ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

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

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

Work package K

Methodology and quality

Deliverable K5: First draft of the methodological report

Final version, 17.06.2020

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

Work package leader:

Alexander Kowarik (STAT, AT)

alexander.kowarik@statistik.gv.at

telephone : +43 1 71128 7513

Contents

0. Prologue/Starting point

This report provides an overview of the current state of art of Big Data methodology. As we have found that the range of what one considers Big Data methodology differs tremendously on the experience and area of expertise of a particular statistical researcher, we decided to apply a very strict (narrow) focus on what is included in this report. Essential starting point of any method described in this report is that it:

i) has been used by researchers involved in the ESSnet Big Data I and II projects,
ii) is included in one of the Big Data based official statistics currently being published,
iii) is used in one of the Big Data based statistics that is in the process of becoming officially published.

Any other method, however interesting it may seem, is not included in this report. For the reader it is important to realize that the methods included in this report are divided in various (sub)sections that relate to the various phases of the statistical process in which they are used. The same phases, input, throughput (1 & 2) and output, were used in the quality guidelines report (WPK Del. K3) published earlier.

1. Overview of Big Data based official statistics

At the moment of writing, 13 examples of statistics are known that make use of Big Data for statistics published by National Statistical Institutes (NSI's). These are listed in table 1. Unfortunately, not all processes have already completely matured and only two are in production. These numbers will undoubtable increase in the near future. This report provides an overview of the current state of art of the methods that have been applied in the ESSnets Big Data I and II and the methods that are currently used for Big Data in official statistics, either in production or as published experimental statistics. Goal of this report is laying the foundation for statistical Big Data methodology. All methods discussed have actually been used in one or more of the statistics listed in table 1.

Table 1. Overview of Big Data based official statistics

1. Consumer Price Index (in production, multiple countries)
2. Traffic intensities (in production, NL)
3. Online job vacancies (towards implementation, ESSnet)
4. Enterprise characteristics (towards implementation, ESSnet)
5. Electricity/energy consumption (towards implementation, ESSnet)
6. Maritime and Inland waterway statistics (towards implementation, ESSnet)
7. Financial transaction based statistics (exploratory, ESSnet)
8. Earth observation derived statistics (towards implementation, ESSnet)
9. Mobile network derived statistics (towards implementation, ESSnet)
10. Innovative tourism statistics (exploratory, ESSnet)
11. Innovative company websites (towards implementation, NL)
12. Social mood on economy index (published experimental, IT)
13. Mobile phone derived outbound tourism (experimental, AU/FI/Estonia)

The first two statistics are in production. For the Consumer Price Index (CPI) either scanner data, web scraped prices or the combination of both sources are used. This statistics makes use of multiple sources, including surveys, that provide the price of products and (when available) the number of items sold. Usually, monthly statistics are produced (Eurostat 2017, Griffioen and ten Bosch 2016). Scanner data is being used by the statistical offices in Austria, Belgium, Denmark, France, Italy, Luxembourg, the Netherlands, New Zealand, Norway, the United States, Sweden and Switzerland. In a considerable number of other countries, their potential is being investigated or access to scanner data is being negotiated. Examples of such countries are Denmark, the United Kingdom and Portugal. Prices scraped from the web are part of the CPI for a considerable number of countries, for example in Italy, the Netherlands and the United Kingdom. Many other countries are considering using it, such as Brazil and Belgium (Del K1 2019).

The second Big Data based statistics in production is Traffic intensity statistics of the Netherlands. The national statistical office of the Netherlands was the first to produce an official statistic completely based on Big data (Statistics Netherlands 2015). For this statistic, vehicle counts data produced on a minute-by-minute bases by the 20,000 road sensors on the Dutch highway network are used (Puts *et al*. 2019).

The subsequent 8 statistical applications in table 1 are or have been investigated in the ESSnet Big Data. The majority of them focus on implementation in the near future by a number of NSI's. For these statistics the typification matrix (Del. K7) has been filled in. This enabled us to get an overview of the most important methodological issues stumbled upon in the ESSnet projects. Most import issue reported by the respondents was that, depending on the level of maturity of the process, it was difficult to answer all questions completely. As a result, it became clear that not all methods were already fully matured for these examples.

An additional assessment revealed that there are a number of Big Data based (experimental) statistics conducted at various NSI's that provide import methodological input. These are: Innovative company statistics based on text analysis of the webpage of both large and small companies in the Netherlands (Daas and van der Doef 2020), the social mood on economy index in Italy which is a daily index based on public Twitter messages (ISTAT 2020) and Mobile phone derived outbound tourism statistics in Austria (to be published with documentation in July 2020), Finland (Nurmi and Piela 2019) and Estonia (Ahas *et al*. 2011).

1.1 Using Big Data

It is important to indicate right from the start that when using Big Data the data in the source is given. The researcher working with the source has no influence on the way by which the data is obtained or from which units. Obviously, the only clear exception to this is when the data is obtained by scraping. But in all other cases "the data (in a Big Data source) is what the data is".

In essence, a Big Data source can provide input for statistics in two fundamentally different ways. The first is the direct use of the data in a source. Here, the data provided by a Big Data source is directly related to the phenomenon studied. For instance, the locations of a vessel or the number of vehicles counted by a road sensor are used to determine trips of vessels to a country (for maritime statistics) or calculate the traffic intensity of a road (for traffic intensity statistics), respectively. Because the data provided directly covers the needs of the NSI, one may expect that -as long as the data source remains available- its use for statistics does not pose huge problems. The large amounts of data available, however, can seriously affect the subsequent analysis part. Methods are needed to deal with these issues (see section 1.3).

The use of Big Data becomes more challenging when the data provided by a Big Data source is indirectly related to the phenomenon of interest to the researcher. This is identified as a derived use in this paper. For example one can use vessel trips and the counts of the number of vehicles to create an indicator for the economy of a country, related to the Gross Domestic Product for instance. Here, obviously, the data in the source is input for a model that eventually produces the economic indicator. Another example is the use of the text on websites to determine, for example, if a company is innovative or not. Here, again a model is used to derive that information. In all cases, this use is the result of a two-stage process in which the first step is the generation of the data and the second step uses a model to obtain the derived result. Clearly, anything that affects the relation between the data generated and the (derived) phenomenon studied influences the quality of the statistics based on such a source. The importance of this becomes apparent when we take a closer look at the use of Google search terms to predict the occurrence of the flu (Ginsberg *et al*. 2009, Lazer *et al*. 2014).

1.2 Lessons learned from Google Flu Trends

In 2009 a publication by Ginsberg *et al*. (2009) described that influenza could be detected by analysing the terms people typed in in the Google search engine. By selecting an appropriate set of search terms, the researchers were able to accurately estimate the level of influenza at a weekly basis around 1-2 weeks ahead of the official number for each region in the USA. However, the exact search terms used were not published. Clearly the search terms used for the analysis correlated with the actual flu cases registered. The correlation observed was around 0.9. The authors stated that the search queries used may not be all submitted by users who are experiencing influenza-like symptoms and that the correlations observed are only meaningful across large populations. Over-time this relation behaved fairly stable, until it became clear that in 2012, the Google flu prediction started to grossly overestimate the occurrence of the flu (Lazer *et al*. 2014). Obviously the relation between the search terms used in the model and the occurrence of the flu was no longer identical the one observed in the work used for the 2009 publication. There are a number of explanations for this which are discussed in the paper of Lazer *et al*. (2014). The most important conclusions are i) searching for flu is not the same as having the flu, ii) both the search behaviour of people and of the Google search engine may change over time and iii) one should be transparent on the exact approach and terms used.

For us, users of Big Data, it is important to realize that whenever a big data source is used to detect a derived phenomenon, there is a risk that the relation observed in the initial study may change over time. Hence, this relation should be carefully checked on a regular basis. If anything changes, the model used to derive the phenomenon of interest should be adjusted to keep the relation as constant as possible. Inspiration for this work can be obtained from the 'concept drift' research currently being performed in the analysis of streaming data (Lu *et al*. 2018).

1.3 Large *n*, large *p* and large *t*

In contrast to the traditional sources used for official statistics, big data contains extreme large numbers of data points. These may be the result of the inclusion of data for many units (large *n*), the inclusion of data for many variables (large *p*) or the combination of both (large *n* and large *p*). Statistical analysis of big data may thus suffer from the increase in the dimensionality of the data (Fan *et al*. 2014). For the majority of the work performed in the ESSnet Big Data, it is clear that dealing with large *n* was dealt with the most. Data sources with large amounts of variables certainly became an issue during texts and images studies (see Chap. 2). Here, selecting the most relevant features becomes challenging. Also, in some cases the number of data points per variable over time was also very high; for instance in the studies of AIS, MNO and smart meter data. We will identify these cases as examples of dealing with a large *t*.

Because of these properties using Big Data comes with a number of methodological challenges. According to the paper of Fan *et al*. (2014), three unique challenges are raised. These are related to the large *p*, large *n* and large *t* topics introduced above.

The first challenge is mostly large *n* related, i.e. analysis of data for many units creates issues such as heavy computational cost and algorithmic instability. Many of the traditional statistical methods used (for small data) do not scale well to massive amounts of data. Analysing all data may take considerable time which will, very likely, result in heavy computational costs. The other issue is the way the outcome of an algorithm is affected by small changes to its inputs. This is known as algorithmic instability. A stable learning algorithm is one for which the prediction does not change much when the training data is modified slightly. Thus, the algorithms applied needs to be fast, should be scalable to large amounts of data (see WP8 Del 8.3 section 3.9 for more details on this) and its outcomes should be stable. This forges cross-fertilizations among different fields including computational science, informatics, statistics, optimization and applied mathematics. When for each unit many parameters are derived, for instance during the analysis of texts, a large *n* issue also becomes a large *p* challenge.

The second challenge is predominantly large *p* related. Because of the large number of variables, e.g. the high dimensional properties, included in some big data sources noise will accumulate, spurious correlations will occur and incidental homogeneity may arise. Methods are needed to better deal with these issues. Data on more variables leads to more noise and more noise leads to a higher change of accidental correlations between variables. In small data sets, dimension reduction techniques and variable selection methods are used to deal with these issues. However, in big data these approaches do not work satisfactory (Fan *et al*. 2014; Bolón-Canedo *et al*. 2015). For example, classification of big data with conventional classification rules using all features may perform no better than random guess. Another example is text classification related. Increasing the number of training examples during model development will often increase the number of features included in the final model without seriously improving the overall accuracy of the model. Furthermore, the spurious correlations introduced by noise accumulation may lead to wrong statistical inference and false scientific conclusions.

The last challenge is mostly large *t* related, i.e. the massive amounts in big data are typically collected at different points in time and may even be obtained from multiple sources collected by different technologies. The latter creates issues of heterogeneity, experimental variations and statistical biases, and requires the need to develop more adaptive and robust procedures. The former leads to the same issues as mentioned under the second challenge. All will affect the inference part of the statistical process and reduce the quality of the outcomes.

1.4 Relation with Quality and Typology reports

This report focusses on methodology, i.e. the methods used to produce official statistics while using Big Data. Two other deliverables of workpackage K (WPK) provide overviews of the work on Quality (WPK Del K.4) and Typology (WPK Del K.7). The first is obviously related to the methods used as these methods aim to reduce and compensate the negative effect of quality issues in the data (and metadata) as much as possible. The chapter on “Dealing with errors” in Big Data in this report (Chap. 4) specifically discusses this relation. The link with the Typology work is perhaps not that obvious. In the Typology report a Typification Matrix is described that makes it possible to assess the maturity level of a big data source to assist the future evaluation of big data sources/projects. The information is provided by the project leader of the research/implementation process. Having gathered all this information, it should be possible not only to achieve an assessment on the Big Data project as well as trace the roadmap for the business case of the exploration of the data source. In addition, the business functions of an architecture for big data (BREAL; WPF Del F.1) are mapped and connected with the Quality and Methodology reports through the columns Challenges and Treatments. For the reader's convenience, Annex A provides an overview of the various Business functions, Life cycle and actors discerned in this architecture. To clearly indicate this relation, the Methodology chapters in this report are linked to the corresponding business functions in the BREAL architecture.

Input Phase

2. Introductory overview of methods for text and image mining

The increased focus on Big Data has greatly stimulated the uses of other types of data sources in official statistic production. In particular, texts and images hare increasingly being studies. Because of this, the methods applied to extract information from these sources have become part of the methodology being used for official statistics production. For this reason this documents starts with an introductory overview of the methodologies commonly applied for text and image mining. Since Natural Language Processing (NLP) is an approach often mentioned for text analysis, this is also included in the text mining section.

2.1 Text mining and NLP

Text mining can be defined as the process of extracting information from textual data (Hotho *et al*. 2005, Jo 2019). Examples of typical text mining tasks are categorization, clustering, entity extraction, sentiment analysis and document summarization. All tasks, with the exception of the latter, have been applied in the ESSnet Big Data. In general, text mining is a special type of data mining. Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages. NLP studies how to program computers to process and analyse large amounts of natural language 'data'. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation. Clearly, text mining can be considered a subfield of NLP.

2.1.1 Text

Usually text is defined as a group of sentences or paragraphs written in a natural language. Sometimes, even what is written in an artificial language, such as source code, may be included in that view, but this is ignored here. Individual words are the basic units of a text. These are combined into a sentence (grammatically) and sentences are organized into a paragraph (logically). A complete text usually consists of paragraphs of variable lengths and to it information such as a date, author, title, and so on may be attached. However, this may not be the case for all texts. For example, social media messages usually lack a paragraph structure. Words, sentences, paragraphs will be briefly discussed below. The whole text is often described as a document in many text mining tasks.

Words: A word is considered as the (meaningful) basic text unit. A single word may consist of several characters but the characters have no meaning on their own; a word has. A group of grammatical words known as stop words are only used for grammatical functions, such as "a" or "the", and do not contribute (much) to the meaning of a text. For this reason, they are usually excluded in many text mining tasks. Words are combined into a sentence by rules known as grammars. In many languages, the words are separated by white spaces; the process that splits a sentence into words is called tokenization. However, this approach cannot be used for all languages, exceptions are, for example, Chinese and Korean. In these languages white spaces cannot be used to split characters without affecting the meaning of a sentence.

