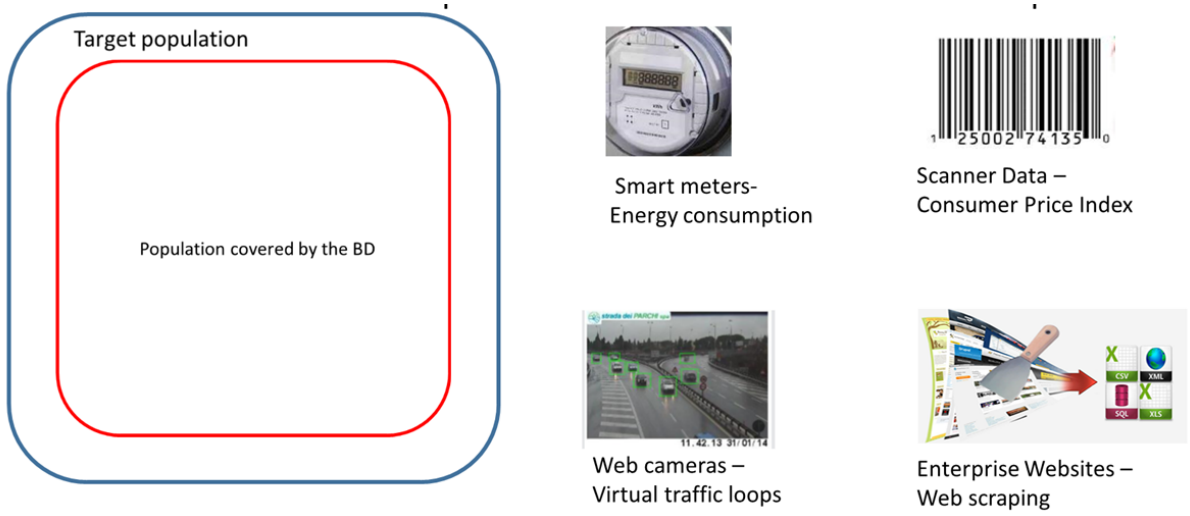

Figure 2 Illustration of Scenario 2

Scenario 3. Complete overlapping between target population and population covered by the Big Data, no under and over-coverage occur

Figure 3 Illustration of Scenario 3

In the third scenario the risk of selectivity is reduced.

Unfortunately, there is an additional scenario that is the worst scenario for Big Data usage, i.e. the total absence of a subset of the target population in a Big Data source.

In the next paragraph, examples of coverage issues and methods to address them and evaluate the related indicators are provided from the working packages WP1- WP7.

2.1.2. Examples and Methods

WP1) WEB scraping for job vacancies

Coverage problems of job vacancies on job portals vary from county to county and it depends on the market of job portals. In some countries, there are just a few job portals while in others there are

hundreds of them. For example in Slovenia there are only two major job portals for all vacancies, while in Germany there are job portals specialized for certain professions. In the case of non-specialized job portals, there is a possibility, that certain professions are not represented there, but are advertised on social networks, forums etc. In any case, if a country does not include a certain source, it can face the problem of undercoverage for certain professions (even if that source is small). Similarly it can happen with respect to NACE activities, areas of employment etc.

But coverage problems arise even if we don't miss out on certain job portals. Those problems are:

- Duplicate job advertisements between or/and within some portals. That can happen if the ad has expired and was reposted on the same portal, but with different description and/or job title. The same is the reason for duplicates between portals at a given point in time.
- The existence of "ghost vacancies", meaning: there is a job advert for a job vacancy but in reality employers are just exploring the labour market and labour supply.
- One advert may represent more than one job vacancy. By scraping one can detect only the number of job ads, but not the number of job vacancies per job ad, if this information is not explicitly written in the ad

WP2) Web scraping Enterprise Characteristics

The WP2 investigates web scraping experiments with the aim of deriving statistics on enterprises from information found on the web, especially the websites of enterprises. Examples of statistics derived from the web scraping experiments are: "Number of enterprises that have a website and use it for online ordering (E-commerce)", "Number of enterprises that have a website and use it for job vacancies advertisements", "Number of enterprises that have a website and that are present on social media". There were four use cases, referring to URL Retrieval, E-commerce, Job vacancies ads on enterprises' websites and to Social Media Presence.

As far as the URL Retrieval use case, the considered data sources are the Business Register, of which a given percentage has a known URL, the responses from ICT survey and web-scraped websites from enterprises. In the following graph a representation is given.

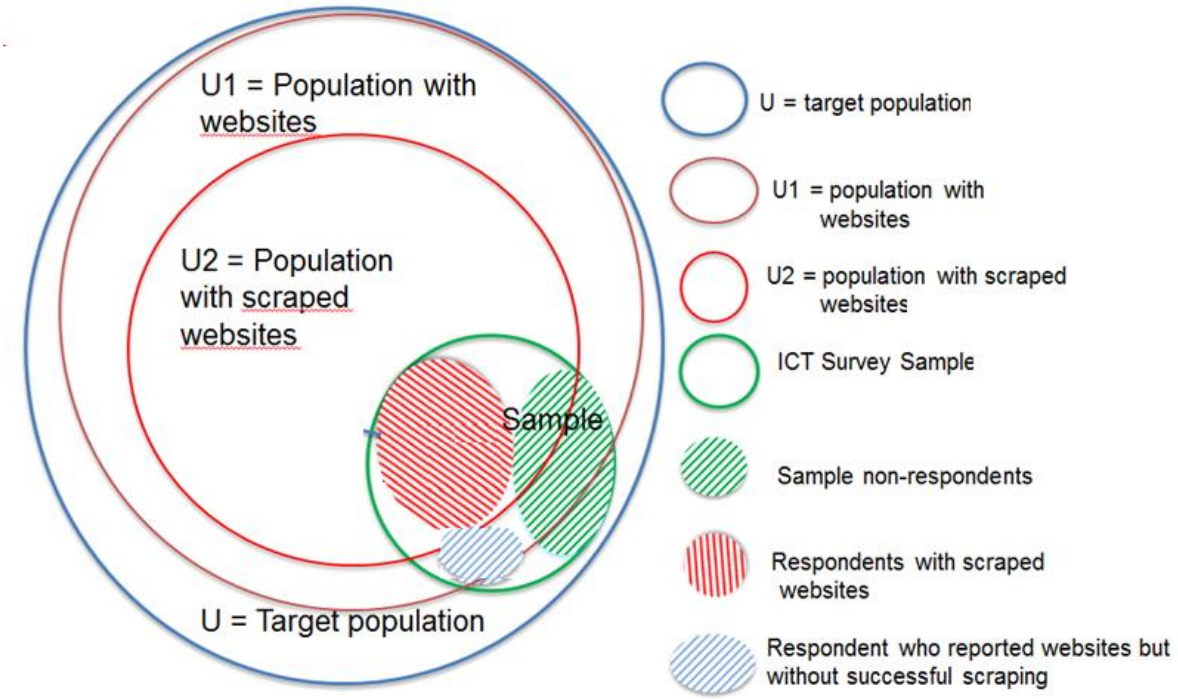

Figure 4 Illustration of different populations

As far as the Social Media use case, the considered data sources are the Business Register, of which a given percentage has a known URL, the responses from ICT survey, web-scraped websites from enterprises and Twitter messages via Twitter API (at least in some countries, i.e. the Netherlands).

The analyses on survey data and web-scraped data highlight some differences with respect to some characteristics of the enterprises. However, the selectiveness aspect was not investigated so far. The different reference time of the considered data and changes in data source – enterprises' websites could also explain some differences.

WP3) Pilot on smart meters

The most obvious coverage problem with smart meter data happens during the roll-out period, when the penetration rate of the smart meter devices is growing, and quite naturally, thereby also the consumption of electricity, measured by smart meter devices, grows. So as to end up with reliable estimators for the overall electricity consumption, the NSIs can either integrate – if possible - the smart meter data with traditional, non-smart meter data, or the NSIs can apply weighting procedures.

A less obvious coverage problem concerns the electricity produced for own consumption, which is not measured by the smart meters, resulting in undercoverage. The members of the WP3 were aware of this coverage problem. In the Estonian case, the Report on data access and data handling reads: "The total amount of unmeasured consumption is negligible compared to total consumption."

WP4) AIS Data

WP4 investigated quality aspects of the European AIS data source from Dirkzwager (a maritime and nautical service provider in the Netherlands) in deliverable 4.3, concluding that the quality is good by means of the quality and metadata framework. Almost all factors of the quality framework are judged as mostly positive. However, "spatial coverage" is judged as insufficient, mainly because not all European coastal areas are covered in the European AIS data. In deliverable 4.3, the AIS data of DZ is compared to national data for Denmark, Greece and Poland. In almost all cases, national data contained (much) more data than DZ data. DZ misses data on complete areas in coastal Europe. Thus, ports visits and journeys cannot be analysed for all European ports and ship routes. The number of messages in areas covered by DZ is usually lower in the DZ data compared to the national data. In general, WP4 is not satisfied with the coverage of the DZ data. Coverage differs per country, but if you want to analyse the whole of Europe it does not suffice. If DZ data does cover a port, the data is sufficient to determine the port visits. However, it is not sufficient to determine ships' journeys, especially in areas with a capricious geography.
Deliverable 4.3 also describes the comparison of satellite AIS data and AIS data from land based stations. The satellite data completely covers the Mediterranean Sea. However, satellite data structurally detects a lower number of ships per port compared to the high quality of the national AIS data from Greek land based stations. Also, the data points per ship are lower in the satellite data. Both findings are probably due to the lower frequency of that data: if ships are in the port shortly, they might be missed by the longer latency of satellites.
All in all, satellite data is indispensable when it comes to following ships across oceans. Then the distance from the ship to the land receiver is too large for the land receiver to still be able to pick up a signal from the ship. However, satellite data is less useful to perfectly model shipping routes in more complex areas. That is, for areas where the geography is more complicated, e.g. due to isles or different water depths, frequency of reporting by satellites might be too low.
When investigating the coverage and representativeness of AIS data it is important to keep in mind that:

- AIS is a radio signal, parts of the messages can get lost or scrambled due to factors such as meteorology or magnetics.
- Receivers have timeslots in which data is received. In busy areas with many ships, not all data from all ships may fit into this time slot. This results in the loss of data on some ships in that time slot.
- Ships can turn off their AIS transponder resulting in the disappearance of a ship.
- AIS receivers on land can only pick up signals within the range of about 40 sea miles. Therefore, land receivers have a very limited coverage of signals transmitted from sea which results in loss of information of ships on open sea.

WP5) Mobile phone data

The aim of WP5 is to investigate whether Mobile Phone data can be used to produce population estimates. In this case, whether the Mobile Phone (MP) data is used at aggregated level or at unit one, the basic assumptions for avoiding selectivity are:

1. High level of MP penetration rate, that is the number of active mobile phone users per 100 people within a specific population, which is technically not a penetration rate as it does not account for users having multiple mobile phones and hence can exceed 100% due to double counting

2. High level of MP area coverage, that is the percentage of the territory over which the MP operators offer MP services
3. To know the MP operator market share, that is the number of subscriptions of a specific MP operator over the overall number of subscriptions, when statistics are based only on data provided by one operator while more than one operator are active in the country.

The importance of representativeness of MPD has been clearly stated by the WP5 even if WP5 does not state so far why it can be sure that representativeness is given.

