Center for
Big Data Statistics

Working paper

Updating the Paradigm of Official Statistics:

New Quality Criteria for Integrating New Data and Methods in Official Statistics

Working paper no.: 02-20

In collaboration with

Sofie De Broe
Peter Struijs
Piet Daas
Arnout van Delden
Joep Burger
Jan van den Brakel
Olav ten Bosch
Kees Zeelenberg
Winfried Ypma

February 2020

Contents

Abstract

This paper aims to elicit a discussion of the existence of a paradigm shift in official statistics through the emergence of new (unstructured) data sources and methods that may not adhere to established and existing statistical practices and quality frameworks. The paper discusses strengths and weaknesses of several data sources, methodological, technical and "cultural" barriers (as in the culture that reigns in an area of expertise or approach) in dealing with new data and methods in data science and concludes with suggestions of updating the existing quality frameworks. Statistics Netherlands takes the position that there is no paradigm shift but that the existing production processes should be questioned and that existing quality frameworks should be updated in order for official statistics to benefit from the fusion of data, knowledge and skills among survey methodologists and data scientists.

1 The paradigm question for official statistics

1.1 Embracing new data sources and new methods

Before the emergence of what has been dubbed 'big data', national statistical institutes (NSIs) almost exclusively used two types of data sources: (1) data collected by means of statistical surveys and (2) data from registers held by administrations for purposes other than official statistics (OS). For statistics based on these data sources an elaborate body of validated statistical methods is available. This is not the case for what is now increasingly being explored by official statisticians everywhere, the potential use of: (3) new data sources.

Statistics Netherlands has been using big data sources for official statistics since 2007. This experience has included experimenting with new data sources such as online price data, scanner data, social media data and sensor data such as traffic loop data to making these new data ready for the production of official statistics, as in the case of the consumer price index. The exploitation of new data sources often involves combining them with survey and administrative data, in some cases giving these more traditional sources new applications.

However, the existing body of validated statistical methods cannot always readily be applied to new data sources, whether or not these are used in combination with survey and administrative data. New methods may have to be considered given the emerging opportunities to produce more timely and more detailed information, or even new types of information – and given the changing information demand and expectations from society (Struijs and Daas 2014). However, official statisticians conversant with the body of proven methods may not be entirely open to using new data sources, especially if this involves reassessing and rebalancing quality requirements. Whereas output-driven survey methodologists are mainly focused on accuracy, input-driven data scientists are predominantly focused on timeliness. The reluctance to change statistical practice is to some extent a matter of "culture'' (as in the culture that reigns in an area of expertise or approach).

After having explored the potential of new data sources within the existing organizational framework for some years it gradually became clear that a new approach was needed in order to be more agile in grasping the new opportunities and overcome "cultural" resistance, while at the same time concentrating on matters of substance. Seeing the potential of big data sources for

making more detailed, real-time or new statistics, and in particular the added value of combining different types of data sources, Statistics Netherlands has therefore set up the Center for Big Data Statistics (CBDS). The mission of the CBDS is to "explore and exploit new data sources, applying state-of-the-art methodology in collaboration with partners, in order to provide timely, comprehensive information on social phenomena relevant to users".

CBDS was an important innovation initiative and as a consequence, Statistics Netherlands now applies different approaches to making statistics: using unstructured data as input often combined with survey and administrative data whilst experimenting with new methods (such as natural language processing and machine learning). Having different cultures and expertise in one organization is desirable when confronted with new circumstances, but this poses challenges and might be experienced as threatening to one's own expertise and known statistical frameworks. Thus, NSIs are faced with the question how to optimally rebalance the quality dimensions and how to maximally exploit all available data sources and techniques from both survey methodology and data science. This is the question addressed in this paper.

1.2 Approaches to official statistics

Before introducing the idea of a paradigm update, it may be useful to mention some basic ideas of Thomas Kuhn (1922–1996). Kuhn acknowledged that what members of a certain scientific community have in common, are not only techniques but also shared values. This refers to a certain generally accepted truth that members of this scientific community take for granted. What about the paradigm of the official statistics community? If big data sources are ignored, the techniques and values are based on the use of survey and administrative data. The process of generating this data is broadly known as NSI are in control of the survey design (even if the distribution that is underlying the data is often unknown), the data is in limited supply and it requires quite some time for statistics to be produced. The body of validated statistical methods relies to a considerable extent, when it comes to survey data, on sampling theory, with survey sampling frames often derived from administrative registers. An elaborate quality framework exists to ensure that disseminated information complies with professional standards. This does not preclude the use of modeling techniques and associated assumptions, but models are primarily used to enhance intermediate statistical outcomes in a way that allows validation. Statistics are designed with specific questions or even output tables in mind.

Naturally, the culture sustaining such official statistics is very much quality oriented. Professional standards are considered essential. Accuracy must be ensured and proven before statistics can be released. There may be preliminary or experimental statistics, to be sure, but even these have to meet high standards and they are considered acceptable only because final official figures will follow in due course. This culture has served official statistics and society well.