Sentence: Each sentence starts with a capital letter and ends with a punctuation mark, such as a period, question mark, or exclamation mark. A sentence may have a single clause which basically consists of subjective and its verbs, or several clauses, depending on its types. A sentence may be decomposed into clauses.

Paragraph: A paragraph is defined as an ordered set of sentences, keeping the consistence for a particular subtopic. A paragraph is decomposed into sentences by punctuation marks. A paragraph usually starts with an indented position and ends with its carriage return; at least in many European texts. Each paragraph has a variable length. Note that sentences are combined into a paragraph by the logical order or the semantic rules, rather than the grammar.

Document: The overall text, such as a news article, a patent, a letter, the content of a webpage or an abstract, is usually referred to as a document. A document can consists of a single sentence, a single paragraph, a group of paragraphs, or several groups. In general, a document is considered the smallest unit for a typical text mining task. Many text mining scholars discern between the processing of large, medium and small sized documents. For larger documents fairly standardized (pre)processing steps and sequences have emerged to fulfil particular text mining task (described below). For small documents such standardization is often not possible. Here, it is required to determine the information loss caused by each individual step in the process.

2.1.2 Text mining tasks

As mentioned above, typical examples of text mining tasks are text categorization, clustering, entity extraction, sentiment analysis and document summarization. Since only the first four tasks have been performed in the ESSnet Big Data, the latter is not discussed here. For the categorization and clustering tasks preprocessing of the text is usually performed.

2.1.2.1 Preprocessing steps

Certainly for large documents, texts are usually processed in a standardized way. First, the relevant part of the text is extracted from the document. This is especially important for documents that contain a lot of lay-out related codes, such as webpages and pdf-files. Next, the language of the remaining (raw) text is determined. There are a number of libraries available for that, such as cldr and textcat in R and langdetect in Python. Depending on the methods implemented, results may vary. Subsequently, the text is converted to lower case and all punctuation marks, numbers, and single character words are removed. Next, depending on the language detected, the stop words are removed. In R the tm package contains this functionality, while Python has NLTK for that. Next, the different morphological variants of the remaining words are mapped to their base form. This can be done via stemming or by lemmatization. Both stemming and lemmatization generate the root form of the words processed but a stemmed word might not be an actual word whereas, the lemma is certainly an actual word. To enable the latter, lemmatization uses a large text corpus to find the correct lemmas which makes this step considerably slower compared to stemming. Stemming can be done in R with the tm package and for Python the NLTK or SnowbalStemmer library can be used. One needs to check if a stemmer for the specific language of the text is available. For lemmatization obviously language specific libraries are needed. Here, the choice is much more limited. In R, the koRpus package provides this functionality, while in Python the WordNetLemmatizer in NLTK can be used. The words remaining after all preprocessing steps are applied form the starting point for a number of text mining tasks.

Although a similar approach can be applied to the text in small documents, it is essential to keep an eye on the loss of information during each step. This is best illustrated by an example. When studying Twitter messages to detect those in which persons indicated that they would like to move to another house, it was found that the removal of stop words and stemming had a deteriorating effect on the information content of these messages. The accuracy of prediction dropped from 87 to 60% when these steps were included. It was found that the effect of stop word removal was mainly caused by the removal of the words 'I' and 'me'; these are included in the stop word lists of many languages. Since these words indicate that the content of the message referred to the person writing the message, it's clear that they contain import and relevant information. As a consequence, a tailor-made (short) stop word list was prepared for this task and stemming was not performed. All in all, this illustrates the need for a tailor-made approach when studying smaller documents.

2.1.2.2 Text categorization and clustering

When documents are classified or clustered they are usually preprocessed prior to analysis. After that, an approach commonly referred to as a Bag of Words approach is often applied. Here, how often a word occurs in a document is counted and the frequency of each word compared to the other words in the document is determined. Because the order of the words is irrelevant in this approach, it is referred to as a Bag of Words approach. Usually a slightly more advanced frequency measure, e.g. frequency-inverse document frequency (tf-idf), is used. Tf-idf is generally considered the best way to identify words that characterize the topics in a text as it reduces the effect of words that occur in many of the documents studied (Gentzkow *et al*. 2019). Next, usually a document-term matrix is created in which the rows corresponded to the individual documents and the columns to the words selected from the text collection. To reduce the size of the matrix, often only words that occurred in at least 100 documents are included. For each word in the preprocessed text, the tf-idf value is included in each cell. Many of the values will be zero. The scikit-learn library (v 0.21.2) in Python can be used for this. To this matrix other variables can be added as binary features. For text classification, the document-term matrix is usually the starting point for a machine learning algorithm that tries to found the optimum between the tf-idf values of the features and the correct classification of the examples included in the training set. Apart from single words, combination of two subsequent words, so-called bigrams, can also be included as features in the document-term matrix. For clustering a similar approach can be applied, except that no predefined class information has to be added and that somewhat different machine learning algorithm may be applied.

As a very interesting extension to this, word embeddings can be used as features for classification or clustering. Word embeddings are based on word co-occurrences and it is found that they often enable an improved extraction of topic information by encoding the semantic and syntactic information of words (Allen and Hospedales 2019, Li and Yang 2018). This may improve classification considerably, possibly in combination with the Bag of Words approach described above. In Python, Word embeddings are implemented in the gensim library (v 3.4.0).

2.1.2.3 Entity extraction and more

When entity extraction is performed the texts in documents are usually studied as is; i.e. no preprocessing is performed. Here, the semantic structure of a sentence is used to detect possible entities, such as persons, companies and places. This is an NLP task that is (obviously) language specific and hence requires language specific entity recognition software. For English and a number of other languages spoken by many people on earth such software is usually available on the web. A good starting point in Python is the spaCy library that supports a considerable number of languages. In R OpenNLP is a starting point.

For some simple extraction tasks, regular expressions (RegEx) can be used very efficiently. By defining specific sequences of characters in a regex a number of search patterns are defined. RegEx are part of nearly every programming language. An example of a RegEx that is able to identify two subsequent words separated by a space that each start with a single upper case letter is:

```
“[A-Z]{1}[a-z]+\s[A-Z]{1}[a-z]+”
```

When applied to a text, this will enable the identification of persons, companies and probably also some street names. Similar approach can be applied to, for instance, detect phone numbers, bank account numbers, URLs and more.

2.1.2.4 Sentiment analysis

Determining the sentiment of a given text is another often applied NLP task. This is sometimes also referred to as opinion mining. Here, the ‘polarity’ of a text is being determined which is usually indicated as positive, negative or neutral. More advanced sentiments, such as angry, sad and happy, can also be determined. After some preprocessing steps, a researcher usually studies the occurrences of individual words in the text for which the polarity is known, after which the overall sentiment of the text (or sentence) is determined. For this a list is needed that contains words classified as either positive or negative or classified on a range scale. By combining the overall score of the words, according to some predefined rules, the net sentiment for the whole text is derived. In general, this provides useful results but it is challenging for cryptic and sarcastic sentences. In R the tidytext packages can be used and in Python the NLTK library can be used for this task.

2.1.3 Examples of text mining applications

Texts have been studied in a number of workpackages of the ESSnet Big Data.

In WP1/B the online vacancy texts were studied to link them to company names in the Business registers (Del 1.3, p. 81), compare the vacancies texts in multiple sources to find duplicates (WP1, Del 2.2, p. 11 & section 4.3.2) and extract education, qualification and skills (Del. B1, p.7).

WP 2/C focussed on enterprise web sites with the following subtasks. Create an inventory of the URL of enterprises web sites, detect E-commerce, locate job vacancy ads, detect social media presence and extract relevant information from the web site. Each of these tasks requires the analysis of texts for instance to compare company names and URLs, to detect words specific for E-commerce, to detect vacancies and to compare social media usernames and profiles with company names. An overview of this work can be found in Del 2.2 of WP2 and in Del C2, p. 37.

In WP7 social media were studied and user profiles and messages were analysed, amongst others to determine their sentiment (Del 7.7, chap. 3). For Italy such an indicator related to the Economy, based on publicly available Twitter messages, has been published (ISTAT, 2020).

In WPG the free text fields in bank transaction data were studied. It was found that these descriptions are not always very informative (see Del G1, p.72).

Another typical example of a text-based Big Data based statistics is the detection of innovative companies derived from the text on their web site. Because of this approach all companies with a web site, so also (non-surveyed) companies with only a limited number of employees -such as startups-, could be studied and classified (Daas and van der Doef, 2020).

2.2 Image mining

Image mining, also referred to as Computer vison, aims to automatically extract patterns from the raw data of images. Raw images or image sequences with a low level pixel representation are processed to enable the extraction of the high level objects and their relationships. In general, two approaches are usually applied: i) the Bag of Visual Words approach (Szeliski, 2010) and ii) Deep learning (Rosebrock, 2019). Both will be discussed below.

2.2.1 Bag of visual words

The general idea of the ‘bag of visual words’ (BOVW) is to represent an image as a set of features. This is very similar to the bag of words approach used for texts (described in section 2.1.2.2). For images, however, the features do not consist of words but of keypoints and descriptors. Keypoints are the “stand out” points in an image and usually information on their positon and coverage area are provided. These are general characteristics and do not enable one to determine how different or similar one keypoint is to another. For example, suppose you have an image of a duck and another image of exactly the same duck but it is twice the size. In both images, the extracted keypoints will be the same (the same parts of the duck) but their location is different because of the difference in size of the images. Hence, one is unable to compare these images based on the information provided by the keypoints alone. Here, descriptors come into play. A descriptor is the description of a keypoint and it summarizes, in vector format (of constant length) some characteristics of the keypoint. For example, it could be the intensity in the direction of the keypoint most pronounced orientation. As such, it adds a numerical description to the area of the image the keypoint refers to. What is special about descriptors is that they are independent of the keypoints position, are robust against image transformations and should scale independently. Because of these properties, the information provided by descriptors enable one to conclude that the duck image example mentioned above are actually identical, though different in size. Detecting features and extracting descriptors in an image can be done by using feature extractor algorithms; examples of the latter are SIFT and KAZE.

By creating a visual dictionary of the keypoints and descriptors of images, one is able to perform image mining tasks such as, image comparison, image classification, image clustering and detecting objects in images. Recent developments in machine learning, i.e. the rise of Deep Learning, have greatly decreased the interest in the BOVW based-approaches.

2.2.2 Deep Learning

Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. Similarly to how humans learn from experience, the deep learning algorithm performs a task repeatedly, each time tweaking it a little with the aim to improve the outcome (LeCun *et al*. 2015). It is called ‘deep learning’ because the convolutional neural networks used for this task have various (deep) layers that enable learning. To learn, large amounts of data and a considerable amount of computing power are needed.

The recent popularity of deep learning started in 2012 when the seminal AlexNet (Krizhevsky *et al*., 2012) outperformed all other classifiers on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by a large margin; 11% to the second-best entry. In the ILSVRC, computer vision algorithms are presented with a classification challenge in which they need to distinguish 1,000 different categories. In subsequent years, new developed convolutional neural networks were used to outperform the benchmark set by AlexNet and achieve almost superhuman performance (Geirhos *et al*., 2017). AlexNet and its successors showed that deep convolutional networks trained with stochastic gradient descent (or variations of it), make excellent end-to-end classifiers. These end-to-end classifiers can be trained with only the raw image data as input and the corresponding classification label as output. The features needed to perform the classification task are learned from the data, bypassing the need for specifying hand-crafted features. The trade-off of using convolutional neural networks is that they need a much larger amount of data as compared to the amount of data when using traditional computer vision with hand-crafted features. However, the performance of convolutional neural networks often drastically outperforms the traditional computer vision approaches.

One of the first successful applications of convolutional neural networks was a document recognition system developed by LeCun *et al*. (1998). In contrast to fully connected networks, convolutional neural networks contain specialized spatial layers: (1) convolutional layers that apply filter kernels to the image by sliding them over the image and calculating results for each position in the image, and (2) pooling layers, which summarize the results coming from the convolutional layers by retrieving the maximum or average value in a certain receptive field. By looking at local patches of the image and commonalities in-between them, convolutional neural networks can learn filter weights that look at edges, textures, colours and can group them in hierarchies that look for specific details in an image. Because filter weights are shared across the whole image, much less weights are needed than in a fully-connected scenario where each input node of the network would correspond to one pixel in the input image. Initially, convolutional neural networks, like AlexNet, consisted of subsequent convolutional layers organized in a linear fashion. However, already soon after, convolutional networks like Inception/GoogleNet, ResNet and Xception, were developed that employed specialized modules and non-linear paths. In recent years, convolutional neural networks have brought about breakthroughs in the processing images, video, text, speech and audio. Because of these developments Deep learning is currently successfully applied for automatic translation, speech and facial recognition, object detection, and in the creation of fake news videos and in many other domains such as drug discovery and genomics. In python, deep learning studies are performed using TensorFlow and PyTorch.

2.2.3 Examples of image mining applications

Images have been predominantly studied in WP7 and WPH of the ESSnet Big Data. In both WP's satellite images and areal pictures were predominantly studied to detect crops, seasonal changes, soil properties, map tillage activities, detect urban sprawl, quality of housing, to determine land cover, and to detect air pollution. In some cases, the findings have been enriched by combining them with administrative data. Results are described in Del. WP7.3, WP 7.7 (Chap. 5) and WPH1.

In some WP's the potential use of images is mentioned. Examples of this are using image data to detect traffic flows in Italy (WPJ, Del J2., p.9) and the use of images, such as logo's, on company web pages for Enterprise statistics (WP2, Del 2.2, p. 61).

2.3 Discussion

The rise of Big Data has greatly stimulated the study of alternative data sources by NSI's. This has lead to an increase interest in methods capable of extracting information from texts and images. At the moment of writing, web pages and social media messages are the predominant text-based data sources studied. The work on images predominantly focusses on sattelite and arial pictures. This work has, amongst others, led to an increasing in the application of machine learning and deep learning methods. Since these methods provide interesting results but they way these findings are obtained are not (always) very transparant, it is expected that future work by NSI's will increasingly focus on the application of Artificial Intelligence and include studies on algoritmic fairness and transparancy (Helwegen and Braaksma, 2020).