In the Netherlands, data provided by one of the three main mobile phone providers was used to estimate the number of festival visitors. Interesting of this approach is that the estimate can be compared to the total number of visitors reported by the festivals organization. This approach was applied to the famous Motorsport event at the TT circuit in Assen (June 28, 2014) and the largest European Truckstar festival (July 26-27, 2014) held at the same location. The mobile phone data based estimate resulted in 104,000 and 23,000 visitors, respectively; on top of the 68,000 people normally living in that region. The official total number of visitors reported by the organisers was 90,000 and 55,000, respectively. The approach followed is described in Tennekes (2017). Comparing both estimates reveals that the last one is clearly overestimated, which suggests that the number of visitors with a mobile phone of the provider studied were either much more interested in trucks than in motorsport or that a lot more local people attended the Truckstar festival compared to the TT.

WP6) Early estimates

In Slovenia, traffic loops data are used in conjunction with other data (administrative data - industry) to estimate in a timely fashion the current levels of GDP. The traffic loops data are used for nowcasting - using present data and estimating missing data to get an estimation of a correlated economic indicator. Traffic loops data are also used for other economic indicators, such as the Industry Production Index, with the same methods. There are studies to further exploit these data for other economic indicators, thanks to the ease with which they are obtainable and the good correlation with the GDP.

In the Slovenian case, the coverage of obtained sensors' data is complete. As it is obtained by the Slovenian Ministry of Infrastructure, all existing traffic sensor data are accessible. Furthermore, the list of all traffic sensors on Slovenian roads is available freely online on the Ministry's internet pages, including the geo-position of every sensor. This is a .pdf file, but an excel file can be received with a request to the Ministry. The sensors are present throughout the country on every speedway and regional road, however it must be said that the number of sensors on the speedways is lacking.

A case for over-coverage could be made from the fact that we are not able to differentiate between foreign and domestic traffic by this data. If this difference is found to be important, some kind of linkage with border traffic sensors will be needed to assess the number of foreign vehicles.

Problems with missing data of road sensors on the micro level were described in more detail by the Slovenian members of WP6 in Chapter 2.5 on Processing Errors.

2.1.3. Discussion

Selectivity has become an important aspect in assessing usability and therefore quality of Big Data. The selectivity of Big Data depends on the characteristics of the data itself (Daas and Burger, 2015). Compared to the traditional survey data, the Big Data may contains a considerable set of data that is not in the scope of the target population and, on the other hand, not all units in the target population may be included in Big Data; in any case, units that are included are not a random sample from the target population. Moreover, often Big Data are composed by events that need to be related to the statistical units in the target population that generated the data.

Therefore, selectivity issue is strongly related to other quality aspects. For instance, the linkability of Big Data can be a prerequisite to assess selectivity and Chapter 2.5 of this report is devoted to it. The role of duplicates (out-of-scope units) and their impact on the "Comparability over the time" is highlighted with several examples in Chapter 2.2 of this report.

The selectivity and coverage errors may become an issue also in weighting procedures, if expected in the Big Data analyses, as illustrated in chapter 2.3 on "processing errors". In fact, often, in the use of Big Data the weighting procedures are mainly devoted to adjust for the representativeness of the undertaken working frame.

2.1.4. Literature

Buelens B., Daas P., Burger J., Puts M., van den Brakel J. (2014) Selectivity of Big Data Discussion Paper, Statistics Netherlands

Daas, P.,Ossen, S., Vis-Visschers, R., &Arends-Toth, J. (2009), Checklist for the Quality evaluation of Administrative Data Sources. Statistics Netherlands, The Hague/Heerlen

ESSnet Big Data (2017) Deliverable 4.3 Report about sea traffic analyses using AIS-data Version 2017-07-21

ESSnet Big Data (2016) Smart meters Deliverable 3.1 Report on data access and data handling Version 2016-07-29

Daas, P., Burger J. (2015) Profiling Big Data sources to assess their selectivity, Proceedings NTTS 2015 Conference, Bruxel

Statistics Canada (2002) Quality Assurance Framework

Tennekes, M. (2017) Statistical inference on Mobile Phone data. Presentation for the European Statistical Training Program. Available at: https://circabc.europa.eu/sd/a/6317d43e-1293-401e-81f2-f50a59127300/Mobile_Phone2.pdf

UNECE Big Data Quality Task Team (2014) A Suggested Framework for the Quality of Big Data

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

2.2. Measurement Error

2.2.1. Introduction

The values included in Big Data may not be all correctly measured; some may contain errors. This affects the outcomes produced, certainly when a systematic bias is introduced. Measurement errors do neither include errors with respect to missing values, nor do they include problems with under- and over-coverage.

In survey based statistics, measurement errors are errors that occur during data collection and cause recorded values of variables to be different from the true ones. The values included in Big Data are normally not collected in the usual sense of the word. They are rather generated, sometimes directly by technical processes or related to processes mapping the interactions between human beings and machines. The collection process by the NSI is mostly downstream to this generation process. But also the generated data may contain errors. This affects the outcomes produced, certainly when a systematic bias is introduced.

Statistical microdata always bear the risk of errors. For example for business statistics, survey based data are a by-product of companies' and government activity. Companies have no incentives to spend much time to fill required forms accurately. This is a burden for them and they minimize efforts for producing data for government statistics (Morgenstern, 1963). Morgenstern did a deep analysis of business data errors. Nowadays we can add to Morgenstern's insights the increasing reluctance of households to respond (unit response), or those who responds tend not to answer some questions (item non response) and when households answer they are less accurate (measurement error). The statistical community is pre-occupied with rising non-response rates but pays no attention to other sources of errors (Bruce D. Meyer et al.).

When it comes to measurement errors, there is an important difference between Big Data and survey data survey based statistical errors happen due to the reluctance and/ or inability of respondents to provide honest and unbiased response; due to the type of survey instruments used for data collection that may lead to the recording of wrong values; due to the influence of interviewers on respondents. Big Data errors share many of the challenges of surveys; however their specificity is of a more technological nature. Big Data is a product of technological development and measurement errors are usually a consequence of machines collecting data from other machines. Big Data rely on direct observations of vessels position, road traffic intensity, posts in social media etc. It is really hard to determine what measurement errors are in Big Data. The main concern is that sometimes one cannot know the details of how the data were collected or generated. Additionally, sometimes definitions of the terms and measures being used are unknown, so replication of measurements and analyses based on measurements might not be possible. This could make a comparison of statistics produced from survey data and statistics from Big Data really hard.

In the next chapter we list only a small fraction of possible measurement errors that may occur.

2.2.2. Examples and methods

The following example of the Mariana trench depth demonstrates how measurement errors in Big Data sources are generated by the usage of technological tools and software products. Mariana trench is in the ocean and the deepest location on Earth. But measurements vary from 11 km, or

10 994 m to 11 034 m of depth, depending on what site you are looking at (National Geographic, Live Science, Geology). Similar varying results are produced when one searches for the question "How high is the Mount Everest?" Being the deepest and the highest places on earth, both Mariana trench and Mount Everest have been subject of extensive research and measurements. One would expect no variation in measurements. However, as these prominent examples show, measurements vary.

WP1) and WP2) Web scraped data: job vacancies and enterprise websites

The example of web scraping of job vacancies on job portals faces measurement errors, that are mainly the result of scraping errors (Scraper may download incorrect data from the web page), errors on the web page or incorrect data on the web page (e.g. employers may upload incorrect data).

WP4) AIS data

Measurement errors in AIS data can be considered as technical errors and human errors.

Technical errors are related to dynamic data such as position of ship, speed, course and rotation which comes from AIS device (sensors, cables and antenna). As AIS is a radio signal, technical errors can also arise due to meteorological or magnetic factors disturbing the transmission of the radio signal. These errors can affect every part of the message.

Human errors are related to static (ship number, ship's name, call sign, type, length) or voyage data (draught, destination) which are manually entered in the AIS devices so therefore are a common cause of errors. Most of these errors are due to faulty or missing input by the ship crews.

Apart from these systematic errors, all of the parameters can be erroneous due to technical issues (e.g. meteorological factors, distance to receiver).

All those factors are the cause for the following additional errors:

- Location errors: data show that vessels are located in areas where there is no river or sea.
- Messages are transmitted encoded. As a result, an error in one transmitted 'byte' can result in an error in one or multiple fields in the decrypted message. Most of the times, these errors are detectable as the result yields an invalid variable, but sometimes they result in valid variables. For instance, coincidentally the resulting MMSI (Maritime Mobile Service Identity) can be a technically valid, but incorrect MMSI, resulting from an erroneous detection. These errors can arise for every variable, so this can for example result in erroneous latitude and longitude, yielding faulty locations that are quite far away from the actual location of the ship. In turn, this can result in a very high journey distance of ship.

WP6) Early Estimates, Road Sensor Data

The most important data from road sensors are the number of vehicles per lane and the type of vehicles. The most common measurement error is that sensors measure more vehicles than it is possible at a given space and time. They sometimes also measure -1 vehicles. Another measurement error depends on the type of the road sensor, with some road sensors being theoretically able to the type of the vehicles, but classifying them in a wrong way (some road sensors recognize more types of vehicles than others, some can't recognize any type at all, but are simply counting).

2.2.3. Discussion

Measurement error is the error of data collection and not the error in the source itself. It is easy to mix it up with the problem of validity or coverage. Contrary to measurement errors, which ask about errors occurring in the process of data collection, validity of a dataset is the extent to which it measures what the user is attempting to measure. For example, if a company is advertising a job vacancy on its homepage, when in reality there is no job vacancy and we scrape this (wrong) job vacancy, this is not a measurement error. If we count this vacancy, we will end up with additional job vacancies (example of overcoverage in Chapter 2.2). The same holds for the problem of technical errors of AIS devices. If there is no transmission due to meteorology or if the ship disappears from the radar due to turning off the AIS transporter, this is considered as non-response and it is not a measurement error.

These examples illustrate that there exist problems which seem to be measurement errors at first sight, but turn out to be a different kind of error on closer examination: They fulfil the first part of the measurement error definition at the beginning of the chapter, which says "error that cause recorded values of variables to be different from the true ones". However, they do not fulfil the second part of the definition, which says, "Data may not be correctly measured". If we scraped, for example, the same web site repeatedly, we would get the same (potentially wrong) information. This shows that this kind of false information is not the result of errors in measuring.

At the end, the quality of data collection depends solely on how good the tools are (that is APPs, scrapers, sensors etc.) that are gathering the data. Generally speaking, it seems to be easier to overcome measurement errors that are the result of technical errors. It is much harder to overcome errors that are the result of human behaviour, weather disturbances etc. These second kind of errors result from problems of the source itself. In this case, we no longer talk about quality of measuring, but problems of validity, which was left out of this quality report.

When using Big Data one has to distinguish between data generation and data collection. Data generation is the process where data comes into existence. In case of Big Data this is normally associated with technical processes. The understanding of the generation process for Big Data is crucial to evaluate and assess possible sources of bias in the context of measurement. On the other hand data collection is a process performed by an NSI and in most of the Big Data cases this is not causing additional measurement errors.

2.2.4. Literature

Geology, Deepest Part of the Ocean: https://geology.com/records/deepest-part-of-the-ocean.shtml

Livescience, Mariana Trench: The Deepest Depths: https://www.livescience.com/23387-mariana-trench.html

National Geographic, The Mariana Trench: Earth's Deepest Place, https://www.nationalgeographic.org/activity/mariana-trench-deepest-place-earth/

Meyer B. D., Mok W. K. C., Sullivan J. X. (2015). Household survey in crisis, NBER Working paper No. 21399: http://www.nber.org/papers/w21399

Morgenstern O. (1963). On the Accuracyy of Economic Observations. Princeton University Press.

Rino Bosnjak et al, Automatic Identification System in Maritime Traffic and Error Analysis, February 2012: http://www.toms.com.hr/archive/vol1/no2/toms_vol1no2_doi002.pdf

2.3. Comparability over Time

2.3.1. Introduction

To produce comparable statistics over time, it is essential that the source remains accessible, relevant and its content remain usable.