If big data and other new data sources come into play, as is notably the case for the CBDS, new techniques are needed, and this has an impact on accepted statistical quality standards, frameworks and expertise and requires an update of the paradigm. The new needs have several causes. First, one has to cope with features of the data such as abundance, variety, timeliness and dynamism, lack of structure, messiness and uncertainty. Second, the data offers the possibility of a data driven approach where one generates information without having a specific question in mind. Third, some of the new techniques are rooted in artificial intelligence, such as machine learning systems that can detect complex patterns and are built to predict rather than explain. Fourth, features can be extracted and relationships can be learned automatically,

which is a radically different approach from the one traditionally used, wherein analysts derive features and model relationships explicitly based on their knowledge of techniques and the (small amounts of) data.

The use of new data sources also requires new approaches to addressing quality issues, such as understanding the quality of the data, understanding the data generating process, guaranteeing data delivery, and harmonization of definitions and concepts. Central quality issues are also understanding observed correlations and outcomes of machine learning algorithms. In turn, the key quality issue of a machine learning algorithm is its ability to predict new, unseen, cases. Another issue of those algorithms is that they should be kept up to date in order to keep measuring changing phenomena. And many more issues arise when combining data sources.

The new techniques, collectively called data science, have a bearing on the culture in an NSI. They even presuppose certain attitudes. A mindset is required in which it is acknowledged that the messiness of data – not knowing how it is generated, what population is covered or what is the actual meaning of the data – does not preclude its use in a process yielding information deemed useful for society. The conditions for its use may still have to be sorted out for many cases, but in this culture, validating new methods for new data sources is seen as a challenge rather than a problem. Still, since models are at the core of data science techniques, it is crucial to validate them and quality is seen as being a variable of the optimization equation of producing statistics, but not the only one.

Moreover, if only survey and administrative data is used, an NSI can in most cases work autonomously (provided the legal requirements to access these data are secured). Acquiring access to and exploration of new data sources may require more openness to collaboration with others. The collaboration with universities, knowledge institutes, other NSIs, local and national government and in some cases private partners allows joining forces in getting access to and combining data, sharing expertise and finding market value for new statistical products. Statistics Netherlands has been working with other institutes in the past, such as universities and NSIs across Europe; the CBDS has a strong focus on external contacts and seeks the interaction with the user on the experimental statistics. This is also driven by the need to have insight in what are the burning societal questions among citizens and local and national government, and to exploit new funding possibilities (demand driven approach). Although the new data sources are used to focus on the societal questions, the exact output that is finally generated is often more input-driven (the possibilities of the data sources are explored without predefining research questions and specific variables) than output-driven. . In a more traditional output-driven approach, first a set of variables and hypothesis are defined and then one decides how these variables can be observed. Here, data is specifically collected for this foreseen purpose. Nevertheless, by putting a demand for information first, the approach could also be seen as an output-oriented orientation.

In short, new data sources and new techniques require an update of the OS paradigm, which poses challenges to existing practices and culture in official statistics. This updated paradigm consists of two elements: use of an input-driven approach and the intensive use of models in the estimation process. There is a real or perceived tension between the exclusive use of validated methods using the known paradigm of inferring from survey and administrative data based on statistical laws and probabilities to yield solid results and the desire to exploit new opportunities. In fact, the new paradigm sometimes meets resistance. Objections to the new approach are linked to the use of unstructured data of which the data generating process is

unknown, pertaining to unclear populations; basing statistical outcomes on observed data correlations; the black-box nature of certain AI applications; the shift from a predominantly output-driven process to a more exploratory input-driven approach; possible disruptions of time-series if new sources are used; the volatility of new data sources; a fear to open the door to opportunistic practices, in particular related to new financial opportunities; and quality concerns in general. Mere resistance to change – and its consequences – plays also a role, of course.

1.3 Towards solving the paradigm question

What is the nature of the paradigm question resulting from the existence and emergence of various approaches to official statistics? Kuhn's (1962) notion of paradigm shift involved discarding old concepts and replacing them by new ones. For instance, in physical science, accepting that the speed of light was a universal constant led to discarding the concept of ether. This paper will argue that there is no need at all to discard any of the validated methods related to the use of survey and administrative data. In fact, it is in the data and methods fusion of both approaches where the enormous added value lies.

In order to make optimal use of the opportunities offered by new data sources these established methods need to be enriched and complemented in order to properly deal with features of data from new sources such as abundance, variety, timeliness and dynamism, lack of structure, messiness and uncertainty. Established methods are not easily applicable to a data driven approach where one generates information without having a specific question in mind. Contrary to established methods, such as design-based and model-based techniques, methods aimed at analyzing data from new data sources may be rooted in artificial intelligence.

This implies that the paradigm question is not one of choosing between an old and a new paradigm, but one of finding a good synthesis of approaches with proven value and new approaches. Finding such a synthesis is not easy, as new approaches and combining them with existing approaches need to be validated. And this is even more difficult as the existing quality framework for official statistics may have to be upgraded as well, in order to be able to properly assess possible outcomes of the integration of new data sources initially used for experimental statistics in the production of official statistics. To further complicate things, making use of new data sources has consequences for the technical way of data processing, from questions of hard- and software to questions of a more architectural nature such as the use of privacy preserving data sharing.