Throughput Phase 1

3. Big Data exploration

To enable the application of Big Data, large amounts of data need to be efficiently checked. This is the case for both exploratory studies and for production purposes. The technique known as Exploratory Data Analysis (EDA) focusses on this task and touches upon pattern recognition. In EDA one of the following tasks is included:

- **D** Describe data distribution: how often does a value or a combination of values occur? What is the dependence on other variables?
- **G** Find groups: is the distribution a mixture of distributions of different ‘natural’ groups in the data? If so, the data can be split.
- **O** Identify outliers: find improbable extreme values. Are they errors? Determine if they should be excluded or corrected in the analysis.
- **M** How many values are invalid or missing? Determine what the impact is on the analysis.

Visualization has proved to be a very useful tool to explore and summarize data. The most famous example to illustrate this is Anscombe’s quartet. This quartet comprises of four data sets that have (nearly) identical simple descriptive statistics, yet reveal very different distributions when plotted (Anscombe 1973, Figure 1).

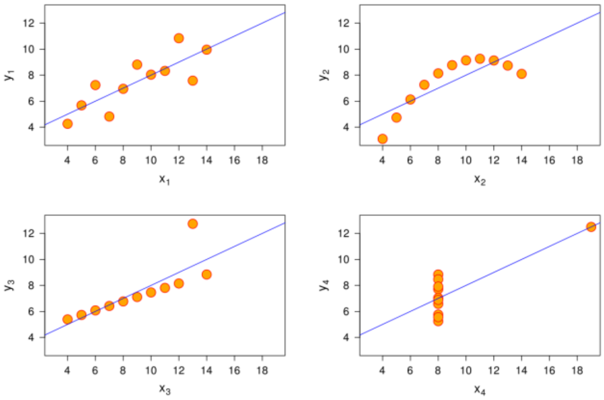


Figure 1. *Plots of Anscombe’s quartet revealing the clearly different composition of the four datasets.*

However, the traditional exploration methods have been developed on relative small data sets and not on very large ones. So while it is very easy to load a relative small dataset into a computer’s memory and completely explore it, this becomes challenging when the dataset becomes very large (Tennekes and de Jonge, 2019). Hence the need for scalable methods for the exploration of large data sets. Here, a large dataset refers to data with a large numbers of records (*n*) and/or large number of variables (*p*). An easy solution in the beginning is to start the exploration process with a random sample, but to get a good overview increasing amounts of data needs to be visualized.

3.1 Visualization methods for large datasets

Some visualization methods have found to be particularly well suited for the exploration of large datasets. The tasks for which they are suited are indicated between brackets.

3.1.1 Univariate data

For univariate visualizations the most informative ones are Empirical Cumulative Density Functions (O), Histogram/frequency plots (DG), Kernel Density Estimators (G), Frequency plots, Pie charts or stacked bar charts (DM), Treemaps (DM) and Calendar plots (DO). The majority of the visualization methods make use of the diamonds dataset (included in the ggplot2 R-package). To this dataset, sales dates and manufacturing dates are added to illustrate various datetime plots. For the Treemap artificially generated Business data is used. The plots are shown in Figure 2 and the findings are described below.

Histogram/frequency polygon plots (Figure 2A) the distribution of a numerical, ordinal or datetime variable using bars. Numerical values are discretized cutting the variable range into intervals (bins) and counting the frequency. It is advised to try several different bin sizes to extract features from the data set. A histogram is well suited for describing a data distribution since it makes no assumptions. Figure 2A shows a histogram of the carats variable. A datetime variable is typically shown in a frequency polygon, which is the line version of the histogram.

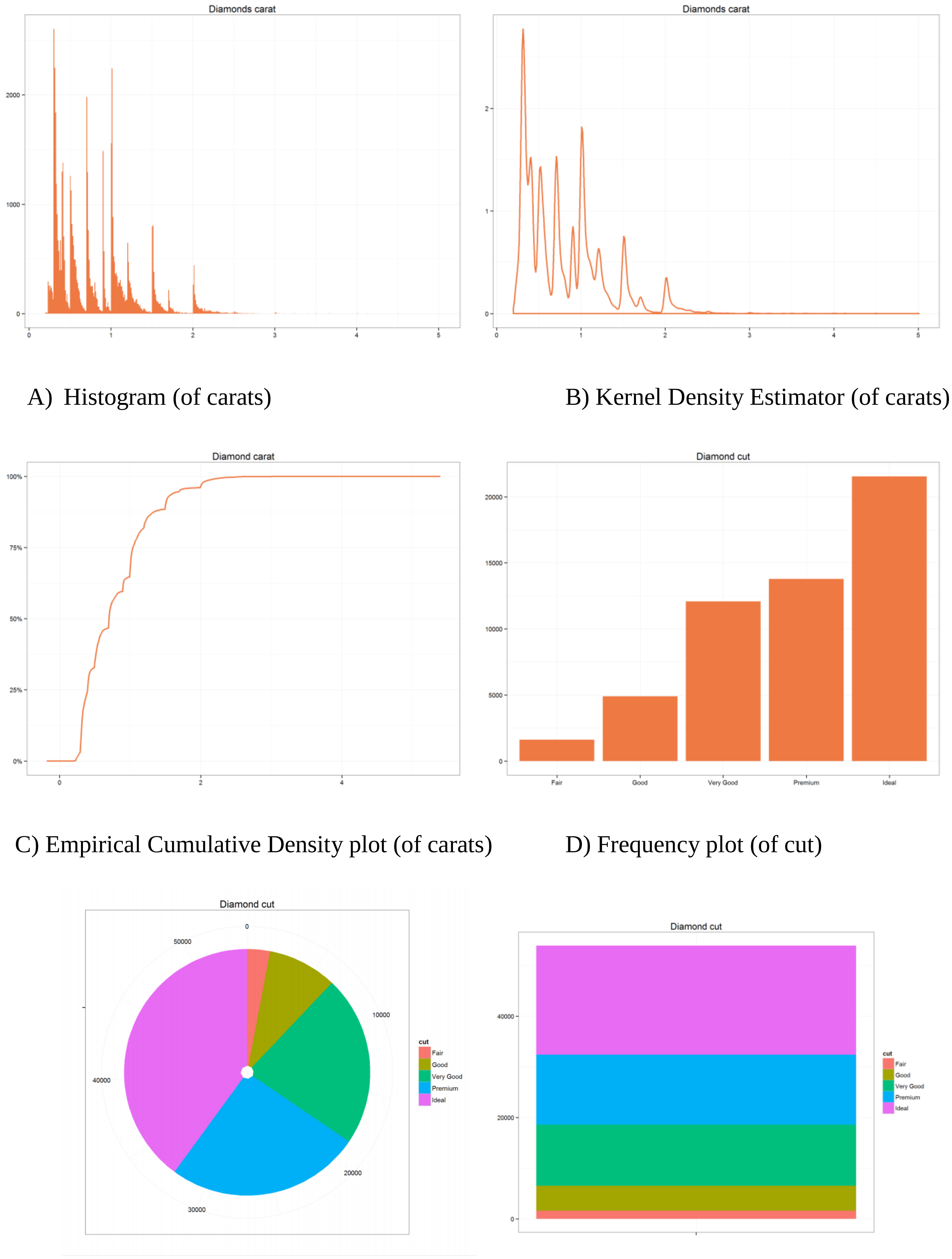
Kernel Density Estimator (KDE; Figure 2B) is a non-parametric technique to estimate the probability density of a numerical continuous variable. Similar to the bin width of the histogram, a KDE has a band width. The method assumes that the variable is continuous, which makes the technique less suited for exploratory purpose: it may distort data by smoothing out relevant (asymmetric) discontinuities and outliers. Figure 2B shows the KDE of the carats variable (compare with Fig. 2A).

Empirical Cumulative Density Function (ECDF) plots (Figure 2C) each value of a numerical variable against its order (on a percentage scale). The ECDF is close to the original EDA techniques that use quantiles to describe data distribution. It is well suited for finding outliers and describing where the 'mass' of the data is. It can therefore be used to truncate the range of a variable, or to determine an exclusion threshold for outliers based on quantiles. It is less suited for viewing the data distribution, although modes can be seen. Figure 2C shows the ECDF plot of the carats variable.

Frequency plot, pie chart or stacked bar chart (Figure 2D, E, F) show the frequency of low cardinality categorical data. Although infamous, the pie chart can be useful to show part-whole relations, but is not well suited for comparing frequencies of categories. Figure 2D shows a frequency plot, figure 2E a pie chart and figure 2F a stacked bar chart of the cut variable.

Treemap (Figure 2G) shows the frequency of a hierarchical categorical variable. This is an effective technique to summarize the frequency of a variable with high cardinality. The treemap may be coloured using a hierarchical colour scheme. Treemaps are also very useful for bivariate distributions (see next section). In figure 2G a treemap of business data is shown.

Figure 2. *Overview of univariate visualizations*



A) Histogram (of carats)

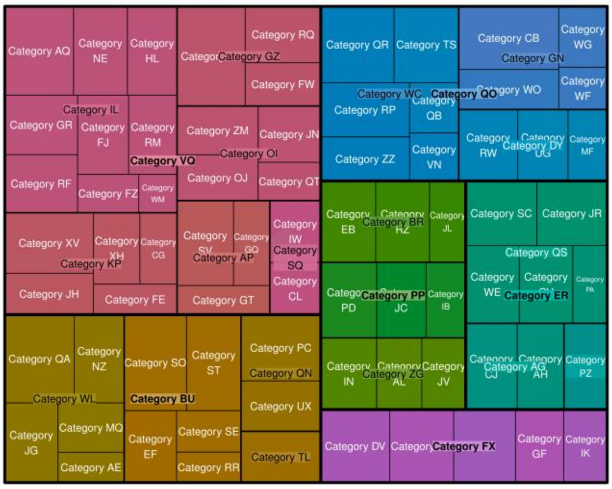
B) Kernel Density Estimator (of carats)

C) Empirical Cumulative Density plot (of carats)

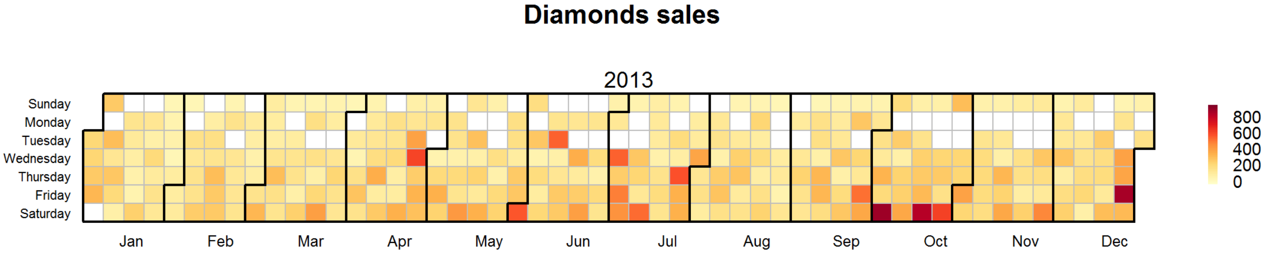
D) Frequency plot (of cut)

E) Pie chart (of cut)

F) Stacked bar plot (of cut)



G) Treemap (of Business data)



H) Calendar plot (of sales)

Calendar plot (Figure 2H) shows the frequency per day of a datetime variable. For variables with a range between one month and a couple of years, a calendar plot is very useful. It can be seen as a one dimensional heatmap in calendar format. High counts are considered outliers. Figure 2H shows the (artificial) sales data added to the diamonds dataset to illustrate the amount of sales during 2013.

3.1.2 Bivariate data

A scatterplot is an example of a visualization method for bivariate distributions that scales poorly and is hence not recommended for large datasets. Good visualization methods for bivariate distributions are Heatmaps (DG), Two-dimensional kernel density plots (DG), Surface plots (DG), Mosaic plots (DM), Tableplots (DOM), Treemaps (D) and Small multiples (G). An overview of these plots is provided in Figure 3 and the findings are described below.

Heatmap (Figure 3A, B) is a very powerful workhorse that copes with the shortcomings of scatter plots. Moreover, it can be applied to numerical and high cardinality ordinal, categorical, and datetime variables. Numerical variables are discretized in the same way as for histograms. Counting the number of occurrences for each combination of (discretized) values results in a frequency matrix. This frequency matrix can be displayed as a heatmap. It makes no assumptions on the data. Alternatively, numeric data can also be binned and visualized in hexagons, which prevents rectangular binning artefacts. Figure 3A and 3B show the heatmaps of carats versus price and sales versus manufacturing dates of the diamond set.

Two-dimensional kernel density plot (Figure 3C) is another visualization method that is scalable for large number of observations in which contour lines are drawn based on the estimated kernel densities. It is especially suitable for numerical, but also for datetime variables. It assumes continuity in both dimensions. Figure 3C shows a 2D kernel density plot of carat versus price.

Surface plot (Figure 3D) is a three-dimensional plot, in which the densities are expressed by height. It is a useful tool to visualize bivariate relationships. Like the two-dimensional kernel density plot, it is useful for numerical, and can also be used for high cardinality ordinal and datetime variables. In figure 3D a surface plot of carats versus price is shown.

Mosaic plot (Figure 3E). Whereas heatmaps are especially useful for high cardinality categorical and ordinal variables, the mosaic plot may be more suitable for low cardinality variables, since the frequency distributions per category are encoded by the width or height of the rectangles, which may be easier to compare than colours. Figure 3E shows a mosaic plot of cut versus clarity. A stacked bar chart (Figure 3F) is similar to a mosaic plot, but does not show the univariate frequency distribution of the column variable. In figure 3F clarity versus cut is shown.

Tableplot (Figure 3G) is a plot that actually is a combination of bivariate plots. Data from two or more variables, that can be numerical, ordinal, categorical, and datetime, are binned according to the quantiles of a numerical variable. For each variable that is either numerical or ordinal with high cardinality, a bar chart with mean values per bin is plotted. For each low cardinality categorical variable or ordinal variable, a stacked bar with the frequencies of the categories is plotted. Figure 3G shows the tableplot of price, carats, cut, color, clarity, depth and table. Price is used as the sorting variable.