Article 12 of EU-regulation 223 defines comparability over time as the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures over time. Therefore, applying common standards is a prerequisite to have comparable data over a "reasonable" time period. The adjective "reasonable" implies that definitions could change, or new units could appear (financial industry is a good example), etc. Comparing the European System of National Accounts (ESA) edition from 1953 which had only 48 pages to the ESA 2010 edition, which has 652 pages, it gives a good idea how many new concepts have evolved since 1953.

The European Code of Practice covers mainly survey based statistics, and to a certain degree statistics based on administrative data. As also noted in the Big Data Framework by the UNECE, the usual quality frameworks focus mainly on the quality of the output, but less on the quality aspects of the input side, namely data acquisition and data accessibility. But the guaranteed and ongoing access to the source is one of the decisive risks when working with Big Data. For the quality dimension Comparability over time, the guaranteed access to the Big Data source plays a crucial role. Often Big Data sources are owned and administered not by the NSI but by an (often private) data owner. The experience accumulated within the WPs clearly demonstrates that neither accessibility nor content usability is ensured in the long run.

The literature is abundant with good practices of employing Big Data: from a city traffic management to evaluation of Internet economy. Most of these cases are research work and the comparability over time is not an issue for pilots only produced once. But the question, if comparability over time is guaranteed will play a crucial role when NSIs have to assess if they can produce Official Statistics based on Big Data sources on a regular basis.

We will make an overview of working packages WP1 to WP7 to see how comparability issues are addressed.

2.3.2. Examples and Methods

The described projects of the WPs were only pilots which explored possibilities to use Big Data sources to produce statistical output; comparability over time did not play a role in the respective WPs. Still, the following aspects with respect to comparability over time would play a role if the Big Data based statistics were produced on a regular basis.

WP1) WEB scraping for job vacancies

Multiple counting and comparability over time

Job portals are the source of information for job vacancies (JV). WP1 distinguishes three types of job portals: *Job boards where employers place job advertisements; Job search engines which consolidate job ads from other sites; Hybrid portals which do both.*

In addition to the above mentioned job portals employers place job ads on their own website. Having so many websites where JV could be placed means that a job ad could be counted many times. Moreover, it is possible a job ad to be still visible after the job is not vacant any longer. Job portals make partnerships either to increase the visibility of their own job ads by allowing partners to publish them on their portals or to expand their hit list by publishing job ads from other sites. These partnerships could augment multiple job ads counting.

Multiple counting undermines comparability over time. It introduces an additional variance that has nothing to do with change in supply and demand on labour market. In good times when the economy grows and demand for labour is high, multiple counting is high as well. In bad times when labour demand is low multiple counting is low as well. In bad times the labour market looks even worse and in good times it looks better than it actually is. Multiple counting variations have a negative effect on comparability over time and comparability across countries. National economies have different economic cycles. An economy of one country could grow at the same time when the economy of a different country could shrink. Different economic cycles can underlie different intensities of multiple counting across countries.

Technological development and comparability over time

Another important source of variation is related to technological development. Recently Facebook developed an application useful for both business and job seekers. The application made it possible for businesses to create job posting and job seekers could apply for opened position directly on the social media. Businesses could post job ads and arrange interviews from mobile devices. Job seekers could post job alerts for positions they are interested in. The application is available in the USA and Canada and spreads to 40 more countries (see: *You might find your next job on Facebook, The Telegraph (28 Feb., 2018)*). It is free of charge and focuses on small business. Facebook application could motivate more small businesses to fill in job vacancies via Facebook and more online job ads will appear compared to previous periods. This is an example of technology induced variance. More job ads appear online because more companies hire online. To be in line with technological trend job ads scraping should extend to social media.

Summing up, both technological development and multiple counting undermine comparability over time. It seems that technological factors are more important for comparability than multiple counting. Technology motivates companies to shift from more traditional channels of hiring (personal contacts, printed media, etc.) to online hiring. Web scraping will register more job ads because more companies hire online and due to a bigger amount of vacant jobs. In contrast, multiple counting will fade over time because scraping tools will get sophisticated enough to detect the same job ad across different sites.

WP2) Web scraping Enterprise Characteristics

The web-scraping methods employed in this work package are entirely based on collection of textual information from web-pages. This may become increasingly challenging as the Internet evolves and more data may be encoded in forms that are harder to extract – audio or video files or increasing use of interactive or user-specific content It may become more challenging to extract information from websites for some enterprises – potentially large enterprises, or those in the creative industries.

Web-scraping may therefore need to be increasingly carried out by specialists outside of NSIs or new skills have to be developed in the NSIs.

Another issue which may affect comparability over time is a lack of continuity of data from the source, e.g. the Internet, as the NSI is unable to ensure that this source is available in the long run to produce statistical products with high quality. The access to the source may also be disturbed by technological changes that are not under the control of the NSI.

WP3) Pilot on smart meters

The source of information is smart meters. There is a strong tendency toward employing more smart meters in the EU. This is a prerequisite to produce unified statistics on electricity consumption based on smart meters. It is good to make difference between business entities consumption and households' consumption. Business entities consumption is intermediate consumption. Household's consumption is a final consumption. Differentiation between final consumption and intermediate consumption is important for national accounts. Still, there are challenges ahead. The technological progress in electricity production and distribution is huge in the last couple of years. Photovoltaic (PV) becomes cheaper and much more efficient. Smart grids distribute electricity with minimum losses. As PVs become affordable more households will be both electricity consumers and electricity producers. The same holds for the business sector. Technological advances allow companies to cover at least part of their electricity consumption from their own production. Smart meters measure electricity that comes from the grid. The electricity, which is produced by household/business entities and which is directly consumed by themselves, does not move through smart meter and is not registered. Missing own produced and consumed electricity will underestimate final electricity consumption (by households) and electricity intermediate consumption (by business). It will hurt comparability over time because own production of electricity tends to grow. Hopefully smart meters will get even smarter in the future and will capture the electricity produced and consumed by households and business.

Smart meters could be considered as a stable source of information and this is good for comparability over time and across countries. However, during the roll out phase, when the number of smart meters is continuously increasing, the comparability of statistics based solely on smart meters is not ensured.

WP4) AIS data

There are a lot of organisations (for example Dirkzwager, Marine traffic and national coastguards) which can provide AIS data for pilot studies. Also getting long term access to national AIS data should not be a problem. For example Norway, The Netherlands and Denmark already have continuous access to national AIS data. The big advantage of AIS data (compare to other data sources) is that it is available on European level, but long term access to European AIS can still not be ensured. A future risk is that AIS communication could possibly be substituted by Internet/WIFI-, as most ships nowadays have equipment to be online and exchange information through the Internet.

WP5) Mobile phone data

NSIs and mobile phone operators have to find common interests to enter an agreement to produce statistics based on mobile phone data. There are both legal and ethical barriers to mobile phone data access. Mobile phone data are sensitive for subscribers and EU strongly defends privacy of data. We need a long-term access to phone data to evaluate the full list of factors that affect data quality: children phones are registered to their parents, some persons may have more than one phone, etc. (*see: Braaksma B, Daas P., Struijs P., (April, 2014) Official statistics and Big Data*). For the time being long term access to phone data is far from being ensured.

WP7) Multiple statistical domain

This work package investigates in depth three domains of interest: sentiments revealed in the social media; agriculture; tourism/border crossing.

The comparability issue is different for the above mentioned domains. Data sources could be considered stable across three abovementioned domains. However stability of source is not sufficient to produce comparable data over time. Sentiment assessment is based on analysis of posts in social media. However, the number of social media active people could change for many reasons: shift to another media, no lasting interest in posting, etc. Changes of population that post in the media could undermine comparability of satisfaction assessment. Another factor that could undermine comparability is related to technological changes. Recently Facebook announced change in algorithm that prioritises the posts, videos, and photos that appears on the users' News Feeds. Users will see more posts from friends that spurred lively debates and less videos, publications and offers to buy nice 4G mobile. The idea is to encourage FB users to participate in debates and to spend more time on sites they find useful (see: Facebook is changing. What Does That Mean for Your News Feed? The New York Times (12 Jan., 2018)). It is reasonable to assume changes will make contacts and debates more focused and intensive. It could affect sentiments as well. This is an example of undermining sentiment assessments comparability driven by technology change.

In the field of agriculture WP7 combines satellite images of land lots with administrative data and survey data about agricultural land. The results so far are encouraging. The EU is very particular on agricultural policy and has built strong institutions to implement union regulation. Both sources of data - administrative and satellite data - could be considered as stable and this is a prerequisite to have EU wide comparable data about land use and crop.

In the Tourism/border crossing domain WP7 developed a methodology to measure intensity of border traffic. Data about border traffic could be compiled via traffic sensors in Poland and neighbouring Schengen member countries. Traffic data owners are government agencies and they have an interest to combine national data with their neighbours' data to have reliable border traffic statistics. This ensures comparability of data over time and across countries.

2.3.3. Discussion

Big Data are a technology driven phenomenon and changes in technology are the most important factor that affect comparability of BD based statistics over time. Technological development motivates more companies to hire online; people to spent more time in lively discussion and less

time in pressing like/don't like button; more smart meters to be installed to measure and bill electricity consumed; etc. Technological changes will affect comparability over time of all seven WPs. Moreover, technology will affect coverage over time as well as representativeness over time. It is reasonable to think that technology savvy persons will be the first to use new opportunities, whereas not so advanced persons will follow them later on.

Moreover, technological changes affect the comparability between countries. Inter-country variation is caused not only by changes in the object of interest (online hiring, peoples' sentiment, online retail trade, etc.) but by new opportunities offered by technology. Technological changes do not take place simultaneously in all countries, but spread in an irregular way: they appear first in a small number of countries and later on penetrate other countries. People also accept technological changes in a different way. Summing up, part of inter-country variations could be attributed to technological changes and spreading of changes across countries.

There is another factor that undermines comparability across countries: differing legal restrictions about the access to the data source. The preliminary analysis in WP1 about the legal framework of web scraping identified more or less explicit restrictions techniques for web scraping. (WP1, Deliverable 1.2). These different legal restrictions across countries lead to less comparable data across countries.

The comparability over time issue is also connected with representativeness. Scrapping social media is a good example. There is a natural bias toward social media active people when assessing sentiments. Maybe it is good to have profiles of different groups active in social media and to track profile changes over time. This is a prerequisite to have an idea about selectivity and how to improve Big Data comparability.

Another important factor for comparability over time is data ownership. Most BD comes from sources owned and administered not by the NSI but by someone else. The owner could allow access to the data for some time, but could decline the access to the data as well after a certain period of time. In most cases data owners are interested in research to understand better their market and how to improve their business model. They are not interested in providing statistics on regular basis. The most important prerequisite for data comparability, long term access to data, is not always ensured.

2.3.4. Literature

Eurostat (2014) European statistics code of practice. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice (accessed 22 May 2014).

Braaksma B, Daas P., Struijs P., (April, 2014) Official statistics and Big Data , SAGE Journals. Available at: http://journals.sagepub.com/doi/full/10.1177/2053951714538417

You might find your next job on Facebook, The Telegraph (28 Feb., 2018). Available at: https://www.telegraph.co.uk/technology/2018/02/28/might-find-next-job-facebook/

Facebook is changing. What Does That Mean for Your News Feed? The New York Times (12 Jan., 2018). Available at: https://www.nytimes.com/2018/01/12/technology/facebook-news-feed-changes.html

2.4. Linkability

2.4.1. Introduction

It is to be expected that Big Data needs to be linked or combined with other data sources. During this process, errors may occur which affect the quality of the output.