And this is not all. If big data and other new data sources come into play, some "cultural" barriers may have to be overcome if one wants to integrate already existing and new approaches. In fact, we had this experience before, when the use of administrative data sources became more prominent. At that time, the paradigm of the time needed to be upgraded as well, resulting in new methods for integrating survey and administrative data, an updated quality framework and new technical solutions such as, in some cases, processing events rather than repeatedly processing whole databases. Architectural principles were also updated, leading for instance to prioritizing the use of data from administrative sources over survey data, as long as quality requirements can be met. This was done given the opportunities that arose to reduce the survey response burden. We are in a similar situation now.

In this paper the paradigm question will be tackled in the following way. Rather than giving an overview of established and new approaches, we will start, in Section 2, with indicating the strengths and weaknesses of the use of various types of data sources, i.e., survey data, administrative data and new data sources. We will then identify the methodological, technical and cultural barriers to using new data sources and methods in Section 3. In addition to identifying the barriers, the question of how to get to solutions will be discussed as well. A new quality framework, on which the criteria for dealing with new data and methods will have to be based, is the subject of the subsequent Section 4. The concluding Section 5 will make up the balance and consider to what extent the paradigm question has been answered. However, the exploration of the growing number of (potential) new data sources is still in its early stages. As a consequence, there are many questions that cannot be answered yet, and new questions arise while others are answered. The last chapter will also propose a way forward, including research priorities.

2 Strengths and weaknesses of the use of different types of data sources

2.1 Strengths and weaknesses of the use of survey data

National statistical institutes traditionally use probability sampling in combination with design-based or model-assisted inference for the production of official statistics. The concept of random probability sampling has been developed mainly on the basis of the work of Bowley (1926), Neyman (1934) and Hansen and Hurwitz (1943). See for example Cochran (1977) or Särndal et al. (1992) for an extensive introduction in sampling theory. This is a widely accepted approach, since it is based on a sound mathematical theory that shows how under the right combination of a random sample design and estimator, valid statistical inference can be made about large finite populations based on relative small samples. In addition, the amount of uncertainty by relying on small samples can be quantified through the variance of the estimators.

Design-based and model-assisted inference means that the inference is based on the probability structure of the sample design that is used to draw a sample from the target population. They are popular by NSIs, since they do not rely on an explicitly assumed statistical model. For decades, there has indeed been the prevailing opinion that official statistics must be free from model assumptions, since model misspecification easily translates into wrong statements about the variables of interest.

A strong advantage of probability sampling in combination with a design-based or model-assisted inference is that it has a built-in robustness against model misspecification. This is useful in a production process where there is not much time for extensive model evaluation. This approach, however, is only useful for the production of statistical information at a relatively high aggregation level, since it requires large sample sizes to obtain sufficiently precise sample estimates. In the case of small sample sizes, however, the design variance of the sample estimates become unacceptably large and makes the built-in robustness against model misspecification

of less use. In such situations model-based inference procedures can be considered as an alternative. This means that the inference is based on a statistical model that describes how a random variable is related to one or more other random variables according to an assumed probability distribution. In the context of small area estimation, model-based inference procedures are applied to increase the effective sample size of a domain with sample information from other domains or preceding sampling occasions (Rao and Molina 2015). This kind of estimation procedures are gradually accepted by NSIs (Boonstra et al. 2008, Van den Brakel 2012).

The accuracy of statistics are measured with its variance and bias. The variance of survey sample statistics, depends on the sample size and will typically constitute a substantial part of the uncertainty of sample statistics.

The selection bias of sample survey statistics is approximately zero under complete response. In practice however, selection bias arises due to selective nonresponse, under coverage of the sample frame and to which extent with the field work strategy the target population is successfully reached.

The measurement bias in a sample statistics typically depends on the extent to which the conceptual variables to be measured, are correctly implemented in the questionnaire, but also on data collection mode and the quality of the interviewers. Problems with measurement bias in surveys arises, since measurements of the variables of interest are indirect in that respondents are asked to report about their behavior, introducing all kind of measurement errors. These problems do not arise with big data if they contain direct measurements of peoples' behavior, but this only holds for very specific examples. A problem with registers and big data sources is that an NSI has no control over the availability and stability of this data source. Major changes in the behavior of the public on social media and internet have a disturbing effect on the comparability of series over time. For example a Google-trend series on search related to vacancies might track an official series on unemployment. It does not measure unemployment, however. Search behavior before the start of the financial crisis in 2009 might be completely different compared to the period directly after the financial crisis, invalidating the concept intended to be measured.

A drawback of sample surveys is that data collection is costly and that its quality is compromised by non-response. In addition survey samples induce response burden, which is particularly an issue in business surveys. Strengths and weaknesses of survey data are summarized below.

Strengths	**Weaknesses**
Valid inference of a large target population based on relative small samples	Sampling error
Sampling error can be quantified via variance estimation	Large variance under small sample sizes (under design-based inference methods)
Established and widely accepted methods	Costly
Precision of the results are controlled by design of the survey	Sensitive for measurement bias
Low risk level since design-based inference is robust for model misspecification under sufficiently large sample size	Slow
The NSI is in control over data sources (stability, availability, and consistency)	Selective non response
	Response burden

2.1 Strengths and weaknesses of survey data

2.2 Strengths and weaknesses of the use of administrative data

An important characteristic of administrative data is that these data are secondary data and not collected for statistical purposes as with survey data (Hand 2018). As such, the method and process by which the data are collected is often controlled by the collecting organization and does not necessarily cover the purpose of the NSI. The aim is to cover the entire administrative target population as opposed to statistical data which is often based on a sample. When administrative data are used for statistical purposes, it often includes a harmonization step to come close to the population parameter of interest. Statistics Netherlands has by law, access to all administrative data that could serve statistical purposes.