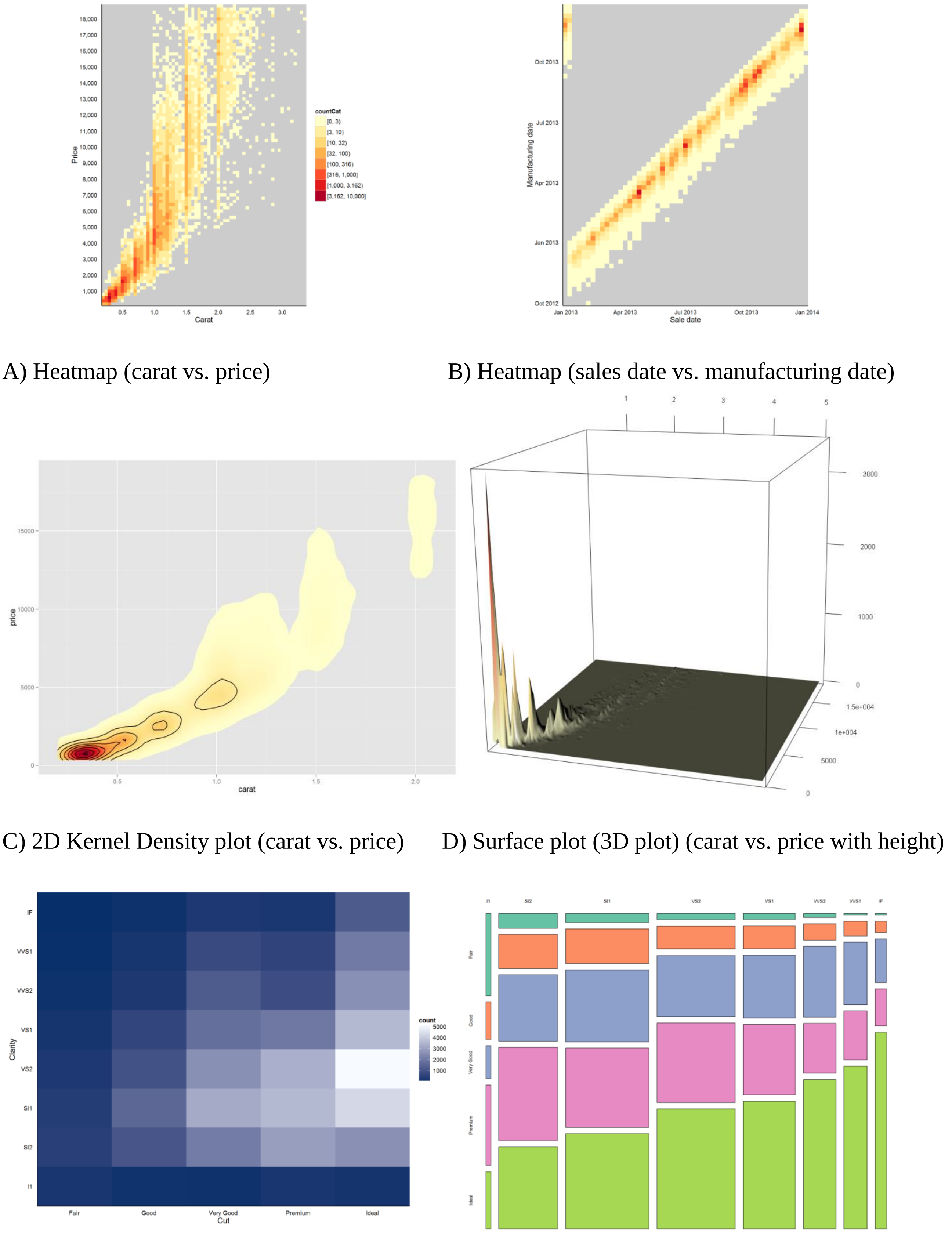
Small multiples (Figure 3H). A very powerful technique to explore bi- and trivariate relationships, is small multiples. Data are split according to the distinct values of one or two ordinal or categorical variables. Per (crossed) category, a small univariate or bivariate plot is shown. Figure 3H shows price versus count for each cut variable in the diamond set.

Treemaps (Figure 3I) are also very useful to visualize the relationship between a hierarchical categorical and a numerical variable. The sizes of the rectangles correspond to aggregations of the numerical variable based on the hierarchy of the categorical variable. In addition, colour can be used to encode a third, numerical, variable, typically with a diverging colour scheme. Figure 3I shows a treemap for business data growth for various business categories.

2.1.3 Other relevant visualization methods

For Big Data containing geospatial data, maps of an area or country are a logical choice. They can be combined with the visualization methods shown in Figure 2 and 3. Examples of such maps can be found in a number of ESSnet Big Data deliverables (e.g. Del 3.1, Fig. 8-11, Del. 4.2, Fig. 2-3, Del 4.3, Fig. 11-15 and Del H1, Fig. 4.6 and 4.8). A bubbleplot is also a very interesting method to display geolocated information (Daas and van der Doef, 2020, Fig. 2). For network data a number of visualizations are available, such as an alluvial diagram or a node-link diagram. In Deliverable 2 of WPJ on pages 16-18 a number of simple network visualizations are shown.

Figure 3. *Overview of bivariate visualizations*



A) Heatmap (carat vs. price)

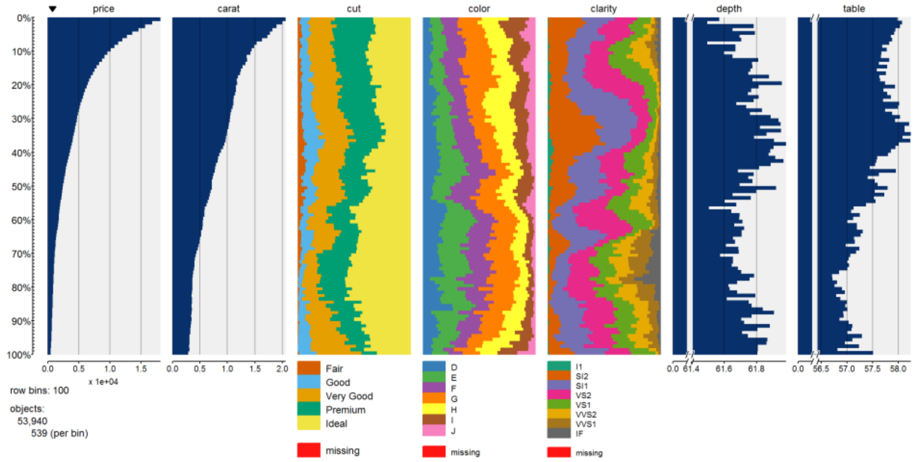
B) Heatmap (sales date vs. manufacturing date)

C) 2D Kernel Density plot (carat vs. price)

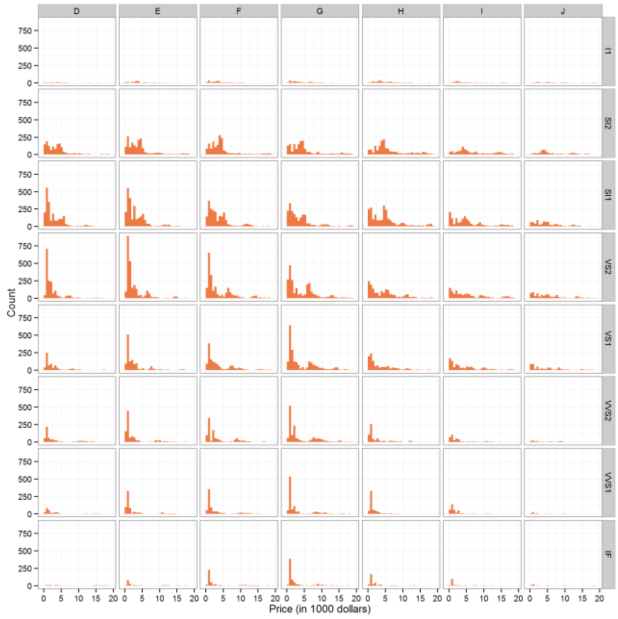
D) Surface plot (3D plot) (carat vs. price with height)

E) Mosaic plot (cut vs. clarity)

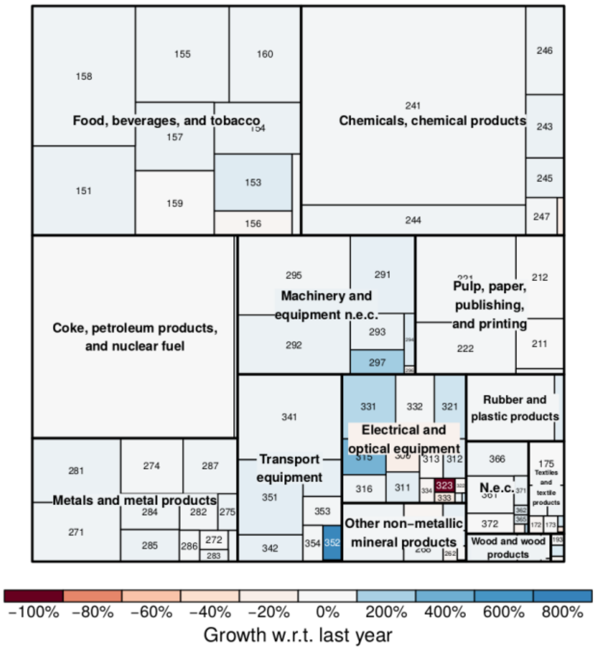
F) Stacked Bar chart (clarity vs. cut)



G) Tableplot (of price, carat, cut, color, clarity, depth and table)



H) Small multiples (price vs. count per cut)



I) Treemap (of business data growth)

2.2 Examples of visualization methods used

In the ESSnet a number of Big Data visualization methods have been applied. Since a lot of this work is of an exploratory nature, not all of them will be included in the deliverables of each WP. For example, creating small multiples of the number of vehicles detected per minute per road sensor is very insightful to get an idea of the quality and distribution of the data, but no such plot has been included in any of the deliverables of WP6 of the ESSnet Big Data 1. However, a number of insightful plots have been included in the deliverables specifically in WP 1/A, WP 3/D, WP 4/E and WP H. These will be discussed briefly below.

Workpackage 1 uses pie charts in a number of deliverables, such as Deliverable 1.3 (Figure 3, 4 & 6 on page 41, 43), and geolocation maps in Deliverable 1.1 (Figure 5 on page 25) and Deliverable 2.2 (Figure 6 on page 35).

In the research on smart energy (WP3/D) a heatmap of Canadian hourly energy consumption per day was shown in Figure 4 (Del 3.1, page 10). In Figure 7 (Del 3.1, page 19) the weekly patterns of energy consumption are shown to illustrate the differences between household and business consumption patterns in Estonia. Deliverable 3.1 also contains a number of geographical maps showing various energy consumption patterns (Figures 8-11).

WP4/E focusses on AIS data. Here a number of maps are used to illustrate the trips ships make (Del 4.3, Figure 6 & 8, page 16 and 17), the location of various harbours (Del 4.3, Figure 7, page 16) and the movement of ships therein (Del 4.2., Figure 3, page 11). A plot illustrating the many geolocation errors in AIS data has not been included in any of the deliverables but has been shown in a number of presentations on this topic during ESSnet BD meetings.

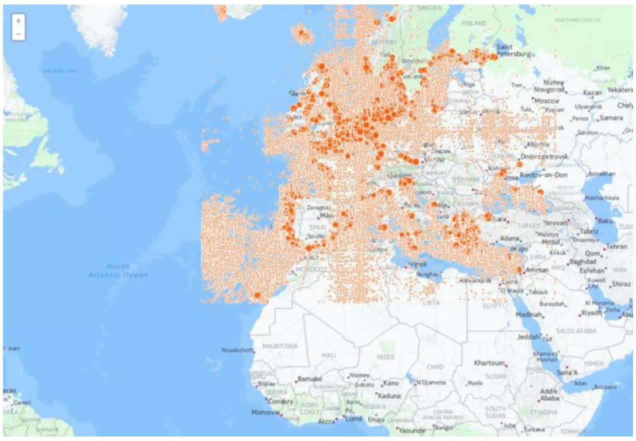


***Figure 4**. Plot of the raw geolocation data of ships European AIS data studied in WP4 of the ESSnet Big data. The plot reveals that sometimes ships are located on land. This is clearly an error.*

In WP H Earth observation studies are described in which satellite images and admin data are combined. It should come as no surprise that Del H1 contains a number of maps illustrating the content of the satellite images (Figure 4.6 and 4.8 on page 15 and 16), maps showing the combination of satellite and admin data (Figure 4.9 and 4.10 on page 17 and 18) and some delineating city maps (Figure 5.1 and 5.2 on page 29 and 30).

3.3 Discussion

This chapter is all about the importance of visualizing data to extract information and to obtain insights. Humans need this because a visual summary makes it easier to identify patterns and trends in data; especially when the amount of data is large. It fits the way the human brain works. The examples discussed above also indicate that visualization methods can also be applied to discover errors in the data.