Official statistics is usually based on units (so called statistical units) which are clearly predefined by some legal basis. This holds especially for European Statistics, where we rely on agreed definitions coming from European regulations. The nature of Big Data is that they are not made for statistical purposes and that they are often of different types (sensors, scraped data etc.). Therefore the assignment to the statistical units relevant for a statistical product and the linking/merging to other relevant datasets can be challenging. Even if the data are of the same type they may have been collected by the help of different methods or definitions, so producing combinations is not necessarily straightforward. Linkability assesses the difficulty of the task of combining different data sources.

Linking different sources is a process with different steps that can be divided into three groups of tasks: schema matching, data matching and data fusion.

Schema matching refers to identifying database tables, attributes and conceptual structures from disparate databases that contain data, which correspond to the same type of information (Christen, 2012).

Data matching (also record linkage) is the task of identifying and matching individual records from disparate databases that refer to the same real-world entities or objects. To be easily linked, data should have the same format, structure and content, which is usually not the case and is making the process more difficult. Therefore, data processing (cleaning, structuring, standardization etc.) is crucial for successful data matching (Christen, 2012). The most straightforward version of data linkage is linkage by a unique identifier that is “shared” by different data sources. Unfortunately, the existence of such common unique identifiers is rather the exception than the rule.

Data fusion is the process of merging pairs or groups of records that have been classified as matches (i.e. that are assumed to refer to the same entity) into a clean and consistent record that represents an entity. In order to have a result of high quality, classifications of entities have to be good. Due to the lack of training data with true match status and large populations when working with Big Data, there is no guarantee that two records correspond to the same entity. This makes linking data harder and makes quality assessment more difficult (Christen, 2012).

When linking data, each record pair falls into one of four categories:

- True positives: record pair has been classified as match and is a true match.
- False positives (or false matches): record pair has been classified as match, but is not a true match. The two records in these pair refer to two different entities.
- True negatives: record pair has been classified as non-matches, and is a true non-match. The two records in pair in this category do refer to two different real-world entities.
- False negatives (or false non-matches): record pair has been classified as non-match, but is actually a true match. The two records in these pair refer to the same entity.

There are various aspects of how to assess the quality of linkability:

- What is the quality of the variables used to perform linkage? Are the kinds of variables needed to perform linkage even present in the data sets?
- Are there unique entity identifiers? One has to be absolutely confident in the accuracy, completeness, robustness and consistency over time of these identifiers, because any error in such an identifier will result in wrongly matched records. Therefore linkability is closely connected to those other aspects of quality.
- If no entity identifiers are available in the databases to be matched, then matching needs to rely on the attributes that are common across the databases.
- What is the level of linking (the level at which linkage must take place); eg. level of classification.
- What is the success of record linkage in terms of percentages of linked/unlinked records? How much of them are strongly linked (deterministic linkage methods were used), and how much softly linked (probabilistic linkage and machine learning methods were used)? How confident are you that data are matched correctly?

There are other, more exact quality measures for data linkability that can be calculated based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Those are Accuracy (not the quality dimension), Precision, Recall, Precision-recall graph, Precision-recall break-even point, F-measure (or F-score), Maximum F-measure, Specificity, False positive rate, ROC curve, AUC (Christen, Goiser, 2007).

2.4.2. Examples and methods

The Table 3 below illustrates what were some of the Big Data sources and how were the WPs linking them in their work. In each row a type of Big Data is presented with additional information:

- Entities represent target variables that we are assessing using Big Data and to which Big Data observations have to be linked. For example when scraping job vacancies ads, the goal is to find the number of ads for each company through information from the internet, therefore we have to link names as they appear on the internet to official company information.
- Additional sources are in-house data that are available at the same time as the main Big Data source, and are used to facilitate linkage or used as one of the source when estimating with a combination of sources. E.g., scraped job titles and descriptions do not usually match to the official ones and classification codes are not included, so we classify them through the ISCO classification and then to the official names/titles. When using traffic sensor data as a regressor in regression model, additional sources may be used as well.
- Entity identifiers are specific standard variables which are used to link Big Data information to other data. These might be classification codes or geographical position and/or selected time points etc.
- Methods of linking are algorithms with which the linkage is executed. This may be accomplished by simple rules to complex text-mining record linkage techniques. Other machine learning classification/regression operations are common as well, they are more in detail discussed in the Methodology Report.

Type of (Big) data	Entity	Additional source	Entity identifiers	Method of linking
Scraped job vacancies adds	Company name	Administrative business register	Registration number	Record linkage
Scraped job vacancies adds	Job title/job description	ISCO classification	Occupation code	Machine learning - classification methods
Road sensor data	Road sensor	Survey data on turnover in industry/ retail trade	Time period	Rule based
Retail scanner data	Commodity group	COICOP classification	COICOP code	Rule based
Smart meter data	Address of metering unit as text	Address information system		90% Automatic text, rest manually
Smart meter data	Metering unit	Administrative data (business register, dwelling register)	Registry codes or address ID-s	
AIS data	Type of vessel	Port visit statistics	IMO number	
AIS data	Reporting port	Selected EEA ports in the port list	Latitude/Longitude	
URLS of Enterprises	Company	Business register	Register number, name, address (depends)	Machine learning

Table 3 List of examples of linking in the WPs

2.4.3. Discussion

When we are thinking about combining different sources, we can do that for two purposes. One is to add additional variables to existing observational unit/entity and extracting statistics from there. The other one is to use different data sources, combining them on an aggregated level and then measuring some kind of concept, but not an entity itself.

Using Big Data might create an assignment problem. Statistical products in official statistics normally rely on predefined statistical units. Therefore, before a possible integration can take place, it must be clear how the data from new data sources can fit into the concepts relevant for the statistical units.

To assess the possible impact on quality of combining different (types of) data sources one must look at each step of the production process. Although combining different data sources affects all quality dimensions, the most significant ones are accuracy and comparability. For example, when linking and determining the target population, the risk is that units will not be linked or will be linked wrongly, which will turn out as under/over coverage (see Coverage in Section 2.1). However, problem of comparability can be reduced to the structural errors generated by some possible statistical biases. On the other hand, the problem of accuracy is bigger. Even if the accuracy of the separate sources of data can be measured, assessing the sensitivity of the accuracy of the final combined data set to the source-specific errors and the integration methods can be very difficult (Vaju et al., 2015).

The quality aspect of Linkability in this quality report is connected to the aspect of Data source integration in the IT report and the aspect of Data linkage in the Methodology report.

2.4.4. Literature

Christen, P. (2012). Data Matching Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection. Springer.

Christen, P., Goiser, K., (2007). Quality Measures in Data Mining: Quality and Complexity Measures for Data Linkage and Deduplicaton. Springer.

Hand D. J. (2018). Statistical Challenges of administrative and transaction data. *Journal of the Royal Statistical Society.*

Vaju, S., Agafitei, M., Gras, F., Kloek, W., Reis, F. (2015). Measuring the quality of multisource statistics. *In New Techniques and Technologies for Statistics 2015*, pp. 456-459. European Commission.

2.5. Processing Errors

2.5.1. Introduction

During the processing of Big Data, errors may be introduced that negatively affect the quality of the data. Examples of this are the way outliers and missing values are treated.

Processes involved for producing statistical output with the help of Big Data sources are as varied as the Big Data sources itself, and thus also the potential processing errors are very diverse. The expression "Processing Errors" itself is not completely clearly defined either. Some error types are counted among processing errors by some authors, other authors do not list them among this category (e.g. imputation errors). Further, the expression "Processing Errors" was defined – as most elements of the usual quality framework for official statistics – for statistical processes based mainly on surveys. The quality framework was gradually extended when multi-source statistics, which are based on more than one data source (e.g. survey data and administrative data sources), became increasingly important. For multi-source statistics, additional processing errors can occur, which did not play a role for unique data sources (e.g. linking errors, errors when integrating data). Still, we start from the (relatively) well-designed definition and classification of processing errors for surveys. Using this as a reference framework, we will examine which kind of processing errors can be simply adopted for Big Data sources, which kind of processing errors need a different kind of treatment for Big Data sources and which kind of processing errors for Big Data sources are missing in a unique-source and survey-based framework.

In "Overview of Design Issues: Total Survey Error", the author Paul P. Biemer names processing errors as one of five major sources of non-sampling errors (the other sources include specification errors, frame errors, nonresponse errors and measurement errors), (c.f Beamer 2010). He states five major sources for processing errors (in the context of a survey as single data source)

- **Editing:** Data editing is the activity aimed at detecting and correcting errors (logical inconsistencies) in data, see OECD's Glossary of Statistical Terms.
- **Data entry:** Data entry is the technical procedure to convert available information to a digital format so it can be further processed, see Statistics Austria's Quality guidelines.
- **Coding:** Coding is the technical procedure for converting (verbal) information into numbers or other symbols which can be more easily counted and tabulated. A coding error is the assignment of an incorrect code to a survey response, see OECD's Glossary for Statistical Terms.
  In (Beamer and Lyvberg, 2013), the authors describe **Automated Coding** as follows:
  Automated coding means that a computer program attempts to match response descriptions to computer-stored nomenclature descriptions (dictionary) and then assign code number when matches are acceptable according to specified matching criteria.
- **Weighting:** The survey weights compensate statistically for unequal selection probabilities, nonresponse errors and frame coverage errors. When they are calculated erroneously, this leads to weighting errors.
- **Tabulation**: Errors in the tabulation software may also affect the final data tables.

Depending on whether "correcting errors/logical inconsistencies in data" is counted among the process step of editing or not, also the process step of **Imputation** can be seen as source for processing errors:

- **Imputation**: Imputation is a procedure for entering a value for a specific data item where the response is missing or unusable3. Imputation errors occur when correct values are replaced erroneously by imputed values or when imputed values themselves are incorrect.

Already at this point, it is clear that some of these error types do not play a (specific) role for Big Data sources. Since Big Data is already in a digital format, the process steps of Coding and Data entry (in its original meaning) fall away. Tabulation does not seem to involve any process steps differing in case of Big Data sources and will thus not be dealt with in the following.

Leaving the single-source framework, also the process step of **Linkage** can cause processing errors when Big Data sources are combined with other data sources such as administrative data or survey data. Linkage errors occur when it is not possible to merge two data sets via exact record linkage (meaning there exists a key variable) and methods such as textual comparisons or statistical matching have to be applied. We will not elaborate on the topic of linkage in the course of the Chapter Processing Errors, because this is covered in **Chapter Linkability.**

Metadata errors are one category of errors, which has become more prominent since statistical institutes have been working with data sources where statistical institutes are not the data owners. The most prominent example for such data sources with data owners other than the statistical institute are **administrative data sources.** Also in the case of Big Data sources, the statistical institutes are (almost) never the data owners. Obtaining the accurate metadata from the data owner can thus be another source for errors. (Japec et al 2015) names the example of inaccurate or erroneous labelling of column data (the variables) in a Big Data set. The role of good metadata is also highlighted in Chapter 3.2 “Metadata management” of the IT report, also created in WP8 of the ESSnet Big Data.

By definition, processing errors include all post-collection operations. When Big Data sources are involved, this definition can still lead to problems, because the collection phase itself can be rather vaguely defined. Think for example of social media data (e.g. tweets) which are used to extract the general sentiment out of them (see WP7a). Another example is the web-scraping of enterprise websites to extract enterprise characteristics (see WP2). As described in (Japec et al 2015), **multiple layers of processing can be required until a data set is produced**. Using the example of extracting sentiments (about the economy) from tweets, this transformation process to obtain a workable data set involves e.g. parsing phrases, identifying words, classifying them as to a subject matter and then further as to positive or negative about the economy.