The advantage of administrative data sources, as they are being collected by other organizations, are cheaper to collect with the added advantage that the data can be exhaustive if the data cover the statistical study population. They are deemed to be of good quality as they are being collected for a specific organizational purpose and could be real time if they would be available as soon as the data are recorded. Data sets tend to be large and detailed with a large coverage of a population so that statistical output is possible at a low geographical level. When administrative data are continuously updated the data also allows longitudinal studies based on time series of the data. However, changing definitions in the administrative data collection process will seriously hamper time series and comparability over time and detailed regional analysis can be affected by small errors in the data.

When the data contain unique identifiers, the data can be easily combined with other sources, such other administrative and survey data (see below). This way multi-variate distributions, relations between variables, can be obtained that have not been observed together (in single data sets).

A number of errors can occur in administrative data partially due to the fact that data analysts are not involved at the data collection stage. Frequent errors are partially or entirely missing records, the latter introducing selection bias in the data. Data entry errors, update errors, reporting errors (Groen 2012) and transmission errors are some of the errors introduced in administrative data. The larger challenge is to detect underlying mechanisms that affect the administrative data. These can remain unobserved if no quality check of the validity and non-representativity of the data or measures of uncertainty around observed changes in the data are in place or provided.

One of the big strengths of administrative data sources is the potential to combine the administrative data with other data such as survey data, other administrative data or even big data sources, which also represents challenges. Most of these other data sources will use a different classifications or use different aggregation of similar variables; data sets will use different periodicity and types of data. These other data sets will be a different measure of the same concept (as often in the case of big data) which makes comparison difficult and leads in some instances to contradictory results. Here lies a challenge to use derived variables or in some cases latent variables through the combination of data sources. There are many methods used when combining administrative data sources and administrative and survey data sources. These methods depend on the way the data are combined: multiple sources may be used to cover the population, the variables may complement each other, or they may overlap, there may be undercoverage, micro- and macro data may be combined or different sets of macro data. A large set of dimensions play a role when combining multiple sources, an overview is

given in De Waal et al. (2019). Although for combining administrative and/or survey data already a broad set of processes and methods has been developed, there is definitely scope for improving the processes and developing new methods in order to combine different administrative data and survey data with new types of sources. The above characteristics determine to a large extent the strengths and weaknesses of the administrative data sources, summarized in the table below.

Strengths	**Weaknesses**
Low data collection costs	Dependence on the data supplier
Potentially exhaustive coverage of the population	Under coverage of the study population
Deemed of good quality	Purpose for data collection does not fit statistical purpose
Potential for more current data	Rarely real time as time delay between data collection and delivery to NSI
Richness of the data in terms of variables and regional detail	Changing definitions in the data collection process
Potential for linkage with other sources	Selection bias due to missing records (units)
Limited selection bias (if only with early estimations)	Potential bias due to reporting patterns and errors
	Administrative units differing from statistical units
	Data entry errors
	Unobserved mechanisms affecting the data

2.2 Strengths and weaknesses of the administrative data sources

2.3 Strengths and weaknesses of the use of new data sources

Data are generated in large amounts in our modern world and often remain to be stored. These new data sources often contain direct measurements of the behavior of people and companies. Because of this, these sources provide a number of opportunities for official statistics.

Strengths of the new data sources

The major strength of new data sources can best be described by three words: speed, volume and new. The clearly biggest advantage is the speed at which new data becomes available. If such data contains information relevant for official statistics, this could enable the production of a very timely statistic and possible even the production of a real-time statistic. None of the other type of data sources currently used for official statistics is able to do this. Another advantage of these new sources is the availability of data in large quantities. This is good for two reasons. The first one is the fact that the availability of large amounts of data increases the chance of having considerable amounts of high quality data available. Even though a part of the data may still be of low - none or less useful - quality, the remaining part might very likely be suited for the purpose foreseen. Having large amounts of data available may make it more likely that a considerable part of the data is produced by units included in the target population. There even are a number of new (big) data sources that completely include the target populations. Examples of this are road sensor data for the whole country and Automatic Information System data of vessels. This, however, doesn't mean that all new data sources completely cover the target population and that their inclusion probabilities are known. Another advantage of new data sources is that they may contain signals not picked up by any other data source (more on this below). This enables the production of completely new statistics, which makes these data sources especially interesting for the development of early indicators of new phenomena.