4. Combining sources

The statistical offices are increasingly using multiple alternative sources jointly for the production of statistical indicators. Not only does this approach make it possible to increase the information otherwise available from a single source, but it also increases the quality of each single source, improving both the quality of its measurements and of its coverage in terms of statistical target units. Indeed, combination of sources allows the NSIs to increase the amount of possible analysis on joint variables that are never observed on the same sources without increasing the statistical burden on respondents. It also allows NSIs to evaluate the quality of measurements on the same variable by comparing the different sources. In fact, when the same variable is measured in multiple data sources on the same units, the multiple observations can be used to identify and correct errors and to harmonize variables definitions. An approach that is often used in practice at NSIs is micro-integration (Bakker, 2011), or alternatively, it may be possible to model the measurement errors, e.g. Scholtus, Bakker and Van Delden (2015), Guarnera and Varriale (2016). At the same time, when a single source is not complete in terms of the recorded units of the target population, combination of sources is a standard strategy to address the issue. See for example Wolter, (1986) and Ding and Fienberg (1994) for population size assessment; see Guidelines on the use of registers and administrative data for population and housing censuses prepared by the Task Force on register-based and combined censuses. Combination of sources can be useful to adjust selectivity often present in Big data sources (see Di Consiglio and Tuoto - Chapter 6 in Deliverable 2.2 Makswell project, 2020)

The combination of sources can be at unit or aggregate level. When the combination is through record linkage we have different scenarios: when unique unit identifiers are available, such as unique personal identification numbers, deterministic linkage can be used (Herzog, Scheuren and Winkler, 2007); otherwise probabilistic linkage might be used. The classical paper on probabilistic linkage is by Fellegi and Sunter (1969). However, when a combination of sources is carried out by linking units, as for example to adjust for coverage or selectivity of the single source, the challenge when dealing with Big data is particularly complex. First, big data records are usually different from the target statistical units and moreover contain very little identifying information on their record. An analysis of common difficulties to face when linking big data with survey and administrative data is in Tuoto *et al*. (2018). In fact, a very relevant issue when treating big data sources is first the identification of the target statistical units from the records of the big data source, sometimes events being recorded instead of target units as in the surveys sources. The effect of the errors in this process, such as duplication of units and/or wrong association to the same statistical units of records, could make the application of statistical methods for analysing variables in different sources, to improve the coverage and adjust the selection bias in big data, very challenging. A discussion of these challenges is, for example, in deliverable 2.3 of the MAKSWELL Project (Makswell, Del 2.3). The effect of linkage errors on estimation of population size is overviewed e.g. in Di Consiglio, Tuoto, Zhang (2019), whereas a sensitivity analysis on the effect of linkage errors on the statistical linear and logistic regression is in Di Consiglio and Tuoto (2018).

4.1 Linking units

4.1.1 Reasons for linking

All the WPs recognize that alternative data sources need to be integrated into the production of official statistics to let these data gain solid ground; and therefore they adopt and recommend linking activities for the objectives identified in the previous paragraph. The unit deduplication is explicitly considered for the online job advertisements in WPB; this operation is included in the phase 1 of the throughput phase, i.e. in the transformation of raw data in statistical data, while the other linking activities (combination with other sources) are executed in the phase 2 of the throughput, once the raw data have been already transformed and usable as statistical data.

In WPB, linkage is proposed between online job advertisements and Job Vacancy Survey micro data or/and Statistical Business Register data, in order to understanding coverage issues of scraped OJAs data, since for many reasons OJAs cannot be considered as exhaustive of the total number of job vacancies that exist at a specific reference date. In WPC - web scraping enterprise characteristics- the importance of identifying websites and companies in the business registers is underlined. For the scraping, the authors often start right from the list of companies in the register, then the identification activities consists mainly in checking if the link obtained from the web is the correct one or not. In WPD, smart meters for electricity consumptions, the linkage between smart meters data and business register is motivated by the need of capturing the NACE (Nomenclature of Economic Activities) code for local units, to produce consumption statistics by economic activity. In WPE, AIS data, the purpose of the linkage between AIS data and administrative data on vessels provided by Maritime Transport Statistics is to identify type and size of the vessels present in the AIS, to produce statistics according to the disaggregation required by Eurostat regulations. WPG, financial transaction, recognizes the increased value of financial transaction data when combined with national account data and business statistics data. WPH, earth observations data, is supporting the added value of combining several primary and administrative sources with earth observations data. The combination of the sources is recognized as crucial. Generally it is proposed at area level or via geocoding/georeferencing. These activities are often a work in progress or still under study in many NSIs. WPJ, tourism statistics combining several sources, is explicitly focused on developing methods for combining several sources to produce tourism statistics. The most widely experimented combination is the one between data obtained from web scraping with the survey frame of tourist accommodation establishments, with the aim of using data from web scraping to improve the frame of the survey of tourist accommodation establishments.

4.1.2 Linking units and linking variables

A common aspect for the linkage procedures in the different WPs is that they involve business/enterprise as linking units. This makes the linkage quite difficult and problematic, due to the intrinsic complexity of real world units such as the enterprises, the different way in which they can be described in different sources. In facts, especially for large enterprises with many establishments and local units, it is possible to distinguish and identify legal units, administrative units, physical units, statistical units. Moreover, some enterprises are subjected to high demography, with closures and reopening, fusion and acquisition, splitting. In addition when dealing with the jobs portal for vacancies, in WPB, one should be aware that jobs are often advertised through private employment agencies and the employer is not usually identified in the advertisement or identified with. These factors jeopardize the success of the linkage.

For business identification, most of the WPs use strong identifiers like Register number, name, and address (including building, postal code, longitude and latitude coordinates). The complexity and the inconsistency in the names and address fields are the biggest obstacle. Therefore, geographical coordinates seem to be a promising solution, encouraging the use of technology that allows geolocation of address data.

Linkage results based on address as linking variables are something affected by uncertainty where on the same address are several big data units (e.g. metering points in WPD) or several statistical units (e.g. companies in WPD, tourism accommodations in WPJ).

Due to the intrinsic complexity of the units and variables involved in the linkage, it is largely recognized that in order the linking activities to be effective, some additional data pre-processing and additional tools are required, in terms of standardization, coding, parsing of the information achievable via the big data sources.

4.1.3 Linking methods

The methodology to execute deduplication and linking activities is mainly the record linkage. The method used is often based on deterministic decision and sometimes on machine learning classification techniques (logistic regression, classification trees).

In WPG some countries identify a good solution via pseudonymization of data that allows linking register-information at the NSIs on payer (person) and enterprise (receiver) with electronic transaction, using the bank account register from the Tax authorities.

4.1.4 Suggestions for improvements on methods

A WPs common suggestion is the need of involving specialists of record linkage and business registers at NSIs, given that linking activities and the knowledge of the enterprises structures play a crucial role for producing statistical outputs. The record linkage experts should be involved in the multidisciplinary team producing statistical output from big data sources.

A second important note regards the advantages of linking activities and the full exploitation of the linking efforts. It is well recognized that the linked sources can enrich each other from the variables perspective; however, it should be stressed that they can improve each other from the unit perspective as well. In fact, they might be affected by under-coverage regarding the units. Therefore, after linking they seem to meet the conditions for the application of proper methods to face under-coverage, e.g. capture-recapture methods, useful to assess the unknown target population size. Many WPs' frameworks seem particularly promising for this kind of improvement.

4.2 Examples of combining sources

In a number of WP's of the ESSnet Big Data several data sources were combined.

In WP1 various sources of online job advertisements were combined and checked for duplicates (WP1.3, p.16). The conceptual model behind this is shown in Fig. 3 of Del. 1.2 (p. 9).

WP2/C studied enterprise characteristics. They ran into a number of methodological challenges when they combined web scraped and survey and administrative data. The need to obtain (and check) the correct URL for each company lead to the development of an URL inventory approach (Appendix 7.1 in Del 2.2) in which methods were developed to check if the correct URL (website) was obtained for each company.

In WP3/D smart meter data was studied. Here, linking smart meter data (the metering point) with administrative data proofed challenging (Del. 3.2, p. 16). Both the business register and the dwelling register were used as administrative input. Goal of combining these sources was to link the observed unit (a metering point) to the statistical unit; a company, household or dwelling. Linking was especially complicated when address data was used as there is a many-to-many connection with this key. Different approaches and strategies were developed and described (Del. 3.2).

In WP4/E a number of AIS data sources were combined that need to be enriched with administrative data of ships. The admin data contains more detailed ships data, required for official statistics. This work is currently under development (Del E1, p.10).

In WP 7 combining sources strategies have been discussed in Del 7.1/7.2/7.3 for population statistics (p. 28), tourism and border crossings (p. 51) and agricultural statistics (p.71). The latter is performed for satellite images and administrative data in WPH. Here, these sources are combined using geolocations (Del H1, p. 17). This resulted in an enriched crop and cadastral parcel data set to be used for classification purposes. Similar approaches were used for the other subtasks in WPH.

4.3 Discussion

Combining data is important as it not only increases the various application of a particular data source but will also stimulate the extraction of new information. Especially for big data, finding ways to combine their data with other big data sources and the sources traditionally used by NSI’s, i.e. survey and administrative data, will increase its use at it may provide ways to get a grip on the population composition of the source. It will also increase the ways in which the big data source can be used for official statistics. Involving multidisciplinairy teams in this linking effort is a suggestion to increase the chance of succesfull linking.

5. Dealing with errors

Not all data in a Big Data source is of high quality. Data may be missing or may simply contain an invalid or imprecise value. In the ESSnet Big Data I and II a number of cases are described in which these errors affect potential outcomes. To enable the creation of high quality official statistics, methods are needed to identify and deal with or correct the poor quality data. This must not only happen at the variable but also at the unit level.

This section is organized along the quality dimensions looked at for throughput phase (part1 from raw data to statistical data) in the quality guidelines (Deliverable K3) and gives examples of methods applied to reduce actual occurring errors and the methods or strategies used to tackle them.

5.1 Error dimension – Coverage

One of the most crucial questions when using Big Data for Official Statistics is the question of coverage. The definition and study of coverage errors require the definition of the target population, that should be explicitly identified in terms of type, time and place.

One example of a coverage error occurs in the form of "ghost vacancies" when webscraping for online job advertisements (see Deliverable 1.2 p.7 of WP1 ESSNet Big Data I). These listings do not correspond to a real job vacancy and are therefore overcoverage for this population. Some of these "ghost vacancies" follow certain semantic patterns therefore through text mining and NLP it should be possible to identify and remove them.

Some similar phenomena can occur for AIS data, not all ships in the data set are relevant for a certain statistical product, e.g. transport statistics. One methodological solution to deal with this overcoverage issue is to use webscraping techniques to match the ships in the data set with external registers. The information from these external registers could then be used to verify the inclusion of a certain shop in the target population. (WP4 Ghost ships/Reference frame Del 4.2. p. 13).

5.2 Error dimension – Comparability over time

Another main concern about introducing Big Data in Official Statistics is the process generating data. This may not be stable over time as technology or behaviour might change. One data source might get irrelevant whereas another gets important for a certain phenomenon.

Missing data can lead to problems in comparing observations over time, this is for example the case in traffic loop data (WP6 Del 6.6 p10). A traditional method to deal with missing data is imputation. For a big data source with a very high level of detail, it might be useful to impute not on the lowest possible level, but on a macro level to avoid precision errors due to small scale data. As it was done in this example on a monthly level for missing road sensor data.

5.3 Error dimension - Measurement errors

Measurement error is the difference between the true value of the measurement and the value obtained during the measurement process. For some Big Data sources, this might actually be a physical measurement or transmission error.

In AIS data a measurement error of the position of a ship can lead to a significant error in the estimate of, e.g., port visits. For some measurement of the location was quite far off causing ships to be outside of the port in one measurement and inside mere seconds later. This problem was solved with applying a median filter over the positions over time, in this example the time interval was set to 10 minutes (WP4 Del 4.2, p 12).

Another example of a measurement error as noise in data are satellite pictures (WPH Del H1, p 15). Similar to the previous example a median filter is applied as method to solve/reduce the problem. This time the filter is applied over the spatial dimension averaging over a 3 x 3 window.

5.4 Model errors (raw data to statistical data)

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. More on this is discussed in section 6.6 of this deliverable.

5.5 Discussion (Link with Quality deliverable and relation with user needs)

The link to the quality related deliverables will be discussed in the final version of the methodological report.

We are also considering adding a section on user needs. This is crucial when talking about quality e.g., for one user it may very well be good enough to know that the unemployment rate is somewhere between 5-15 percent but for another user that may not be good enough. The latter user may need a better precision say unemployment being somewhere between 7-9 percent. Another example is the concept that we get from a big data source e.g., online job vacancies that may be good enough for some users while others may want the exact ILO-definition (which may not be what they get anyway from a survey due to measurement errors etc). Striking this balance of different user needs is one of the main challenges for producers of official statistics.

Throughput Phase 2

6. Inference

This important task is close to the end of the statistical process. The paper of De Broe *et al*. (2020) and Salgado and Oancea (2020) provide overviews of the ways in which Big Data can be used in official statistics production and the accompanying challenges in the estimation process. Let it be clear that not all of the 13 examples listed in table 1 have already reached this stage. This makes it quite challenging to provide general recommendations based on just a few examples.

6.1 General considerations

In general, it is clear that Big Data can be used in a number of ways to produce official statistics. The cases that can be discerned are:

i) Use Big Data as the most important input for the estimation process,
ii) Use Big Data as an additional source
iii) Use Big Data as part of a combination of a number of sources that all contribute more or less equally to final output.

A more detailed overview is possible but the general division shown above suffices to illustrate a number of essential differences. When one looks at the 13 examples listed in table 1 it becomes clear that each of these can be assigned to a particular use case. Examples of using big data as the most important source (case i) are: Traffic intensity statistics, innovative companies, AIS-based maritime statistics and electricity consumption statistics (persons and businesses). Examples of using big data as an additional source (case ii) are: AIS-supported inland waterway statistics (in the Netherlands), Enterprise characteristics and Online job vacancies as proxy for vacancies statistics. Examples of using big data together with other equally important sources (case iii) are: CPI based on a combination of multiple sources (most often scanner and survey data), agricultural statistics based on satellite/areal pictures and administrative data and Innovative Tourism statistics based on multiple sources.

The methodology applied in the last two cases, for a large part, greatly resembles those traditionally used in official statistics. For case ii in particular, when it is so that survey data forms the basis, (standard) survey methodology can be applied. It stands to reason that this methodology need not be discussed in much detail here. Readers are referred to Särndall *et al*. (1992), Beręsewicz et al. (2018), and the Makswell project (Del 2.2 & 2,3) more details on those views. However, the reader also needs to realize, that in all cases mentioned information needs to be extracted from Big Data to enable its use. This may require specific Big Data methodology of which some have already been described in the previous chapters of this report. What remains are a number of important considerations when using Big Data as the main source of input, the use of models for Big Data and a number of essential topics identified by the members of WPK. These are discussed in the remainder of this chapter.