Each of these processing steps can introduce all kinds of source-specific processing errors. As noted at the beginning, the diverse nature of Big Data sources leads to completely diverse potential processing errors, which can hardly be described in a generic way. In analogy to the survey-data framework, where **data entry** was defined as the technical procedure to convert available information to a digital format so it can be further processed, one could subsume these Big Data - processing steps needed until one has a workable data set as **errors related to (Big) data entry.**

[3] See the OECS’s Glossary for Statistical Terms: https://stats.oecd.org/glossary/detail.asp?ID=3462

(Japec et al 2015) provide a Big Data Process Map, where they distinguish between three general stages: 1) Generate 2) Extract/Transform/Load (ETL) and 3) Analyze. The authors explain these three stages as follows:

> In the generate step, data are generated from some source either incidentally or purposively. In the extract/transform/load (ETL) step, all data are brought together under a homogeneous computing environment in three stages. These stages are the extract stage, where data are harvested from their sources, parsed, validated, curated, and stored; the transform stage, where data are translated, coded, recoded, aggregated/disaggregated, and/or edited; and the load stage, where data are integrated and stored in the data warehouse. In the last step, analyze, data are converted to information through a process involving two stages. The first stage is the filtering (sampling)/reduction stage, where unwanted features and content are deleted; features may be combined to produce new ones; and data elements may be thinned or sampled to be more manageable. The second stage is the computation/analysis/visualization stage, where data are analyzed and/or presented for interpretation and information extraction.

The authors visualize the flow of data along these steps as shown in Figure 5, taken from the paper (Japec et al 2015)

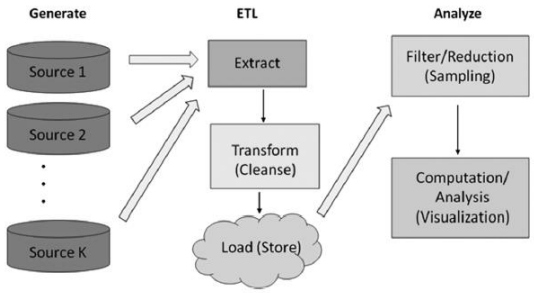

Figure 4. Big Data Process Map (graph created by Paul Biemer).

Figure 5,Illustration of the Big Data Process Map, taken from (Japec et al 2015)

Looking at this process map, it is clear that Processing Errors as we describe them will appear mostly in the "ETL stage". The authors name the following errors which can occur in the ETL stage: **specification errors (including errors in metadata), matching errors, coding errors, editing errors, data munging errors, data integration errors.** These error sources are very similar to the ones already considered by us.

Comparing these three stages with the five major sources for processing errors in (Beamer 2010), where we started from, we see that there is an overlap between the sources of processing errors and the "Analyze stage". The authors name the following errors which can occur in this stage: **modelling errors, inadequate or erroneous adjustments for representativeness, improper or erroneous weighting, computation and algorithmic errors.**

Similar to surveys, **weighting errors** can also play a role when working with Big Data sources, since weighting procedures can be necessary to encounter problems with coverage and selectivity. For more information, we refer to **Section "2.1 Coverage, Accuracy and Selectivity".**

Editing is a source for process errors which is relevant for survey data, administrative data and Big Data sources alike. In (De Waal et al 2014) and in (Puts et al 2015), the authors address specifically the topic of data editing of Big Data. The authors describe several alternatives to the lengthy **manual editing,** which were developed also for administrative data source so as to improve the efficiency and effectiveness of statistical data editing. (Puts et al 2015) give a quick overview of the editing forms from a "small data perspective"

- **Interactive Editing** is also a form of manual editing, where in principle all records are examined and if necessary corrected, but the effects of adjusting the data can be seen immediately on a computer screen
- **Selective Editing** identifies the records with potentially influential errors and restricts interactive editing to those records only
- **Automatic Editing** is done by computers without any human intervention. Random errors can be detected by outlier detection techniques or by deterministic checking rules.
- **Macro-editing** checks whether the data set is plausible as a whole.

When turning to Big Data, the authors in (De Waal et al 2014) first note that traditional editing techniques can sometimes be applied to smaller data sets which are the outcome of multiple layers of processing of a Big Data source.

Further they note that **manual and interactive editing** are for most Big Data sources impractical and often even impossible due to the sheer volume of the data sets. The authors also note that substantive subject-matter-knowledge is often lacking for Big Data sources and it will be difficult to specify edit rules that are suitable to guide through the editing process.

Next, the authors describe the pitfalls for **selective editing** of Big Data sources: So as to be able to measure the influence component one needs to compare values with **an estimated total**. For many Big Data sources, such estimated totals are not available due to a **selectivity bias** in the Big Data source.

Also the applicability of traditional automatic editing techniques can be limited for Big Data sources, because they often require detailed knowledge on the structure of the data, which is often lacking for Big Data. Further, defining a detailed set of edit rules required for automatic editing is expected to be challenging for Big Data.

The authors also name a general editing technique which can work for Big Data sources: **Automatic detection of outliers or suspicious observations** and then deleting these observations can work if there are only few outliers or suspicious observations.

Macro editing can work for Big Data sources as well. The authors name two specific methods, namely the **Aggregation method** and the **Distribution method**. For the latter, they also present a visualisation technique called **Tableplot** which is suitable for Big Data (see Examples and Methods).

As a conclusion, the authors state that most traditional editing techniques are prone to fail when applied to Big Data sources. Instead, it is necessary to develop **taylor-made solutions.**

We note at this point that missing observations are quite common for Big Data sources, but similar to automatic editing, automatic imputation methods are likely to fail. For example, the popular k-

nearest neighbour imputation method will often not be applicable because we have no further information about the population units.

2.5.2. Examples and Methods

Examples and Methods from the Literature

A Markov Chain Model for Traffic Loop Data

In (De Waal et al 2014, Puts et al 2015) the authors describe a taylor made solution developed by them for editing traffic loop data. The high frequency of the data enabled them to apply signal processing techniques for editing and imputation purposes. In particular, they estimated a Markov chain model for each road sensor.

In (Puts et al 2015) the Markov chain model to correct for missing observations is described as follows, also the Figure below is taken from there:

> In Figure 3, Y_t (t = 1, 2, …) denotes the observed signal at time t, that is, the observed (but possibly incorrect) number of vehicles that passed the sensor during the last minute before time t, and X_t the true (unobserved) signal, that is, the true number of vehicles that actually passed the sensor. The observed data Y_t (t = 1, 2, …) is used to estimate the transition probabilities to go from one state X_t to the next X_{t+1}. The most common kind of error that occurs in road sensor data is that observations are missing due to the fact that the sensor is temporarily not working properly. The Markov chain model can be used to automatically correct for this kind of error. Namely, in cases where the observed signal Y_t is missing, the Markov chain draws a value for X_t using the previous true state X_{t-1} and the estimated transition probabilities.

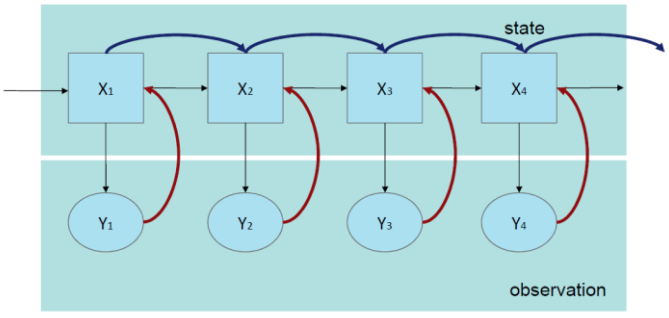

Figure 3. A Markov chain model for road sensor data

Figure 6, Illustration of a Markov chain model for road sensor data, taken from (De Waal et al 2014)

Tableplot, a Method for Detecting Errors in Big Data Sources

(Puts et al 2015) and in more detail (Tennekes et al 2013) describe the visualization tool **Tableplot.** Tableplot is a graphical macro-editing tool which can detect implausible or incorrect values, and it also monitors the effects of the editing process. Tableplot is especially suitable for (very) large data sets. The authors describe the mode of operation as follows (c.f. (Puts et al 2015), also Figure 7 is taken from the paper)

In a tableplot, a quantitative variable is used to order the data for all variables shown. The ordered records are divided into a certain number of equally sized bins. For each bin, the mean value is calculated for numerical variables, and category fractions are determined for categorical variables, where missing values are considered as a separate category. These results are subsequently plotted. A disruptive change in the distribution in a tableplot can indicate the presence of errors. Moreover, a non-uniform distribution over the columns can indicate selectivity. Finally, the distribution of correlated variables can be examined by looking at the value distribution in the unsorted columns.

Figure 7 [...] shows a tableplot for the Dutch annual Structural Business Statistics (SBS), based on unedited […] data […]. These relatively small data sets – in comparison to Big Data, that is – are used to illustrate the benefits of applying visualisation methods for monitoring the editing process. [...] Figure 7 was created by sorting on the first column, "turnover", and dividing the 51 621 observed units into 100 bins, so that each row bin contains approximately 516 records. A subset of approximately 49 000 records was deemed suitable for publication purposes.

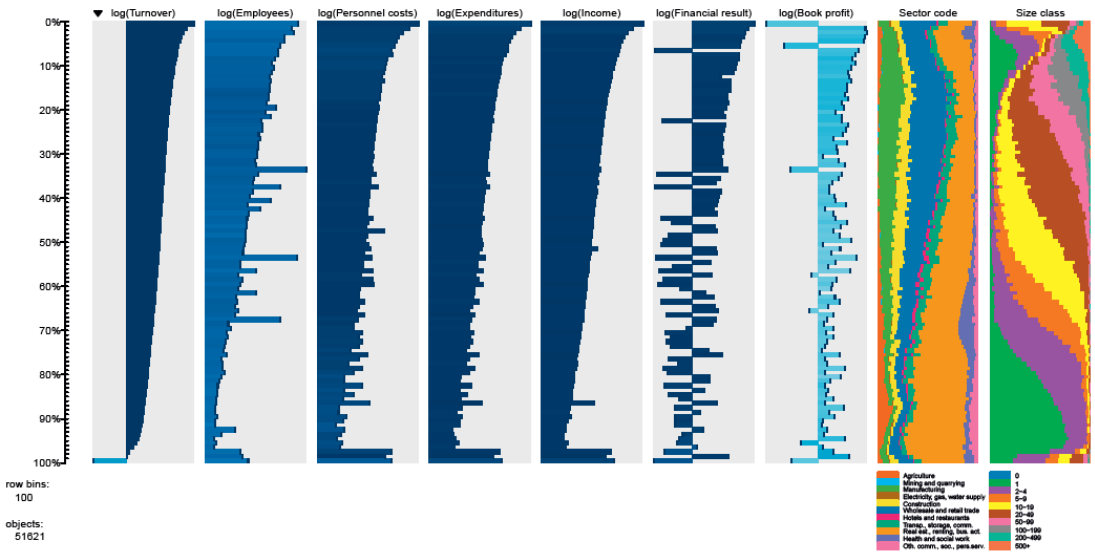

Figure 1. Tableplot of unprocessed SBS data. When sorted on turnover (left-most column) a considerable number of the other numeric variables display a clear – and predictable – downward trend occasionally distorted with large values[4]

Figure 7, Illustration of the visualization tool "Tableplot", taken from (Puts et al 2015)

Examples and Methods from WPs 1 to 7

WP1) Job Vacancy Web Scraping - Information from the Final Technical Report from SGA1

As typical for Big Data sources, multiple layers of processing were required until a workable data set is produced.

The raw data obtained from a job advertisement generally requires a lot of processing before it can be analysed. For example, job title fields often contain extraneous information, such as job location, key skills, and salary. This is because employers try to attract potential job

seekers by stacking the job title field with other key information like skills required and salary.