Weaknesses of the new sources

When one looks at the downsides of new data sources, the following words best describes them: access, stability and units. Getting access to new data is not always easy. Many new data sources are generated by private companies, which makes getting and keeping access to these sources a challenge. Many NSIs have made legal arrangements to assure access to (governmental) administrative data without any costs involved. This enables long term stability of data access for these type of data sources. However, at the moment, access to new data sources is not included in any law by any NSI. The only exceptions to this are data sources maintained by semi-governmental organizations, such as road sensor data, as they can legally be considered administrative data. The lack of access and stability seriously hampers their use and also increases the chance of associated access costs . The latter is not a preferred situation for NSIs. Because of this, a number of NSIs are considering expanding the data access law in their country to particularly include new data sources, such as mobile phone (network) data and bank transactions data. Stability of the data is another major concern. Because new data sources can be considered a more or less- byproduct of new technology, any progress in this area may affect the kind of data collected and the way it is stored; in other words, they affect the variables included. It requires flexibility of the NSI to deal with these changes. Another important downside of new data sources are the lack of metadata on the units included in the source. For example, background characteristics of the units included, such as for instance on the users of mobile phones, are often lacking or stored in another privacy protected system. The same holds for the inclusion probabilities of those units (see above). However, in a number of cases, new data can be used in such a way that it becomes less dependent on these downsides. The social tension indicator (CBS 2018) is an example of this.

Important considerations

However, apart from the strengths and weaknesses listed above, there is a combination of important considerations that need to be discussed as well. New data sources can only be useful for official statistics if, and only if, they conform to the following three conditions:

1. The data contains, in some way or another, traces of information (i.e. a signal) on a topic relevant for official statistics;
2. The data can be processed in such a way that this information can be extracted in a stable, reliable and reproducible way;
3. The relation between the information extracted and the topic for which it is used is the result of a well investigated and understood relationship.

Each of these conditions is essential because only the combination of all of them enables the production of a reliable and reproducible statistical product based on a new data source. However, when all conditions are met, it is not trivial to prove that the findings are valid. The most straightforward way to check them is to triangulate the findings with those observed in other types of data sources; i.e. via external validation. In this way an attempt is made to measure the same or a (very) similar concept in another data set. Such external checks are usually done by comparing the findings that are based on the new data with those provided by survey or administrative data. Any confirmation will obviously increase the trustworthiness of the new data based findings. However, any dissimilarities observed may not mean that the findings based on the new data are worthless. They may very well point to differences in i) the way by which the concept is measured in the data sources compared, ii) the population composition included in each source, or may even hint to issues by which iii) the concept is measured in the more traditional data source. As such, the results based on the new data source may very well provide new insights.

Reproducibility and the stability of the phenomenon observed are other essential properties. The above statements have made clear that when the detection of the phenomenon is not stable, the new data source cannot be used to produce reliable results on the long run. The most famous example of such a case is the use of Google search terms to predict occurrences of the flu in the US (Ginsberg et al. 2009). Only after this relation was tracked over a number of years, it became clear that the Google team had developed something that more resembled a seasonal detection engine than a real-time flu predictor (Lazer et al. 2014). This observation indicates the biggest weakness when using new data sources: it may start with an interesting finding, a correlation in this case, but it cannot be based on this alone. For instance, any correlation found needs to be checked over a longer period and –even better– regarding the underlying cause or logic of the phenomenon observed. An example in which this has been done is the relation between Consumer Confidence in the Netherlands and the sentiment in Dutch public social media messages (Van den Brakel et al. 2017). Here, social media sentiment improved the precision of the survey-based estimates in a structural time series model-based approach. By applying such an approach it can be assured that both series are related in a way that goes beyond correlation alone. It is, however, not easy to get a grip on some of the correlations observed, such as the one between traffic intensity and GDP (Daas et al. 2016; Loumaranta et al. 2018). One should certainly prevent pure correlation based mining when studying new data sources (Hero 2013; Benedikt 2019).

Most challenging is the situation where a direct comparison of new data based findings with observations in other data sources is not possible. Here new data sources are picking up a signal or signals not observed in any other data sources. This opens up possibilities to produce statistics on completely new phenomena. However, when there is no resembling signal in any other data source, the findings cannot be triangulated with traditional data sources. The best way to check the validity of such findings is by first checking the internal validity of the approach used by going through each step meticulously. Next, the stability of the finding over time needs to be evaluated. The need for long time series in the latter case makes it challenging to rapidly convince any critical observer, such as many official statisticians, that the phenomenon measured is real, interesting and worth pursuing. Here logical and causal arguments can be used to confirm or disprove the case. Examples of new data sources in which phenomena that were previously not observed, can be observed are mobile phone network data. It is challenging to validate these findings and we refer to Daas et al. (2019) for an example of such an approach. The strength and weaknesses of new data sources are summarized below.

Strengths	**Weaknesses**
Real time	Often unstructured
Large amounts of data	Access to the (private company held) data
Direct measurements of behavior	Poor quality or lack of meta data
Detailed and high frequent measurements	Unknown data generation mechanism
Detecting new (previously unobserved) phenomena	Unknown inclusion probabilities (when a part of the population is included)
Some fully cover the target population	Instability of the source
	Lack of auxiliary variables

2.3 Strength and weaknesses of new data sources

3 Challenges in using new data sources and new methods

3.1 Methodological challenges

Using big data in the production of official statistics

The non-probability nature of some big data sources requires dedicated methods of inference to produce statistics about a larger intended, finite target population. Broadly spoken, there are two ways to use non-probability data sources in the production of official statistics. The first approach is to use them as covariates in model-based prediction methods for survey data. The second approach is to use them directly as a data source for official statistics and correct for possible selection bias. A third approach, still to be explored further is how to produce official statistics through data fusion and using new techniques such as machine learning and text mining. The third section offers a perspective.