6.2 Big Data as the main source of input

When the data in a Big Data source provides the main input for the statistic produced, in essence two different ways of using the data can be discerned. The first is that one aims to include as much of the data generated in the final output, as this enables very detailed statistics on small areas or subgroups. This approach is typical for smart meter data (Del 3.1), AIS based transport statistics (Del 4.3), mobile phone data (Del 5.3 and WPI) and Road sensor data (Del 6.1, Puts *et al*., 2019). In all these cases, in the end, use a model to infer from the data. The model is used to compensate for missing data, data of low quality, and/or differences in the population composition of the Big Data source and the target population. This need becomes particularly obvious when very detailed, regionalized, statistics are being produced. Here, because of low coverage in some areas, the original abundance of data reduces to a small amount (for some areas) and hence one need to compensate for this.

When Big Data is used as the main source of information, extracting knowledge from such data does not mean that a total new way of drawing inference needs to be developed. However, because of potential selectivity issues, special attention needs to be paid to causes of bias. Variance is considered less of an issue here as large amounts of data are being used. Therefore, a researcher needs to pay attention to the effects of (low) data quality and any of the decisions made during the processing of Big Data on the final outcome. In a perfect world, where Big Data contains perfect quality data including all units, simply adding all values up should suffice to obtain the total of the population. In reality, however, missing data (for particular areas or units) or including data from units not belonging to the target population, need to be dealt with. Because of this, a seemingly precise Big Data based estimate may still be way off (Buelens *et al.*, 2014, section 6.6).

The difficulty when huge amounts of data are used for statistical inference is that the data in the source needs to cover the entire target population to enable model-free inference. This is (likely) what the (in)famous paper on “The End of Theory” in Wired magazine (Anderson, 2008) wanted to indicate. However, in practice, model-free inference is hardly possible as quality issues or simple data delivery issues may prevent the continuous inclusion of the entire population (Puts *et al*., 2017). Despite these considerations, there are examples –that could be verified- that indicate that simply following the adage of including as much data as possible (and correcting for the bias to the best of one’s knowledge) can provide estimates very close to those obtained by other, more traditional ways, of inference. Estimating the number of large innovative companies is an example of that (Daas and van der Doef 2020). In this study, the texts on the websites of all large Dutch companies were scraped and classified with a model to determine if they were innovative or not. The final results obtained were nearly identical to the survey derived number; i.e. 19,276 ± 190 for the web based method vs. 19,916 ± 680 for the survey based method.

Applying a model enables one to deal with the real-world fluctuations that occurs in a data-deluged world. From the above it is clear that Big Data models need to be developed that are able to:

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

Current research focusses on these exiting areas. Studies on non-probability sampling touch the unknown designs issue (Buelens et al., 2018) as do catch-recatch studies (Del 5.3). Applying Bayesian inference methods is another suggested approach (Del 5.3). However, one should not assume that Big

Data is simply a large sample (Doherty, 1994) and be careful when (mis)using the 'representativity' argument (see section 6.4). Another important consideration is that one's needs to realize that "all models are wrong but some are useful" (Box, 1979).

6.3 Conceptual differences

Not every data source measures exactly the same 'thing', i.e. measures the same concept. The concept included in a Big Data source may differ from what is measured by a survey or an administrative source. It's best to illustrate this difference with the job vacancy example from the ESSnet Big Data. In both the first and second ESSnet Big Data on-line job vacancies were studied and compared to the vacancies measured in the official European job vacancy survey. Both sources can be used to produce statistics on labour demand, but the survey is setup to directly measure job vacancies whilst a web portal provides information on on-line job advertisements, i.e. job advertisements published (online) by a company in search of a new employee (WP1, Del 1.1 p. 4). Hence both sources do not measure the exact same concept. It is essential that a researcher using Big Data is aware of conceptual differences and checks if and where they occur to assure the proper comparisons are made at the start of a study. Let it be clear that this differs from the indirect measurement of a concept described and discussed in section 1.2.

During meta-analysis studies, i.e. a statistical analysis in which the results of multiple scientific studies are combined, one also often needs to deal with concepts that differ. Matching those concepts is usually referred to as 'harmonization' in these kinds of studies (Griffith *et al*. 2015). In various data sources, identical concepts may be operationalized in different ways. For instance, age may be a continuous variable in study 1 and a categorical variable in study 2. Even categorical variables may be defined differently; e.g. gender 0 is male in study 1 and female in study 2. When the data from both studies need to be combined, the variables need to be harmonized. This can, for instance, be done by recoding the gender variable from study 1 into the coding used in study 2 and converting the continuous age variable from study 1 into the categories used in the study 2. Harmonizing variables may, however, not always be that easy and it is challenging to completely harmonize away all possible effects. For Big Data, where sometimes the metadata is not very clearly described this can quite a challenge. Also, it may not always be very easy to derive how a concept is defined in a Big Data source. The large amounts of data available can be very useful to derive that.

In the above examples concepts were compared at a single point in time. However, it is also important to compare the concept measured in a single Big Data over time. This to get an idea of its stability.

6.4 Representativity (selectivity)

A central issue in designing the statistical methodology with new digital data sources is the so-called representativity of the data set(s). For example, with MNO data used to estimate present population the expert will immediately enquire about how much population is really represented in the data both with data from a single MNO and from several MNOs. For other Big Data sources the situation is similar. This issue indeed involves several subtleties and has been referred to in many disguises in the literature (representativity, selectivity, selection bias, coverage, extrapolation, . . .).

In our view, the central issue revolves around the problem of **statistical inference**, understood as Box (1958) wrote:

> A statistical inference will be defined [. . .] to be a statement about statistical populations made from given observations with measured uncertainty. An inference in general is an uncertain conclusion. Two things mark out statistical inferences. First, the information on which they are based is statistical, i.e. consists of observations subject to random fluctuations. Secondly, we explicitly recognise that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved. Fisher uses the expression 'the rigorous measurement of uncertainty'.

Clearly, when considering the issue of representativity we are enquiring about the connection between our observed data set(s) and a target population under analysis. In this line of thought, we prefer avoiding the use of the term *extrapolation*, since in Mathematics this refers to the calculation of an estimate of the value of some function outside the range of known values (a very concrete problem), where no consideration about uncertainty, random fluctuations, and statistical nature of populations is explicitly made. However, many other terms such as selectivity, selection bias, undercoverage, overcoverage, . . . are meaningfully rightful since they refer to different aspects highly entangled and embedded within the statistical inference problem.

Also, we want to underline how uncertainty and the assessment of uncertainty is inextricably linked to the problem of statistical inference. In rigour, population estimates avoiding measures of uncertainty are not a solution to the statistical inference problem connecting the observed data with the target population.

Different understandings and approaches to statistical inference (thus to representativity) are rooted on the underlying interpretation of probability and in general on the use of probability theory beneath the inferential statements (Kyburg, 1974). In this line, for considerations regarding the production of official statistics it seems natural to us to start considering the standard inferential paradigm in this industry: design-based inference and survey methodology, in general.

Certainly, survey methodology is limited with new data sources, but it offers a template mirror for a new refurbished production framework to look at, namely a set of modular statistical solutions for a diversity of different methodological needs along the statistical process in all statistical domains (sample selection, record linkage, editing, imputation, weight calibration, variance estimation, statistical disclosure control, . . .). We shall concentrate in this section on the connection between collected data sets and target populations, which is firmly rooted on scientific grounds using design-based inference.

How is representativity dealt with in survey methodology? The short answer is design-based inference. As T.M.F. Smith (1976) already pointed out, the design-based inference seminally introduced by J. Neyman (1934) allows the statistician to make inferences about the population **regardless of its structure**. Also in our view, this is the essential trait of design-based methodology in Official Statistics over other alternatives, in particular, over model-based inference. As M. Hansen (1987) already remarked, statistical models may provide more accurate estimates **if the model is correct**. Apparently, sampling designs avoids making prior hypotheses about the population, which is sometimes difficult to justify and to communicate.

In mathematical terms, in the practice of survey methodology this essential trait translates into the use of (asymptotically) design-unbiased linear estimators of the form $\hat{T} = \sum_{k\in s} \omega_{ks}(\mathbf{x})\, y_k$ where s denotes the sample, $\omega_{ks}(\mathbf{x})$ are the so-called sampling weights (possibly dependent on the sample s and on auxiliary variables $\mathbf{x}$) and y stands for the target variable to estimate the population total $Y = \sum_{k\in U} y_k$. A number of techniques does exist to deal with diverse circumstances regarding both the imperfect data collection and data processing procedures so that non-sampling errors are duly dealt with (Lessler and Kalsbeek, 1992; Särndal and Lundström, 2005). These techniques lead us to the appropriate sampling weights $\omega_{ks}(\mathbf{x})$. Sampling weights are also present in the assessment of accuracy (in terms of variance, confidence intervals, etc.).

The common understanding of sample representativity is rooted on sampling weights. The interpretation of a sampling weight $\omega_{ks}(\mathbf{x})$ is extensively accepted as providing the number of statistical units in the population U represented by unit k in the sample s. This combination of sampling designs, linear estimators, and interpretation of sampling weights provides a robust and solid comprehension of representativity in this realm. When sampling designs cannot be used (as with Big Data sources), we lose this combination, hence also our traditional notion of representativity.

Currently, Big Data methodology for the production of official statistics is not mature enough to provide a single substitute replacing our traditional view of representativity for all Big Data sources. More business cases, examples, and Big-Data-based statistical products need to be produced, but this does not mean that rigorous statistical tools do not already exist to face this challenge. Indeed, we claim that the central goal in the production of official statistics in relation to statistical inference is to produce estimates as accurate as possible together with an assessment of the achieved accuracy. Design-based inference is indeed a concrete approach to solve this problem and not to build a notion of representativity, which should be considered a (dangerous) by-product of this approach.

Let us include a short digression trying to show how design-based inference actually produces accurate estimates and really avoids the issue of representativity. Firstly, the notion of representativity was already analysed by Kruskal and Mosteller (1979a,b,c, 1980) showing how imprecise and slippery this notion is. Indeed, a mathematical definition in classical and modern textbooks is not extensively found, providing Bethlehem (2009) an exception in terms of a distance between the empirical distributions of a target variable in the sample and in the target population. This definition has two immediate short comings: (i) we never know the distribution of a variable in the target population (thus reducing the utility in practice), (ii) we may have a sample being representative for a given variable but not for another.

Secondly, the construction of estimators based on the aforementioned interpretation of sampling weights is indeed mathematical flawed and extremely poor and misleading. This construction reasons as follows. If $\omega_{ks}(\mathbf{x})$ amounts to the number of population units represented by the sampled unit k, then $\omega_{ks}(\mathbf{x}) \cdot y_k$ is the part of the population aggregate accounted for by unit k in the sample s. Thus,

$\sum_{k\in s}\omega_{ks}\cdot y_k$ is the estimate of the total population aggregate. This argument is indeed behind the restriction upon sampling weights for them not to be lesser than 1 (interpreted as a unit not representing itself) or for them to be positive in sampling weight calibration procedures (see e.g. Särndal, 2007). In our view, the interpretation of a unit k in a sample as representing ω_{ks} units in the population can be impossible to justify even in such a simple example as a Bernoulli sampling design of probability $\pi = \frac{1}{2}$ in a finite population of size $N = 3$: if, e.g., $s = \{1, 2\}$, how should we understand that these two units represent 4 population units?

Thirdly, the randomization approach does allow the statistician not to make prior hypotheses on the structure of the target population (Smith, 1976). But this does not necessarily entail that the estimator must be necessarily linear. Given a sample s randomly selected according to a sampling design $p(\cdot)$ and the values $\mathbf{y}$ of the target variable, a general estimator is any function $T = T(s, \mathbf{y})$, being linear estimators a specific family thereof (Hedayat and Sinha, 1991). Thus, nothing prevents us to use more complex functions provided we search for low mean square error. As a matter of fact, a linear estimator may be viewed as a homogeneous first-order approximation to an estimator $T(s, \mathbf{y})$ such as $T(s, \mathbf{y}) \approx \sum_{k\in}\omega_{ks}\, y_k$, but why not a second-order approximation $T(s, \mathbf{y}) \approx \sum_{k\in}\omega_{ks}\, y_k + \sum_{k,l\in s}\omega_{kls}\, y_k y_l$? Or even a complete series expansion $T(s, \mathbf{y}) \approx \sum_{p=1}^{\infty}\sum_{k_1 \dots k_p s}\omega_{k_1 \dots k_p s}\cdot y_{k_1 \dots k_p s}$ (see e.g. Lehtonen and Veijanen (1998))?

Fourthly, the linearity of design-based estimators is indeed motivated by a non-statistical reason, which is basically communication. Multivariate estimates of target populations need to be produced, usually broken down according to diverse classification variables (sex, age group, economic activity, etc.). These classification variables are common among many surveys in a statistical office. Thus, given the public dimension of Official Statistics usually disseminated in numerous tables, **numerical consistency** (not just statistical consistency) is strongly requested across all disseminated tables, even across different surveys. For example, if a table with R+D investments is disseminated broken down by economic activity (e.g. NACE code) and another table with number of employees is also disseminated broken down by economic activity, the number of total business units by economic activity inferred from both tables must be **exactly** equal. Linear estimators can be made straightforwardly satisfied this restriction by forcing the so-called multipurpose property of sampling weights (Särndal, 2007), which amounts to use the same sampling weight $\omega_{ks}(\mathbf{x})$ in the estimator of all population quantities in a given survey. For inter-survey consistency, sometimes calibration techniques of sampling weights are used, even dangerously (not all target variables *y* hold the same correlation with auxiliary variables **x**). This circularly reinforces the above interpretation of sampling weights and the flawed reasoning about the construction of linear estimators.

As a final element of this digression, let us remind that sampling designs are commonly thought of as a life jacket against model misspecification. For example, even not having a truly linear model between the target variable *y* and covariates **x**, the ratio estimator is still asymptotically unbiased (see e.g. Särndal *et al.*, 1992). But (asymptotically) design-unbiasedness does not guarantee a high-quality estimate. A well-known example can be found in Basu's elephants story (Basu, 1971). Dropping out issues about the inferential paradigm, this story clearly shows how a poor sampling design drives us to a poor estimate, *even using an exactly design-unbiased estimator*. A design-based estimate is good **if the sampling design is correct**, something which is not duly assessed in practice (what's the equivalent of a model goodness-of-fit for a sampling design?)