De-duplication seemed to be a major issue during the processing phase:

> Duplication of job ads is a fundamental quality issue with multiple job portals. It can also be a problem within portals, particularly for job search engines that pick up job ads from other portals. Job search engines may take steps to remove duplicates but the effectiveness of these procedures is often variable. Duplication methods were explored as part of a "virtual sprint" held on 28-29 July. Essentially, these methods involve matching common fields, comparing text content and then calculating a similarity metric to establish the likelihood that two job advertisements are the same.

Linkability and Data Integration: WP1 scraped job vacancies on several different job portals, and tried then to integrate the data from the different sources. Thereby, data integration errors could have happened. Further, they also tried to link the data gained from web scraping with the data from the Job Vacancy Survey (JVS):

> Four of the country pilots have explored the matching of online job ads with their own JVS micro data as a means of starting to understand coverage issues. The results have been somewhat mixed. Matching rates between data from the Swedish Employment Agency (SEA) and the JVS has been very high since SEA job ads contain the same organisation identifier as used in the JVS. In contrast, other matching has been done on company name only, which has proved much more difficult. Common problems include use of abbreviated names, trading names rather than the legal enterprise name, and misalignment between the company names and the JVS reporting unit.

WP2) Webscraping of Enterprise Characteristics, all information from Deliverable 2.2 Methodological and IT Issues and Solutions

Both, generic scraping of whole websites and specific scraping of predefined content was used to collect information about the following use cases

1. Enterprise URLs Inventory.
2. E-Commerce in Enterprises.
3. Job vacancies ads on enterprises' websites.
4. Social Media Presence on Enterprises webpage

The members of the WP2 designed a generic reference logical architecture made of several building blocks organized into the following four main layers: Internet access, storage, data preparation and analysis. "Data preparation" is the layer which is interesting in the context of processing errors. In the data preparation layer, the members of WP2 described the following "blocks", which can be seen as data process steps, in each of them errors could occur, but have not been specified in the report.

> The **Feature extraction** block is responsible for localizing and retrieving from a scraped resource a set of predefined features of interest (e.g.: addresses, telephone numbers, names, VAT codes, etc). Usually it is implemented in a SW program.

The URL scorer block is used to assign a score to an URL on the basis of some predefined parameters such as the presence of some features of interest inside the URL's content. Given a list of URLs related to an enterprise, this block can be used alone or in conjunction with other block in order to identify the most probable official URL for that particular entity.

The **Tokenization block** processes the textual content of the scraped resources by transforming it in a text that becomes input for further processing such as parsing and text mining or for analysis blocks. Normally, in lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

The **Data parsing block** focuses on the analysis of the tokens produced by the Tokenization block by searching for specific regular expressions, matching sentences etc.

The **Word filters block** is used to filter out some words/tokens (if a list of words to be filtered out is provided) from the scraped textual content or to enrich it with a list of go words.

The **Language specific lemmatization block** lemmatizes the tokens found in the scraped textual content in order to reduce the number of textual elements to be analyzed. Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form. In this case (computational linguistics), lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. When it is not possible to desume the intended meaning usually the base form of a token is obtained by using a stemmer that compute the base form of a token by operating on a single word without knowledge of the context.

The **Term document matrix generation block** is responsible for producing a TDM (Term Document matrix) to be used by analyses blocks. Normally each cell of the matrix contains the number of occurrences of a token in an enterprises' website

WP3) Smart Meter Data

For the Estonian smart meter data, the following editing steps were described in the Deliverable 3.2:

> Due to the vast amount of data, fast and efficient algorithms have to be developed to identify and correct possible problems in the data set. For detecting errors one needs to determine set of editing rules that usually include rules to detect outliers. However, in practice it is very difficult to come up with editing checks because the data hub description is rather brief and variable' descriptions are lacking. [...]
>
> In general, not many errors were discovered in Estonian data. In Estonian data there were two cases when the end date of the grid agreement differed from the supply agreement's end date. The dates were corrected when the table was transferred into Hadoop file system. Also, there were two duplicate agreements in the initial dataset, of which one was deleted during transfer. No other changes were made in the metering data.

For Estonian smart meter data, the following processing steps were necessary before the use of smart meters data for statistics as described in Deliverable 3.1

- Geocoding of metering point addresses
- Transformation of the timestamp of metering fact into readable timestamp format,
- Anonymization of private personal data in the customers table

WP4) AIS Data – Deriving port statistics and linking data from Maritime statistics with AIS-data

In the Deliverable 4.2, the authors describe all kind of technical errors as well as human errors which can be found in the AIS data. The authors describe that many measures were implemented to detect and remove erroneous data; these errors are also listed in Chapter on "Measurement Errors":

> For example, unreadable messages are already filtered out during pre-processing. Issues concerning the ships' identity we deal with by constructing a reference frame.

The authors further suggest building a common database of known errors which is then used for validation, processing and analysing data.

Deliverable 4.3 describes that WP4 also involved some validation checks of AIS data by comparing data from different providers.

WP6) Early Estimates:

WP6 has developed a nowcasting model so as to calculate early estimates of the time series of interest (e.g. GDP, Industrial Production Index, Consumer Confidence Index) with the help of enterprise microdata (and potentially additional Big Data sources such as traffic loop data):

1) Principal Component Analysis (PCA) is used to extract principal components from enterprise data (or other microdata). For each enterprise included in the model, time series of data without any missing values is needed. Then, first few principal components are chosen. PCA is one of the most used methods to summarize most of the information contained in the microdata with a small number of variables (first few principal components).
2) Linear regression is used: the time series of interest (e.g. GDP) is the dependent variable (Y) and the chosen principal components are the predictors (X_1, ..., X_n). Seasonal component and other predictors can be added (e.g. predictors calculated from traffic loop data). Linear regression is used as one of most used methods in statistical learning to estimate a continuous dependent variable with the help of continuous predictors. Continuous estimates of the dependent variable are obtained and regression coefficients measure the relationship between the dependent variable and the predictors. Also some other methods were tested.

In the Slovenian use case of WP6, traffic sensor data was used as additional regressor in the nowcasting model. The Slovenian colleagues contributed the following description of the processing steps involved with the traffic sensor data and potential errors specifically for this report:

The data are already partly edited: the actual files contained aggregated numbers of multiple sensors on the same stretch of road if they existed. These objects are named **count spots**.

- **Editing**: the data that each count spot represents is organized into two channels: each way on a road is represented by one channel. However, the sensors only have the ability to count on two lanes at most; therefore the sensors on multilane speedways are only able to count one way. Where available, count spots contain both lanes, however some sensors only count the traffic in one way and the data for the other way is non-existent. Their count spots are filled in zeros for the corresponding channel. All of this has to be taken into account when analysing data, and has to be reliably detected.
- **Weighting:** After analysing the data, it was discovered that many count spots lack longer periods of data (such as weeks, months or even whole years). This may be because the count spots were not established yet or they were malfunctioning and turned off for a period. It was decided upon using only those count spots that contain at least 85% of the whole date and have no missing years. Weights were therefore 0 and 1. As newer data were gained this may lead to include newer count spots and exclude old ones.

  The errors considered here are whether excluding count spots in this way may lead to losing information about traffic in part of the country. This was decided to not be an issue on regional roads, where traffic is similar in scale and shape, but might represent problems on speedways which were already under-represented. However to us this is a minor issue, as it was discovered that speedways do not play a big part in our chosen target variable assessments and actually more often than not return worse results.

  Additional weighting happens in the PCA and linear regression process when combining traffic sensor data with other data sources for the purpose of assessing some target variables. These weightings are fully automated by the process and dependent on the parameters used. The traffic sensors are used in two ways. Firstly they can either play the part of main regressors (usually with other data sources), in which case the PCA method selects the most important count spots through their principal component weight. Secondly they can be secondary regressors, where they are summed together to form one time series that is used fully in the regression.

  In both cases the regression adds additional weights to the regressors that explain how big their part in the assessment of the target variable is.
- **Imputation**: Imputations have received a big consideration in our work. As said above the focus was on regional roads and the imputations were tailored towards them. The imputations were used when missing data was detected and only after aggregation to a monthly basis was done. This was done in order to avoid precision errors due to small scale data (one to two cars more in one period may represent growth of 100% in the mornings, while this may only be 1-2% growth in the peak hours) and to limit the data size on disk to manageable degree. Imputations of a count spot data are based on the growth of neighbouring count spots.

  The possible errors that were detected or anticipated were errors that happen due to algorithmic errors and errors that occur due to the differing scales of traffic in count spots in comparison to their neighbours' traffic. An extensive analysis of scale and shape of data and performance tests of four different algorithms were concluded.
- **Metadata Errors**: Metadata errors caused quite a lot of problems. Even though the received data were semi-prepared, the process of preparation clearly had some issues of their own, as quite some errors occur in the metadata. The first problem was indiscriminate use of spaces

as both delimiters in files as actual spaces in variable names. The format of other metadata changed throughout time such as:

- time periods of traffic counting (our choice of the earliest year of obtained data as 2011 was due to the changing time periods of counting from 10 minutes to 15 minutes which occurred inconsistently in count spots data during the previous years),
- different notations of one-way counting on speedway roads in the files,
- errors in the time notations (from entirely missing time periods to wrong time periods; e.g. one time interval was given at 41 minutes past a full hour when every other interval was strictly at 0, 15, 30 or 45 minutes past an hour),
- in one instance the variable names were inserted in the middle of data,
- some information was given in non-structured format and had to be extracted using regular expressions (for example types of sensor that make up a count spot; whether the count spot represented regional or speedway sensors, etc.).

What was gathered during processing of data and the errors is that quite a lot of errors arise from specific characteristics of given datasets (metadata errors, weighting circumstances, editing) and cannot be generalized to the whole or at least to a big part of the Big Data field. The solutions to most of these individual problems need to be specifically made for them. Lucky for us the source of traffic sensor is fairly consistent and the solutions we devised in this pilot can be applied in the future as well without (much) changes. Furthermore some of these problems are sensor-specific and while each problem needs to be addressed individually, they are part of a common set of problems that arise when working with other smart sensors as well. As such a lot of solutions on other smart sensor data can probably be adapted from todays (a good example are missing data due to malfunctions).

2.5.3. Discussion

"Processing errors" were identified as a one of the seven most important quality aspects in the ESSnet Big Data. However, the seven aspects listed in this report cannot be considered separately, but are interconnected or even overlapping. For example, procedures such as linking or actions such as weighting and de-duplication to counteract coverage and selectivity problems are described in the respective Chapters on Linkability (2.4) and Coverage, Accuracy and Selectivity (2.1), but seen from a business process perspective, they belong to the process stage, and errors occurring in this stage count as processing errors.

Further, as noted above, the process steps involved with Big Data sources differ considerably to process steps when working with Survey Data and vary also among different Big Data sources. So far, there exists no well-established quality framework for Big Data sources. The UNECE quality framework for Big Data (2015) structures its quality indicators along the three stages "Input", "Throughput" and "Output", and it is clear that "Processing" is part of the "Throughput" stage. Due to the huge variety of processing steps depending on the Big Data source, the UNECE quality framework is rather vague when describing the Throughput stage. Instead of naming specific quality indicators, they describe general principles such as "System Independence", "Steady states" and "Quality gates".

In this report, which covers the specific Big Data projects in WP1 to WP7, we chose another approach. We started from the well-established categories of processing steps and tried to figure out

which of these categories also play an important role for Big Data sources, and further, if Big Data specific categories of process steps could be identified and described. Based on these theoretical considerations we tried to describe in the Subchapter "Examples and Methods" specific processing steps in the work packages 1 to 7 and actions to counteract processing errors in the respective processing steps.