Big data as auxiliary information

One potential application where big data can be used as covariates in prediction models for sample survey data are small area estimation models. A lot of the big data sources are fuzzy and volatile and the records typically do not coincide with the units of an intended target population or the sampling units of a probability sample. Complications with linking units in big data sources with sampling units in a probability can be avoided, at least partially, by using area level models and time series models where the direct estimates are the input of the model.

Several authors proposed methods to combine survey data with non-probability data sources available from, e.g. sensor data and mobile phone data with the purpose to make detailed regional predictions for wellbeing and poverty. Many applications apply machine learning algorithms to establish the relation between survey data and sensor or mobile phone data and use the latter data set in a second step to make detailed regional predictions. Noor et al. (2008) analyzed the correlation between night-time light intensity from satellite images and survey sample data on household income in Africa. Engstrom et al. (2017) used day time satellite images to predict well-being, using deep learning. Blumenstock et al. (2015) applied machine learning methods to combine mobile phone data with survey data on poverty and used this to predict poverty and well-being on small regional level in Rwanda.

Examples where big data are used as auxiliary information in cross-sectional small area prediction models are Marchetti et al. (2015) who used mobility patterns of cars tracked with GPS as a covariate in a Fay Herriot model for predicting poverty for small regions in Italy. Schmid et al. (2017) use mobile phone data as a covariate in a Fay Herriot model to predict literacy in Senegal.

Multivariate structural time series (STS) models are an alternative to borrow strength over both time and space. Auxiliary series derived from big data sources can be used in these models to combine time series obtained with repeated sample surveys with auxiliary series derived from registers or big data sources. This serves two purposes. Extending the time series model with an auxiliary series allows modeling the correlation between the unobserved components of

the structural time series models, e.g. trend and seasonal components. If the model detects a strong correlation, then the accuracy of domain predictions will be further increased. Harvey and Chung (2000) propose a time series model for the Labor Force Survey in the UK extended with a series of claimant counts.

Information derived from non-traditional data sources like Google trends or social media platforms are generally available at a higher frequency than series obtained with repeated surveys. This allows to use this time series modeling approach to make predictions for the survey outcomes in real time at the moment that the outcomes for the big data series are available, but the survey data not yet. In this case the auxiliary series are used as a form of nowcasting. Van den Brakel et al. (2017) applied a bivariate STS model to estimate the Consumer Confidence Index, based on a monthly cross-sectional sample, in real time using an auxiliary series derived from messages left on social media platforms. Schiavoni et al. (2019) proposed a dynamic factor model to estimate monthly unemployment figures in real time with claimant count series and Google trend series.

Big data as the primary data source

If non-probability data sources are considered as a primary data source for compiling official statistics, then the question raises to which extent results obtained with a non-probability data source can be generalized to an intended, larger target population. Different methods are proposed in the literature to account for selection bias in non-probability samples. Some authors apply standard weighting and calibration methods known from classical probability sampling to non-probability samples, which is referred to as pseudo-design-based inference methods (Baker et al. 2013). Another class of methods to correct for selection bias is to apply a statistical model to predict the units not in the sample (Royall 1970, Valliant et al. 2000).

Some methods combine a non-probability sample that contains the target variable of interest and auxiliary variables with a reference sample that is based on a probability sample and only contains auxiliary variables. One approach, quasi randomization, is to construct propensity models to estimate selection probabilities for the non-probability sample (Isaksson and Forsman 2003, Valliant et al. 2013). Sample matching is also applied as an attempt to reduce selection bias in opt-in Web panels using covariates obtained in a small reference sample to construct propensity weights without collecting observations for the target variables (Vavreck and Rivers 2008, Rivers and Bailey 2009). These ideas are related to approaches that are also used in microsimulation to match probability samples with population or census data (Tanton and Edwards 2013). These correction methods are all based on strong ignorability assumptions and can lead to serious bias if these assumptions are not met. Indeed, Buelens et al. (2018), who compared pseudo-design-based, model-based and algorithmic methods, conclude that auxiliary information typically available for weighting and calibration are demographic variables like age class, gender, regional classifications, do not sufficiently explain the data generating process of a non-probability sample to correct successfully for selection bias.

Another issue is that these correction methods assume that the records in a big data source contains besides the target variable a set of auxiliary variables which correspond with the units in a target population or a reference sample. Unfortunately, these conditions are seldom met. Most big data sets are fuzzy, records do not correspond with units in the target population or a reference sample and auxiliary information is generally not available since owners of the big data source are reluctant to provide them due to privacy issues. It could therefore be the case that big data will be of most use for official statistics in combination with survey and/or census data (Meng 2018).

New big data methods in official statistics

Big datasets provide measurements of phenomena at a level of detail far exceeding that of the more traditionally used sources. The variability of the phenomena at this level of detail is, however, also often found to be large (Puts et al. 2019). Sampling of the 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). New approaches are needed to make optimal use of new data sources. The methods required for this are, certainly within the official statistical realm, new. Apart from the technical barriers discussed in the next section, the essential challenges for which new methods are needed are:

1. Dealing with noisy, dirty and unstructured data
2. Dealing with selectivity
3. Correlations and beyond

Each of these challenges is briefly discussed below.