To sum up, design-based methodology produces accurate estimates by resorting to asymptotically unbiased estimators with a variance as low as possible upon using auxiliary information both in the construction of the sampling designs and the estimators. Any consideration about sample representativity is indeed a by-product which must be rigorously understood in the theoretical context. We defend the idea that any estimation procedure with Big Data source must also pursue the construction of accurate estimates with due measures of uncertainty.

To end, apart from a rigorous understanding of representativity, Big Data sources also bring an important novelty regarding the nature of statistical outputs potentially produced with them. At least, we envision two complementary directions widening the scope of official statistical products. On the one hand, the data deluge is bringing the opportunity to go beyond the traditional scope of estimation in a finite population, i.e. beyond population totals and functions thereof. Since we are to use statistical methods, they can also be used to understand, model, and report on random phenomena (e.g. analyses using social media and/or web scraped data). On the other hand, this abundance of data makes us envision the analysis of interactions between unit populations (also subjected to randomness), a novel piece of information not covered by traditional sources. In all these regards, the representativity issue, i.e. the quest for accuracy, will also be present.

6.5 Sampling

Many statisticians were raised and trained in the golden age of survey methodology which makes them huge fans of studying samples. This 'paradigm' is not always the best to follow in the case of big data as indicated by a quote from the paper of Franke et al. (2016).

> "Sub-sampling from a large dataset is often presented as a solution to some of the problems raised in the analysis of big data; one example is the 'bag of little bootstraps' of Kleiner *et al*. (2014). While this strategy can be useful to analyse a dataset that is very massive, but otherwise regular, it cannot address many of the [Big data] challenges that were presented …."

Simply put, sampling of a Big Data set to reduce its size, with the aim to speed up of processing and analysis, only increases the variance of the estimates of the phenomena, thereby forfeiting the benefits Big Data has to offer (Daas and Puts, 2014). This becomes even more relevant when the phenomenon studied follows a power-law distribution, such as a Zipfian one (Powers, 1998). An example of the latter is the occurrence of words in text corpora (see Figure 6.1). Their rank-frequency distributions are very well approximated by a Zipfian distribution. Studying samples of data with such distributions will seriously underestimate or even miss the less frequently occurring words.

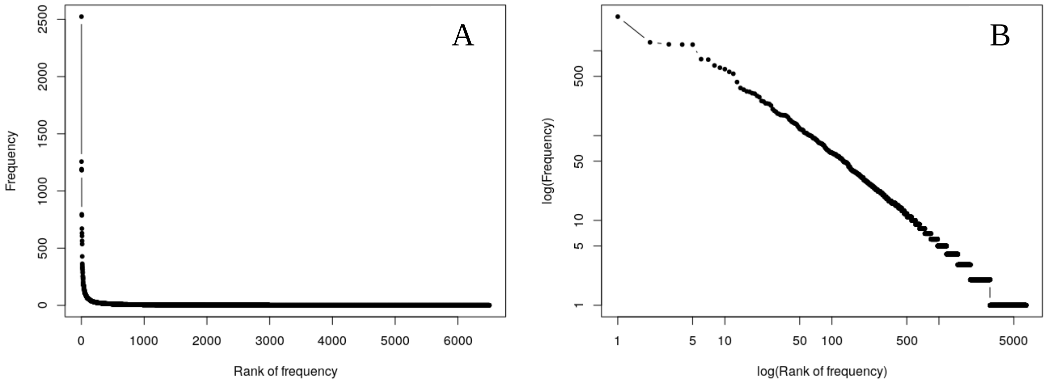


Figure 6.1 Frequency plot of the words in the book “The Hitchhikers Guide to the Galaxy” on a normal (A) and a log-log scale (B) demonstrating a typical Zipfian distribution. The most occurring word is ‘the’ (2524) followed by ‘of’ (1257) and ‘a’ (1191). 3507 words are only used once.

It is therefore advised to -whenever possible- study the data in a big data source as a whole, which is certainly doable in the current day and age considering the increased access to computer power and the parallelizing capabilities of modern computers. If a data source cannot be studied in a single run, considering splitting it up in parts that are separately analysed and combine the findings afterwards. However, this does not mean that sampling is bad. It is still a great way to quickly gain insight in the content and quality of a Big Data source during an exploratory study. However, it should merely be considered as a first step in such a study.

6.6 Accuracy assesment (Bias and variance)

The accuracy of a statistic is measured by its variance and bias. The variance is inversely related to the size of the data used. Because of the large volume of big data, it is to be expected that the variance is rather small. But one should be aware that it is also affected by the similarity between the concept a researcher wants to measure and the concept provided by the source (see section 6.3).

Bias is a much greater concern for big data sources. We discern between selection, measurement and model bias here. The first two biases are discussed first. The more the units in a big data source resemble (represent) the target population, the less of an issue selection bias becomes. Selection bias is expected to be zero when all units of the target population are included and the data of the units not belonging to them are successfully removed. Obviously, when limited information is available on the units included in big data this is a challenging task. This is the main reason why many researchers try to link big data with datasets of known population composition (Chapter 4). The measurement bias in big data-based statistics predominantly depends on the extent to which the conceptual variables the researcher wants to measure are (correctly) provided by the data in the source. Problems may arise here when the variable(s) of interest is derived from the variables available and is, hence, indirectly measured.

In this respect, the mathematically formalization of the error of estimates derived for non-probability samples by Meng (2018) forms in interesting view point for big data based estimates. Non-probability samples should be interpreted here as data sets composed of units with unknown inclusion probabilities which are -very likely- selective for the target population. According to Meng (2018), the error estimate for such samples is composed of three components: 1) a measure for the correlation between the target variable and the response/recording indicator, 2) a measure for the fraction of the target population covered, and 3) a problem difficulty measure which is the standard deviation of the target variables. This shows that selection-bias becomes a serious issue when the correlation between the target variable and the response indicator is low and when the fraction of (target population) units included in the source is low. From the unit point of view, the first issue is predominantly affected by the difference between the composition of the units in the source and those of the target population. Both can be reduced by maximizing the inclusion of (target population) units in the sources. It is interesting to note that this approach has been successfully applied in the webpage based detection of innovative companies in the Netherlands (Daas and van der Doef, 2020). Here, the big data (web based) based and the survey based estimates where nearly identical. The third component identified, the problem difficulty measure, affects the variance of the estimate.

Model bias is relevant when a model is used in obtaining the final estimate. The effect is easiest to explain for a classification model. Suppose as researcher is using a model to classify if a unit belongs to group A or B. The researcher, obviously, wants to use a model that is as accurate as possible for this task. In this case, a high accurate model means that nearly all of the examples in the test sets (a dataset with units of known composition) are correctly classified by the model; e.g. a true A-unit is classified as A and a true B-unit is classified as B. However, this will not be the case for all units, so some A's are classified as B's and the other way around. If these two misclassified groups are not of equal size, the model outcome will be biased towards one of these groups. Hence, the model overestimates either A or B; i.e. the model is biased. By using different metrics during model training one can try to reduce this effect as much as possible (Kuhn and Johnson, 2013) and/or one can determine the unbalance of the model and subsequently correct the estimate for it (Meertens et al., 2019).

6.7 Additional considerations

Causality is an important topic for future research, certainly in the area of Big data. It is important because causation indicates a relationship between two variables where one is affecting the other. Since the data generation mechanism of Big data is usually unknown, causal studies may shed light on this and help to extract more and more reliable information from such sources. It will also extend any of the model based inferences in this area.

Another topic that needs to be mentioned here is the advantages of using a Bayesian statistical approach for Big data sources. Examples of this are the work of WPI and the model used to correct for missing data in the road sensor work (Puts et al., 2017). A Bayesian approach is also beneficial whenever a Bog data source is used that indirectly measures the concept of interest.

6.8 Discussion

Statistical inference is key for official statistics. In the case of Big data based statistics is becomes even more challenging as nearly always a model is involved, the population can be quit dynamic, the data generation mechanism may be and the data could be indirectly used. The most important issues have been listed above and their current state-of-art has been discussed. In addition two topics have been identified that certainly will require more attention in future work. All of this to assure the successful application of Big data in official statistics.

7. Conclusions

Round up of document (for the final version).

References

Ahas, R., Tiru, M., Saluveer, E., Demunter, C. (2011). Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. Paper for the New Techniques and Technologies for Statistics conference 2011, Brussels, Belgium.

Allen, C., Hospedales, T. (2019). Analogies Explained: Towards Understanding Word Embeddings. Proceedings of the 36th International Conference on Machine Learning, June 10-15, 2019. Long Beach, USA. Available at: https://arxiv.org/pdf/1901.09813.pdf.

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

Anscombe, F.J. (1973). Graphs in Statistical Analysis. American Statistician 27(1), pp. 17-21. doi:10.1080/00031305.1973.10478966.

Bakker, B.F.M. (2011). Micro-integration. State of the art. In: ESSnet on Data Integration WP1 State of the Art on Statistical Methodologies for Data Integration (pp. 77-107). (ESSnet). Luxembourg: Eurostat.

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling, Part One*. In: DasGupta A. (eds), Selected Works of Debabrata Basu. Selected Works in Probability and Statistics. Springer, New York, NY.

Beręsewicz, M., Lethonen, R., Reis, F., Di COnsiglio, L., Karlberg, M. (2018). An overview of methods for treating selectivity in big data sources. Statistical Working Paper, Eurostat. Available at: https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf

Bethlehem, J. (2009). The rise of survey sampling. Statistics Netherlands Discussion Paper 09015. Located at: https://www.cbs.nl/-/media/imported/documents/2009/07/2009-15-x10-pub.pdf.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2015). Feature Selection for High-Dimensional Data. Springer, New York.

Box, D.R. (1958). Some Problems Connected with Statistical Inference. The Annals of Mathematical Statistics 29(2), pp. 357–372.

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In: Launer, R.L., Wilkinson, G.N. (eds.), Robustness in Statistics, Academic Press, pp. 201–236, doi:10.1016/B978-0-12-438150-6.50018-2.

Buelens, B., Burger, J., van de Brakel, J. (2018). Comparing Inference Methods for Non-probability Samples: Inference from Non-probability Samples. Int. Stat. Review. 86(2), pp. 322-343. doi:10.1111/insr.12253.

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

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

Daas, P.J.H., van der Doef, S. (2020). Detecting Innovative Companies via their Website. Statistical Journal of IAOS, *accepted for publication*.

Davis, J.C., Michael IV, L.G., Coghlan, C.A., Servant, F., Lee, D. (2019). Why Aren't Regular Expressions a Lingua Franca? An Empirical Study on the Re-use and Portability of Regular Expressions. In: Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19), August 26 - 30, 2019, Tallinn, Estonia. Doi: https://doi.org/10.1145/3338906.3338909

Di Consiglio, L., Tuoto, T., Zhang L.C. (2019). Capture-recapture methods in the presence of linkage errors. In Analysis of Integrated Data, eds. L.C. Zhang and R.-L. Chambers. CRC Press, Statistics in the social and Behavioral Science Series, pp.39-71, ISBN 9781498727983.

Di Consiglio, L., Tuoto, T. (2018). When adjusting for linkage errors: A sensitivity analysis, Statistical Journal of the IAOS 34, pp. 589-597.

Ding, Y., Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. Survey Methodology 20, pp. 149-158.

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

Eisenstein, J. (2019). Introduction to Natural Language Processing. MIT press, Cambridge, USA.

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data. Working paper, September. Located at: https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

Fan, J., Han, F., Liu, H. (2014). Challenges of Big Data analysis. National Science Review 1, pp. 293-314. doi:10.1093/nsr/nwt032.

Fellegi, I.P., Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association 64, pp. 1183-1210.

Franke, B., Plante, J-F., Roscher, R., En-shiun, A.L., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M.M., Grosse, R., Hendricks, D., Reid, N. (2016). Statistical Inference, Learning and Models in Big Data. International Statistical Review 83(3), pp. 371-389. doi.org/10.1111/insr.12176

Geirhos, R. Janssen, D.H.J., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. CoRR, vol. abs/1706.06969. Located at: http://arxiv.org/abs/1706.06969.

Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as Data. Journal of Economic Literature 57(3), pp. 535-574. doi:10.1257/jel.20181020.

Ginsberg, L., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature 457, pp. 1012-1014.

Griffioen, R., ten Bosch, O. (2016). On the Use of Internet Data for the Dutch CPI. Paper for The Meeting of the Group of Experts on Consumer Price Indices, 2-4 May, Geneva, Switzerland.

Located at: https://www.researchgate.net/publication/324896024_On_the_use_of_Internet_data_for_the_Dutch_CPI

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

Guarnera U., Varriale R. (2016). Estimation from Contaminated Multi-Source Data Based on Latent Class Models. Statistical Journal Of the IAOS, vol. 32, no. 4, pp. 537-544

Hansen, M. (1987). Some history and reminiscences on survey sampling. Statistical Science 2, pp. 180-190.

Hedayat, A., Sinha, B. (1991). Design and Inference in Finite Population Sampling. Wiley.

Helwegen, R., Braaksma, B. (2020). Fair Algorithms in context. CBDS Working paper 04-20. Statistics Netherlands, the Netherlands. Located at: https://www.cbs.nl/-/media/_pdf/2020/22/cbds_working_paper_fair_algorithms.pdf

Herzog, T.N., Scheuren, F.J., Winkler, W.E. (2007). Data Quality and Record Linkage Techniques. New York: Springer-Verlag.

Hotho, A., Nürnberger, A. and Paaß, G. (2005). "A brief survey of text mining". In Ldv Forum, Vol. 20(1), p. 19-62

ISTAT (2020). Social Mood on Economy Index. Methodological note, dd. 06-02-2020. Located at: https://www.istat.it/it/files//2018/07/Methodological_Note.pdf

Jo, T. (2019). Text Mining: Concepts, Implementation, and Big Data Challenge. Kacprzyk, J. (ed.) In: Studies in Big Data Vol 45, Springer, New York, USA.

Kleiner, A., Talwalkar, A., Sarkar, P., Jordan, M.I. (2014). A scalable bootstrap for massive data. J. Roy. Stat. Soc. B, 76(4), pp. 795-816.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems Volume 1, ser. NIPS'12. USA: Curran Associates Inc., pp. 1097-1105. Located at: http://dl.acm.org/citation.cfm?id=2999134.2999257.