Still, we are aware that completely different processing errors could occur when working with other Big Data sources than in WP1-WP7, since completely different processing steps might be necessary.

2.5.4. Literature

De Waal et al (2014), *Statistical Data Editing of Big Data*, Paper for the Royal Statistical Society 2014 International Conference, Sheffield, UK

Puts, M. et al (2015), *Finding Errors in Big Data,* Significance 12(3), 26-29, DOI: 10.1111/j.1740-9713.2015.00826.x

Japec, L. et al (2015), *Big Data in Survey Research – AAPOR Task Force Report*, Public Opinion Quarterly 79 (4), 839-80

Biemer, P. (2010), *Overview of Design Issues: Total Survey Error,* in Handbook of Survey Research, edited by Marsden P. and Wright, J., Second Edition

Biemer P. and Lyberg, L. (2003*), Introduction to Survey Quality,* p 236, John Wiley & Sons

Tennekes, M. Et al (2013), *Visualizing and inspecting large datasets with tableplots*, Journal of Data Science, 11, 43-58

Daas, J. and Puts, M. (2014), *Social Media Sentiment and Consumer Confidence*, Statistics Paper Series, 5, European Central Bank

ESSnet Big Data, WP8 (2018), *Report describing the IT-infrastructure used and accompanying processes developed and skills needed to study or produce Big Data based official statistics*, Deliverable 8.3

ESSnet Big Data, WP1 (2017), *Final Technical Report (SGA-1),* https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/20/Deliverable_1_3_main_report_final_1.0.pdf

ESSnet Big Data, WP2 (2017), *Deliverable 2.2 Methodological and IT Issues and Solutions,* https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/66/WP2_Deliverable_2.2_2017_07_31.pdf

ESSnet Big Data, WP3 (2017), *Deliverable 3.2 Report on production of statistics: methodology*, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_2#Data_editing

ESSnet Big Data, WP3 (2017), *Deliverable 3.1 Report on data access and data handling*, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_1#Cleaning_data

ESSnet Big Data, WP4 (2017), *Deliverable 4.2 Deriving port visits and linking data from Maritime statistics with AIS data*
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/8/8d/WP4_Deliverable_4.2_2017_02_10.pdf

ESSnet Big Data, WP4 (2017), *Deliverable 4.3 Report about sea traffic analyses AIS data*
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_21_v1.0.pdf

Statistics Austria (2017), *Qualitätsrichtlinien* (Quality Guidelines), translated by Magdalena Six

OECS's Glossary for Statistical Terms: https://stats.oecd.org/glossary

2.6. Process Chain Control

2.6.1. Introduction

In a Big Data process it is very likely that multiple partners are involved. To assure a stable and timely delivery of data of high quality, the entire process needs to be controlled.

A process consists of several intermediate steps that, to achieve optimal performance, each need to be executed as well as possible. In a Big Data process the general processing steps Collect, Process, Analyse and Disseminate are discerned (IT-report, 2018). Controlling the overall process requires tight control of each step; certainly when large amounts of data are being processed. This is, for example, illustrated by the lessons learned in the system developed to process and analyse 650 TB of human genome data (Golman, 2018). Some state that because large amounts of data are processed a higher quality of the end product can be achieved and the process can be better controlled (Marr, 2015). However, to avoid paralysis, process chains need to process and deliver the right set of data and correctly identify potential problems. If not, the process may become very inefficient or may even grind to a halt.

2.6.2. Examples and methods

In each step in the process it is important to consider:

i) the amount, type and quality of the data needed to solve the problem
ii) the correct type of analyses, skills and tools required

Overall goal is to tackle the problem in each step in the most (cost) effective manner. For instance, when Big Data is being introduced into the process, it is essential to check the input quality of the data in the most efficient way. When only data of sufficient quality is used this assures that resources are not spent on (large amounts of) unusable data. The quality checks should be performed in a time and data efficient way. Data that does not meet the quality standard or cannot be edited to the quality standard envisioned should be ignored. The effectiveness of such an approach is illustrated by the quality checks included in the processing of road sensor data used for the Road Traffic Intensity Statistics in the Netherlands (Figure 6). When the day has ended, the data produced by each of the 20,000 sensors during that day is transformed and the quality of the data is expressed in 5 quality indicators per sensor (Puts et al., 2016). These indicators are implemented very efficiently and can therefore be determined very rapidly which hardly affects the overall processing time. The findings of these indicators are subsequently used to decide to in- or exclude each road sensor - for that particular day - in the remainder of the process. For instance, when a sensor has produced less than 720 measurements, effective half of the minutes in a single day, its data are not be used.

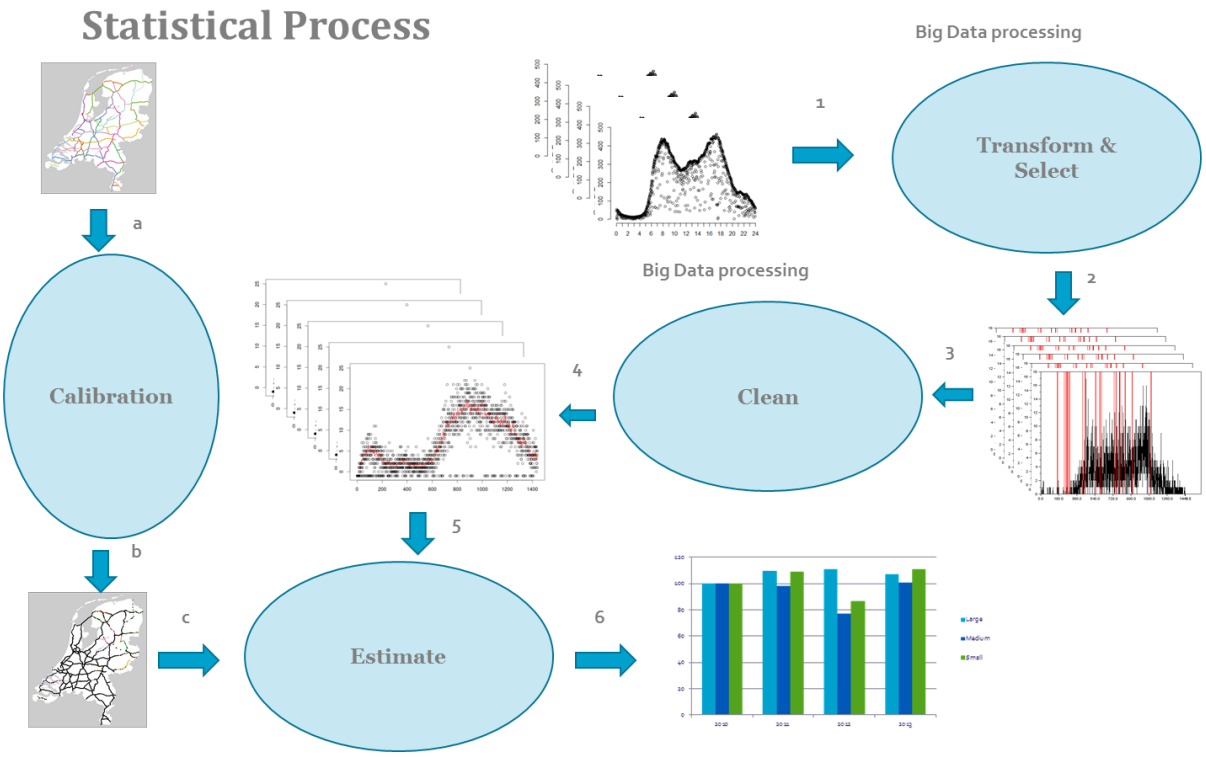

Figure 8 Illustration of the intermediate steps in the Road sensor based Traffic Intensity statistics process of Statistics Netherlands. The Collect step is not shown. The Process step consists of a Transform and Select sub step (preprocessing) and a data Clean step.

The overall idea of only including data that meets specific standards is repeated in each step of this process. For instance in the second step, the measurement values of the data are checked and missing data is imputed in a model-based approach. In another step, called calibration, the assignment of a sensor to a specific highway, road section and driving direction is meticulously checked. In the end, this approach resulted in a considerable quality improvement of the final output of the process. Difficulty with this data-driven way of working is that the effect of each quality decision cannot always be directly evaluated after each individual part of the process. This can only truly be evaluated when the sequence of the final process has become apparent.

At the moment of writing only WP2 and WP4 of the ESSnet Big Data have reached the stage that they approach the dissemination step of the Big Data processing life cycle. This means that in each WP the overall setup of the process is not known yet and limits this discussion to any relevant process evaluations of the intermediate steps. In WP2 the complete processing step for data scraped from enterprise websites is described (WP2 Del 2.2). Quality considerations are discussed in the methodological section of the report (chapter 4) and predominantly focus on the quality of the output and the training and test set; i.e. not on the individual steps. In WP4 Automatic Identification System (AIS) data needs to be (pre-)processed to be of sufficient quality (WP4 Del 4.1). Its quality is discussed in the next deliverables (WP4 Del 4.2, Del 4.3; Chapter 2 in both reports). The overall focus of the quality studies is on the coverage of the ships and journeys made and not on the stages in the production process.

2.6.3. Discussion

Big Data is increasingly used to control process chains either directly or indirectly. Because of the large volumes of data, it is important to use efficiently implemented quality indicators or predictors. The control of the Big Data process starts at the first step of the process (UNECE, 2014); which is to be expected when one works in a data-driven way. This introduces a considerable change within statistical organisations since these are -by tradition- predominantly focussed on the quality of the last step; the output. To assure high quality output when producing statistics based on Big Data, it is important that the quality of each step in the process chain is carefully controlled.

2.6.4. Literature

[1] Mayer-Schönberger, V., Cukier, K. (2013) Big Data, a revolution that will transform how we live, work and think. John Murray.

[2] Goldman, N. (2018) Genomics Research at EBI: Challenges in Statistical Scaling. Link: https://www.turing-gateway.cam.ac.uk/sites/default/files/asset/doc/1803/Goldman_TuringGateway.pdf, SLIDE 21-23.

[3] Hines, E. (2016) How Big Companies Use Big Data. Blog: https://www.fronetics.com/big-companies-use-big-data/

[4] Marr, B. (2015) Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. Wiley.

[5] 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.

[6] 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.

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

[8] 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

[9] 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

[10] WP4 Del 4.3 (2017) 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

[11] UNECE Big Data Quality Task Team (2014) A Suggested Framework for the Quality of Big Data. Final version. Link: https://statswiki.unece.org/display/GSBPM/Issue+%2322%3A+Metadata+Management+-+GSBPM+and+GAMSO [as of 5.01.2018]

2.7. Model errors and Precision

2.7.1. Introduction

Big Data based estimates are likely produced by models. The specifications of these models may be incorrect which negatively affects the reliability of the estimates.

As with traditional data sources models can be applied in different steps of the statistical production process. The discussions in previous chapters show the special need of using models in order to provide unbiased estimates based on Big Data.

Working with Big Data sources instead of survey data, NSIs do mostly not find the information about the target variable in the data source directly; instead the information of interest has to be inferred from other variables in the data. This is not a completely new situation; the deduction of information about the target variable from other variables is also a common process when working with administrative data. Also text mining algorithms were already known before new data sources were considered, e.g. when working with open questions in a survey. Still, modelling the information about the target variable plays a prominent role when working with Big Data, and often, new models have to be developed. The complexity of the algorithms and models needed to arrive at the needed information about the target variable depends on how directly the information of interest and the available information from the Big Data source are connected.

Model errors in the different steps of production can range in their potential influence. Incorrect models in the data editing or data validation phase might lead to incorrect observations on a micro level, but incorrect models in the estimation phase, e.g. to correct for the selection bias of a Big Data source, might lead to biased final estimates.