Noise

As is the case for many secondary data sources and certainly for sources containing huge amounts of data, not all records are relevant for the purpose the researcher has in mind. The irrelevant records, the 'noise', may even negatively affect the relevant ones. We found that the signal-to-noise ratio of big data sources is often rather low. In general the data can be considered as mostly noise; only a mere fraction of the data is of interest–the signal (Silver 2012). Finding methods to reduce the noise in big data, thereby increasing the signal-to-noise ratio, is vital for obtaining a successful result. Aggregating or applying a filter, such as a Kalman or a filter based on Poisson distributed noise, has been successfully applied, as are constructing queries that prefer the inclusion of relevant records. Another approach routinely used is removing all records with data that are clearly corrupt (dirty records) or contain only very small amounts of data. If one analyzes 100 million records, removing 100 erroneous records is easier than considering the relevance of each 'outlier', their effect on the estimate, and attempting to correct them. One needs to be pragmatic when analyzing big data. Applying dimension reduction methods, such as principal component analysis, factor analysis or self-organizing maps, are other ways to decrease the size of the data without losing much information. For unstructured data, such as texts or pictures on web pages, the first step of analysis differs. These types of data are usually transformed into a form more appropriate for statistical analysis; such as word frequencies and their distribution or the clustering of pictures into similar groups. Methods for performing this initial step can be found in areas of science more accustomed to dealing with this kind of data such as machine learning. We can also learn a lot from Google or other internet giants here. Information theory (Shannon 1948) can provide more insight into the nature of big data. Besides noise, as mentioned above, redundancy also contributes significantly in the total volume of big data. In all cases, information is extracted from big data, usually reducing its size and bringing it more in line with the size of the files statisticians are accustomed to dealing with. In subsequent steps many familiar statistical methods can be used.

Selectivity

Despite the huge amounts of data present in big data sources they may still not cover the complete target population considered. Big data may therefore be selective. This is an important issue when using big data for official statistics. There are several important points to consider.

The first one is the realization that selectivity may vary tremendously per big data source, per target variable and per question asked in the survey. Although for some sources selectivity of the units of the target population is definitely an issue , for others coverage is almost or essentially complete, due to the nature of the process through which the data come about. Even when coverage is partial, this may still result in a considerable amount of data for that particular group. Methods that could be used resemble post-stratification. This suggests that the complete absence of a particular part of the population is an important issue to determine accurately. Here, both survey and administrative data can assist.

The second point is related to the first one. As there is no sample design for big data, ways to correct for missing data are needed and will very likely require a model-based approach. Model-based methods require estimating the model parameters. This pivotal task is challenging for data sources with hardly any auxiliary variables and when the data generating mechanism is unknown. Approaches inspired by the area of data-mining and machine-learning may provide solutions and NSIs are in a good position to make these methods as open, transparent and honest as possible.

The third point is–when applicable–an issue that needs to be addressed first. A lot of big data sources actually register events or –more correctly–aggregates of events. This is the reason why many of these sources are big. Examples of (aggregates of) events are: the content and time when a social media message is written, the start and end times and location(s) of a call made by a particular mobile phone and the number of vehicles passing a road sensor at a particular location at a particular point in time. In fact, a considerable part of the big data sources studied at Statistics Netherlands was found to be event-based and hardly contained information on the statistical units of interest. Since the events stored are (indirectly) caused by the statistical units, e.g. people or businesses, dealing with them in a traditional way requires these events to be converted to the corresponding units first. This may not be easy, as a limited amount of identifying information is available. Perhaps the additional use of other (big data) sources may assist here. Considering the above, this suggests a two-step approach in which the first step consists of deriving profiles from events to identify units or subpopulations (groups of units). The subsequent step uses inference methods based on the information provided by these. In this context, it could be that the part of the population included in big datasets might actual be representative for the population for one particular variable studied. Or a big data source might still be used to produce (biased) estimates providing that they strongly correlate with existing statistics, for example to improve accuracy and speed of the existing statistics or merely to reduce the sample size.

Correlation and beyond

Because of the huge amounts of data available, comparing it to data from another (survey) source may result in a correlation between a particular (derived) big data variable and a variable in another more traditional source. This does not imply a direct causal relation. In fact having access to more data increases the chance that correlations are found. It is the difficult task of the researcher to fully investigate this 'relation' and try to distinguish a true from a false correlation which may also be referred to as coincidental or spurious. Methods that can be

used here are time series analysis, cointegration and various forms of causal analysis (Pearl and Mackenzie 2019). Be aware that this can be challenging as it requires analyzing lots and lots of data and may even require combining several (Big) data sources.

Another important issue that needs to be considered here is the fact that the phenomenon observed in the big data source might fade over time. This is known as "concept drift" in machine learning and points to a gradual loss of signal in the data source studied (Lu et al., 2018). It is suggested that this is typical for (prediction) models with a high number parameter/variable included. Such models are often used in machine learning. Solutions suggested to deal with such drift are re-training the model on new data after a specific period (such as every 3 months) or by continuously adding new records to the training and test set. We think a more rigorous approach might help here which requires a careful study of the importance of each of the variables included.