Kruskal, W., Mosteller, F. (1979a). Representative sampling, i: Non-scientific literature. Int. Stat. Rev. 47, 13–24.

Kruskal, W., Mosteller, F. (1979b). Representative sampling, ii: scientific literature, excluding statistics. International Statistical Review 47, 111–127.

Kruskal, W., Mosteller, F. (1979c). Representative sampling, iii: the current statistical literature. International Statistical Review 47, 245–265.

Kruskal, W., Mosteller, F. (1980). Representative sampling, iv: The history of the concept in statistics, 1895-1939. International Statistical Review 48, 169–195.

Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer-Verlag New York.

Kyburg, H.E. Jr. (1974). The logical foundations of statistical inference. Dordrecht (Holland): Reidel Publishing Company.

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

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. In Proceedings of the IEEE, pp. 2278-2324.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature 521, pp. 436-444. doi:10.1038/nature14539

Lessler, J., Kalsbeek, W. (1992). Nonsampling error in surveys. New York: Wiley.

Lehtonen, R., Veijanen, A. (1998). Logistic generalized regression estimators. Survey Methodology 24, pp. 51–55.

Li, Y., Yang, T. (2018). Word Embedding for Understanding Natural Language: A Survey. In Guide to Big Data Applications, edited by S. Srinivasan, 83-104. New York: Springer.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G. (2018). Learning under Concept Drift: A Review. IEEE Transactions on Knowledge and Data Engineering 31(12), pp. 2346-2363. doi:10.1109/TKDE.2018.2876857.

Makswell project, deliverable 2.2, (2020) Methodological aspects of using big-data. https://www.makswell.eu/projectoutput/deliverables.html.

Makswell project, deliverable 2.3, (2020). Report on identification of future research needs in terms of statistical methodologies and new data.

Meertens, Q.A., Diks, C.G.H., van den Herik, H.J., Takes, F.W. (2019). A Bayesian Approach for Accurate Classification-Based Aggregates. Proceedings of the 2019 SIAM International Conference on Data Mining, May 2-4, Calgary, Canada. p. 306-314. doi:10.1137/1.9781611975673.35.

Meng, X.L. (2018). Statistical paradises and paradoxes in big data (I). The Annals of Applied Statistics, 12, pp. 685-726.

Nadeau, D., Satoshi, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. R. Stat. Soc. 97, pp. 558-625.

Nurmi, O., Piela, P. (2019). The Use of Mobile Phone Data in Tourism Statistics. Paper for the 62nd ISI World Statistics Congress 2019, STS-571 session, Kuala Lumpur, Malaysia. Located at: http://media.voog.com/0000/0042/3095/files/STS0571_Pasi.pdf

Powers, D.M.W. (1998). Applications and explanations of Zipf's law. In D.M.W. Powers (ed.) New Methods in Language Processing and Computational Natural Language Learning, ACL, pp 151-160. Located at: https://www.aclweb.org/anthology/W98-1218.pdf.

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

Puts, M.J.H., Daas, P.J.H., Tennekes, M., de Blois, C. (2019). Using huge amounts of road sensor data for official statistics. AIMS Mathematics 4(1), pp. 12-25.

Rosebrock, A. (2019). Deep Learning for Computer Vision with Python, 2nd ed.PyImageSearch.com.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. Survey Methodology 33, 99–119.

Särndal, C.-E., Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester: Wiley.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. New York: Springer.

Scholtus, S., van Delden, A., Bakker, B. F. M. (2015). Modelling measurement error to estimate bias in administrative and survey variables. Paper presented at New Techniques and Technologies for Statistics, .

Statistics Netherlands (2015). A13 busiest national motorway in the Netherlands. Statistics Netherlands report. Located at: https://www.cbs.nl/en-gb/background/2015/31/a13-busiestnational-motorway-in-the-netherlands.

Serrano-Guerrero, J., Olivas, J.A., Romero, F.P., Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences 311, pp. 18–38.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. J. R. Stat. Soc. A 139, pp. 183-204.

Szeliski, R. (2010). Computer Vision: Algorithms and Applications, 1st ed. Berlin,Heidelberg: Springer-Verlag, 2010.

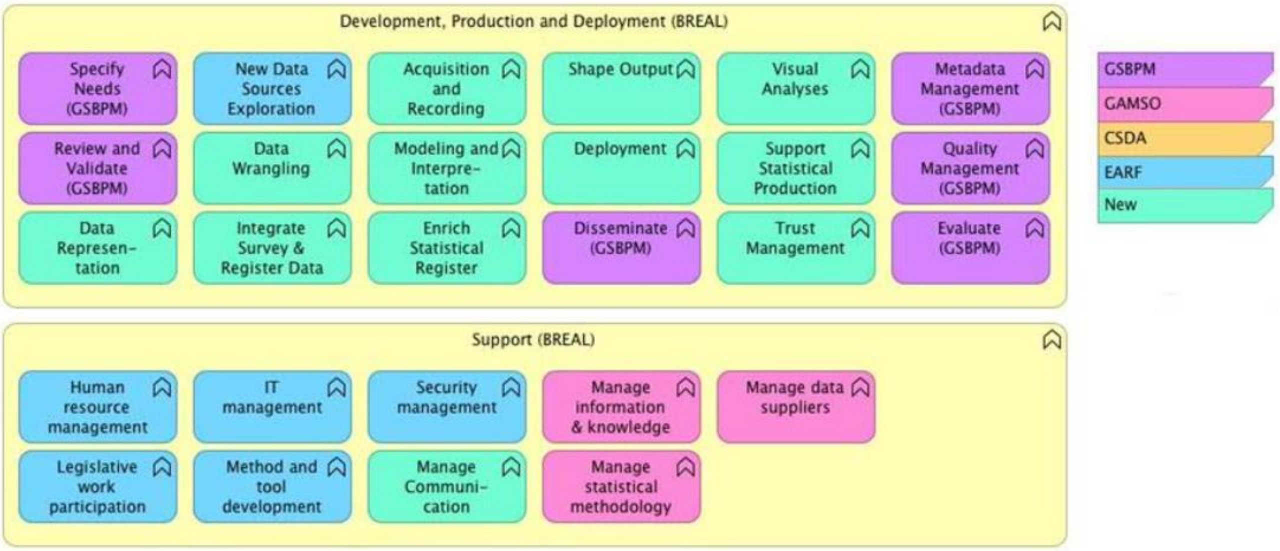
Tennekes, M., de Jonge, E. (2019). EBDA: Visual Exploratory Big Data Analysis for Tabular Data. CBS report, Statistics Netherlands.

Tuoto T., Fusco D., Di Consiglio L (2018). Exploring solutions for linking Big Data in Official Statistics, C. Perna *et al.* (eds.), Studies in Theoretical and Applied Statistics, Springer Proceedings in Mathematics & Statistics 227, pag. 49-58, ISBN 978-3-319-73905-2.

Wolter, K.M. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81, pp. 338-346.

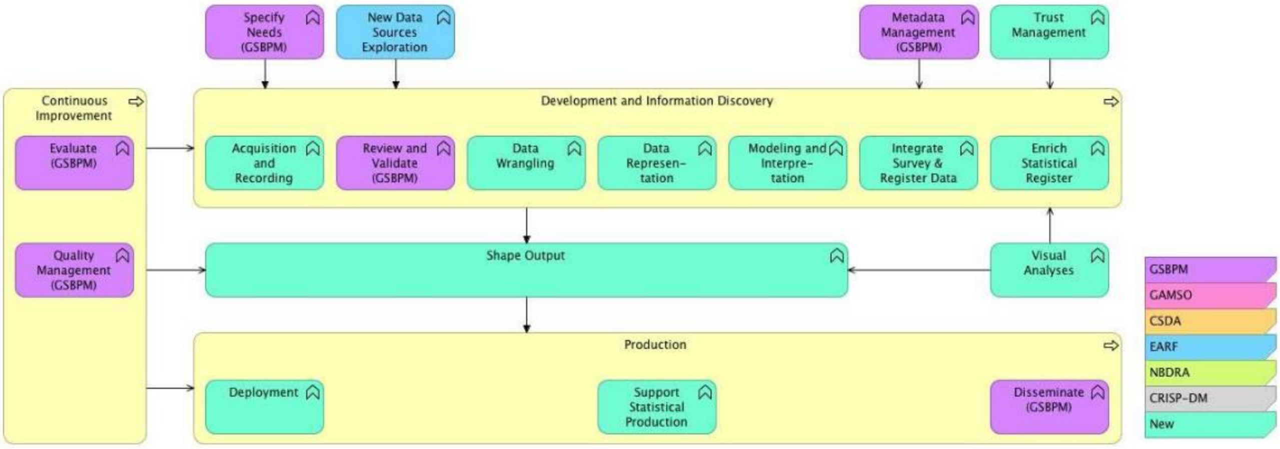
Annexes

Annex A: BREAL (Big data Reference Architecture and Layers)



Development, Production and Deployment (BREAL)
Specify Needs (GSBPM)
New Data Sources Exploration
Acquisition and Recording
Shape Output
Visual Analyses
Metadata Management (GSBPM)
Review and Validate (GSBPM)
Data Wrangling
Modeling and Interpre-tation
Deployment
Support Statistical Production
Quality Management (GSBPM)
Data Represen-tation
Integrate Survey & Register Data
Enrich Statistical Register
Disseminate (GSBPM)
Trust Management
Evaluate (GSBPM)
Support (BREAL)
Human resource management
IT management
Security management
Manage information & knowledge
Manage data suppliers
Legislative work participation
Method and tool development
Manage Communi-cation
Manage statistical methodology
GSBPM
GAMSO
CSDA
EARF
New

Figure 1: BREAL Business Functions



Specify Needs (GSBPM)
New Data Sources Exploration
Metadata Management (GSBPM)
Trust Management
Continuous Improvement
Evaluate (GSBPM)
Quality Management (GSBPM)
Development and Information Discovery
Acquisition and Recording
Review and Validate (GSBPM)
Data Wrangling
Data Represen-tation
Modeling and Interpre-tation
Integrate Survey & Register Data
Enrich Statistical Register
Shape Output
Visual Analyses
Production
Deployment
Support Statistical Production
Disseminate (GSBPM)
GSBPM
GAMSO
CSDA
EARF
NBDRA
CRISP-DM
New

Figure 2: BREAL Big Data Life Cycle

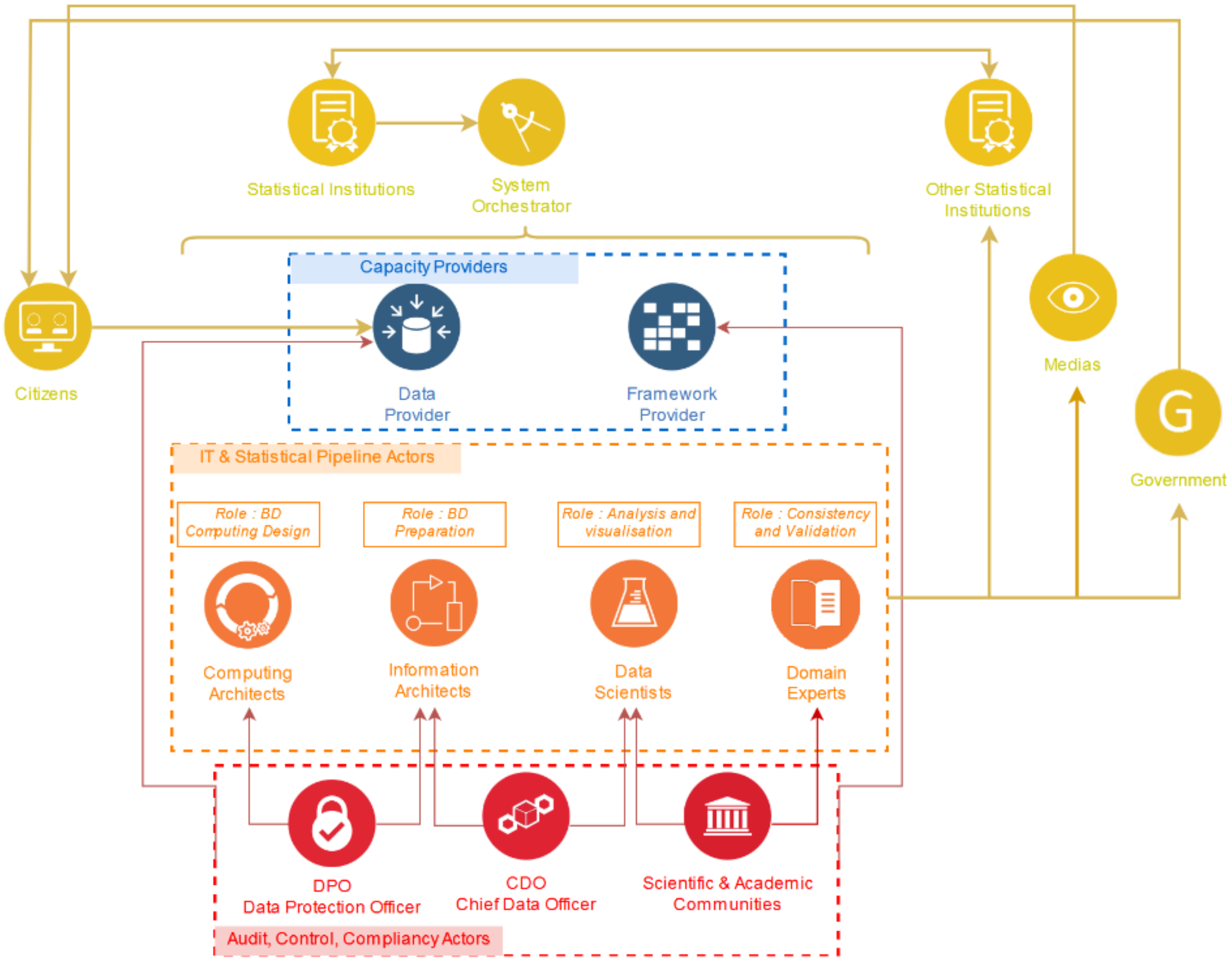


Figure 3: BREAL actors