Statistical models have the special property of a foundation in probability theory, whereas machine learning algorithms are often of an adhoc/heuristic nature. With very large data sets, conclusions based on small-sample statistical inference can be ineffective at best and misleading at worst (Lin et al 2013). Based on probability theory we know that consistent estimators have standard errors that shrink as the sample size increases. With a very large dataset, the standard error becomes extremely small, so that even minuscule distances between the estimate and the null hypothesis become statistically significant. In other words, for very large data sets, the p-value becomes so small that the null hypothesis (e.g. “the coefficient of a predictor variable from a regression model equals zero”) is always rejected.

A model in general is a simplification or an idealized form of the data-generating process (the truth), so model misspecifications can occur for classical statistical models, e.g. linear regression, but also for advanced machine learning algorithms, like random forest or deep learning, e.g. simply by not including an important variable.

Estimating the quality of models is of great importance, therefore techniques like cross validation, out-of-sample tests, etc. should be applied wherever possible.

In (Breiman 2001) the author claims that the use of algorithmic modelling (e.g. random forest) instead of data models like linear regression, logistic regression, etc. might be better suited for large and complex data sets. However the problems of

- Noise accumulation,
- Spurious correlations and
- Incidental endogeneity

as cited by (Japec et al 2015) from (Fan et al 2014) can affect all these modelling techniques, just due to the massive size and dimensionality of the data.

Models might be successfully applied to Big Data sources, but still selectivity of Big Data source might hamper the estimation of model parameters (De Waal et al 2014).

In the Report on Methodology, which is deliverable 8.4 of WP 8, a whole chapter addresses the topic of machine learning in Official Statistics (2.6).

2.7.2. Examples and methods

WP1 Job Vacancy Web Scraping.

Models were used for de-duplication:

> The Slovenian pilot has a duplication issue with their weekly collection where jobs advertisements will typically span more than one reference period. However, these duplicates can be identified quite easily as any job ad that had the same URL as an existing record. To remove duplicate adverts that do not have a unique URL, the first step is to prepare and standardize the data fields that are common and that can be compared to all data sources (i.e. job title, location, company name, date posted, job description). This involves **text normalisation procedures** such as**, removal of white spaces, case standardisation and removal of stop words or other extraneous text**, typically using regular expressions (regex). The next step involves calculating a **similarity metric** to identify any likely duplicate job ads. One approach explored by the UK involved using Python Dedupe, which is designed to identify duplicate records using supervised learning methods. This uses **an initial match using logistic regression and then identifies marginal cases for clerical resolution**. The decisions of this clerical process are then reincorporated into the machine learning algorithm, to be applied for automated removal of duplicates. Other matching methods were explored by Sweden and included, Levenshtein distance and longest common substring distance, although Jaccard similarity performed best. The initial focus was on the structured data fields rather than the unstructured content of the full job description. This was mainly because this information is often difficult to scrape from websites in full and often only a "snippet" of a certain number of characters is readily available. This information may often be needed to achieve a good quality de-duplicated data set, especially with a large number of records

WP2 Scraping of Enterprise Characteristics

Both, generic scraping of whole websites and specific scraping of predefined content was used to collect information about the cases as listed in Section 2.3.2

Statistical models were applied in the Analysis Layer (also see Section 2.3 Process Errors). These models were needed, since it was not possible to go individually through every scraped website so as to check if the website contains information on job vacancies, if the enterprise is engaged in e-commerce or if there exists information about social media presence. Therefore, models were needed so as to predict the dependent variables from scraped information and correctly classify the enterprise. The members of WP2 describe that machine learning approaches as well as deterministic techniques were applied:

> We speak about machine learning approaches in scraping for official statistics when algorithms or models are derived from a set of training data which is supposed to be reasonably representative for the problem at hand. The parameters of the model are usually tuned with a validation set before measuring its performance on a so called test set with known characteristics. Finally the model is then applied to other sets of data of which we do not know anything but for which we suppose the model performs well, in order to produce statistics.
>
> We speak about deterministic approaches in scraping for official statistics when algorithms are designed from a set of rules with known characteristics of the sites and data in mind. Put in a different way, the knowledge of an expert is used to design an algorithm to process and interpret input data from web and other sources into statistical target variables. We call this method deterministic, because the algorithm applied to the same data will always have the same (deterministic) result, where the result of a machine learning approach heavily depend on the training set being used.

It is clear, that especially in the machine learning case, the quality of the model depends on the quality of the training data and on the test data so as to make validation possible. This is also described in the report:

> One thing to be noted here is that in machine learning approaches the availability of training data of sufficient quality is essential. This happens to be a challenge in many cases. In some of the pilots this training data is available or can be derived from earlier surveys. This might be true the moment when a machine learning approach is introduced in official statistics to (partly) replace a traditional statistical process, however on the long run, survey-based training data might become a rarity and other means have to be found to (re)train machine learning models. Obviously, deterministic approaches do not have this challenge, but have other pitfalls.

WP3) Smart Meter Data

Models are tested for classification and clustering. They can play an important role for data validation/editing and also for estimation.

One task is to try to distinguish between households and businesses based on the metering data. Different machine learning methods and statistical models were compared and up to 86 percent of the test data could be classified correctly. **The classification rate is computed by applying repeated k-fold cross validation.**

The tests were also extended to other variables such as urban/rural regional classification, household size or the NACE code for businesses.

Unsupervised learning methods (clustering) were also applied to the metering data to identify similar groups in the metering points and to later classify them as being vacant, primary residence or secondary residence.

WP6) Early Estimates

WP6 compared thousands of models with the help of different model errors. As a rule, the best model is decided to be the one with the lowest value of RMSE (root mean square error). The best model also has to be unbiased: ME (mean error) has to be close to 0. Some other criteria are also important: low MAE (mean absolute error), low MaxAE (maximum absolute error), low MAE (mean absolute error) in year-on-year growth rate; if microdata are levels (in contrast to growth rates), relative errors are also important. If the number of parameters divided by the number of time points is high, over-fitting is considered. For more information, see Deliverable 6.2 of ESSnet Big Data, SGA-1 (2016).

WP7a) Population Domain: everyday citizen satisfaction from social media data such as twitter (combined with others)

Text mining and machine learning tools can be applied so as to extract sentiments. Still it is clear that even a very well adapted model will have problems to handle problems to classify specific forms of expressions which can be easily recognised by humans, but very hard to detect by machines, see also Section “2.6 Measurement Errors”

> Secondly, the clarity of statements/comments may not be so obvious. It is of utmost importance to develop a glossary and to adopt precise definitions. Moreover, the fact that some statements can be ironic must not be neglected. The survey results obtained will also depend on the actual algorithm and learning patterns. Erroneous inclusions or omissions will therefore be unavoidable. However, according to human behaviour conception, negative comments tend to be more frequent than the positive ones. On the other hand, it becomes increasingly more common to post comments which do not reflect personal opinions but constitute a marketing attempt at eliciting specific reactions.

Figure 9 is taken from Deliverable 7.1 (2016), and emphasizes the importance of a training dataset.

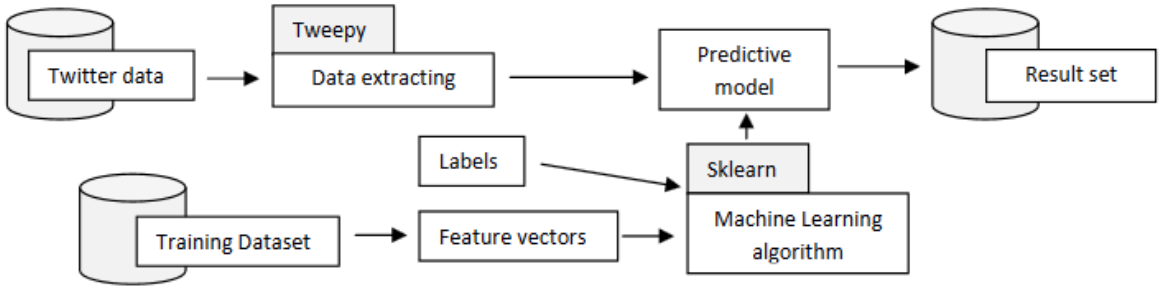

Fig. 5. *Big Data Framework used for Use Case 1.1.*

Figure 9 Illustration of the Big Data Framework for Use Case 1.1, taken from Deliverable 7.1 (2016)

2.7.3. Discussion

There are lot of possibilities to apply what is called statistical modelling. The question is if some of the error scenarios described already in previous chapters may distort or influence the use of models. Another issue is if we can modify the estimators based on Big Data in a way to counteract possible sources of bias. One of the very important methods coming out from sampling is the calibration as final weighting step. Since we can agree that Big Data is not the same as a sample but as well not covering the target population to the full extent it could be taking into consideration to invest in researching calibration methods for Big Data.

2.7.4. Literature

Lin, M., Lucas, H.C. and Shmueli, G. (2013), *Too Big to Fail: Large Samples and the p-Value Problem*, Information Systems Research, Articles in Advance, pp 1-12

Japec, L. et al (2015), *Big Data in Survey Research – AAPOR Task Force Report*, Public Opinion Quarterly 79 (4), 839-80

Fan, J., Han, F. and Liu,H. (2014), *Challenges of Big Data Analysis*, National Science Review 1: 293-314

De Waal et al (2014), *Statistical Data Editing of Big Data*, Paper for the Royal Statistical Society 2014 International Conference, Sheffield, UK

Breiman, L. (2001), *Statistical Modeling: The Two Cultures*, Statistical Science, 16: 199-231

ESSnet Big Data, SGA-1 (2016), Work Package 6, Deliverable 6.2, *Recommendations about the IT tools for collection of data for purposes of Consumer Confidence Index and NowCasts of Turnover Indices.,* https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/60/WP6_Deliverable_6.2.pdf

ESSnet Big Data, SGA-1 (2016), Work Package 7, Deliverable: 7.1 Report for Population domain, 7.2 Report for Tourism/Border crossing, 7.3 Report for Agriculture, "https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/15/WP7_Deliverable_7.1_7.2_7.3_2017_02_01.pdf"

3. Conclusions

This report gives an overview of seven quality aspects, which were identified as especially important in the WPs 1-7 of the ESSnet Big Data. The report describes each of the quality aspects and lists what role they played in the respective Big Data projects of the work packages. The report further tries to relate the listed quality aspects to already existing quality frameworks.

The experience gathered in the ESSnet Big Data, and especially in this WP 8, shows that on an abstract level, the same quality aspects can be considered for different Big Data sources, but the diverse nature of Big Data sources can make it very difficult to apply standardized quality measures for different Big Data projects.

It is also important to emphasize that for new Big Data projects and for new Big Data sources, other quality aspects than the ones listed in this report can be decisive.

Still, we hope that this report can be an initial guide about which quality aspects need to be considered when NSIs are planning new Big Data projects in the future.

4. Abbreviations and acronyms

AIS – Automatic Identification System
AUC – Area under the Curve
CoP – Code of Practice for European Statistics
BD – Big Data
ESS – European Statistical System
ESSnet – European Statistical System network
ESA – European System of National Accounts
ETL – Extract/Transform/Load
FPA – Framework Partnership Agreement
GDP – Gross Domestic Product
MP – Mobile Phone
MMSI – Maritime Mobile Service Identity
NACE – Nomenclature statistique des activités économiques dans la Communauté européenne
NSI – National Statistical Institute
NSO – National Statistical Organisation
OECD – Organisation for Economic Cooperation and Development
PCA – Principal Component Analysis
SGA – Special Grant Agreement
TB – TerraByte
TDM – Term Document Matrix
UNECE – United Nations Economic Commission for Europe
URL – Uniform Resource Locator
WP – Work Package

5. List of figures and tables