3.2 Technical challenges

The technical barriers to overcome to add new data sources to official statistics can be numerous. We describe some of them: connectivity, data handling, tooling / modeling and dissemination.

Connectivity

Where survey data tend to be small and are processed in waves and administrative data are structured and relatively well organized by nature, new (big) data or web data on the contrary are in many cases less structured, more volatile and updated continuously. Also the data may be too big to copy, one may need to connect to streaming data directly, it may be necessary to develop and maintain autonomously operating web scrapers (Bosch 2018) or one may choose to perform statistical pre-calculations at data providers' computer centers, also known as pushing computation out (Ricciato et al. 2018). In all these cases it is essential to have good connectivity, not only in terms of a big pipe to the internet but also in terms of powerful servers close to the internet with the right firewall configurations and people with the right skills to manage all this. And it is not only the data access that demands for improved connectivity. Modern data science tools expect seamless web integration more and more to provide for updates and access to package repos (see next section). In addition the extended collaboration with national and international data- and knowledge- partners requires a more open but still data-secure infrastructure as well. To support all this it is essential to change the connectivity paradigm from *collecting* to *connecting*. Instead of collecting data and processing it in an isolated environment a statistical office needs to be heavily connected to new data sources, advanced software repositories and knowledge partners during all phases of the statistical process. Updating this paradigm is a technical barrier.

Data processing and analysis

Extracting relevant and reliable information from big data sources and incorporating it into the statistical production process is challenging. When one wants to analyze big data, one needs to have a computing environment that enables the rapid checking and processing of large amounts of data. Such an IT environment is needed because the information content of many big data sources is low, meaning that not all data included is relevant for the research question under study. The low information content is the result of a combination of reasons,

the most important ones are: the fact that the data is not generated for the particular purpose the user has in mind (it is secondary data); the data is often unstructured and noisy (more on this below); only a limited number of variables are available; the phenomenon in which the researcher is interested may be rare. Hence, to make successful use of big data, large amounts of data need to be processed. Analyzing big data is routinely performed in our office with R or Python. The most important skill here is knowing how to write a program that is able to access all the data in a big dataset within a reasonable amount of time. Having a secure computer environment with many fast processors, large amounts of RAM and fast disk access certainly helps. The two IT environments we use in our office are: i) a secure private SPARK cluster to process data in parallel and ii) general-purpose computing on graphics processing units (GPGPU). The first is mainly used to pre-processes huge amounts of data resulting in much smaller datasets with a much higher information content. These are subsequently studied in the traditional way. The second environment is mainly used for deep-learning purposes; i.e. machine learning based on artificial neural networks. We also see an increase in the use of NoSQL databases, parallel processing and databases specific for network analysis. The outcomes of any of the methods used in the above mentioned IT-environments should not be negatively affected by the (distributed) way in which these methods are run and should be reproducible.

Tooling / modeling

Part of the process of adopting new data sources is to be open to model-based statistics. Instead of designing and implementing algorithms developed on a study of the data sources at hand, the data itself feeds and trains the production machine. This asks for a different tool landscape. Where traditional systems are implemented by IT specialists using tools targeted at stability and minimizing change, new systems need to be developed using modern data science tools with support for all phases of machine learning including analyzing and retraining models. As an example the new tool landscape would comprise tools such as R, Python, Nodejs, Scala, supplemented with ML tooling such as scikit-learn, PyTorch etc. and interactive environments such as RStudio and Jupyter notebooks etc. The advent of data science tooling brings in a modern data science culture such as the straightforward notion to "not re-implement anything that is already out there". Thus, data scientists require immediate access to fast developing packages in the package management systems of their choice such as CRAN, Anaconda/pip and NPM. To be effective they may need direct access to pre-trained models and text mining corpora developed at universities or major research companies around the globe. Part of the job is to make sure to use the latest models that accurately describe the ever changing world. It is clear that setting up such a professional data science environment supporting research, product development and production is a technical challenge.

Dissemination

Another area where the advent of new data sources and the corresponding data science way of working also breaks some barriers is in statistical dissemination. The production of statistical output usually involves multiple steps, such as putting data into a statistical dissemination database, documenting the methodology and other background information in a pdf, writing a web article in a content management system and developing some interactive visualizations using specialized visual design tools. With the advent of modern data science tools a new - simpler - workflow can be adopted. Using markdown-based tools and visualization frameworks such as ggplot (both in R and Python) it becomes possible to generate professionally looking content from the data itself. Multiple examples of the adoption of this process in statistical organizations were presented recently (Lesur 2019 and Grutter 2019). Also, the rise of R-Shiny as a tool

for generating high quality dashboards where complex phenomena can easily be explained to end-user gains new possibilities for shortening and improving statistical dissemination. Adopting these new tools and workflows to go from *designing* output to *generating* output can be a technical barrier, but also an opportunity.

3.3 Cultural challenges

What "cultural'' barriers have to be overcome in order to arrive at an optimal synthesis of already existing and considered new approaches? The purpose of overcoming "cultural" barriers is not to indiscriminately embrace all new approaches considered, without proper assessment and validation. Rather, the aim is to do the assessment with an open mind, in a rational way with the aim to make statistics real time, more detailed and efficient, eager to make the best use of the opportunities that arise from the emergence of new data sources. These developments often go hand in hand with a more external orientation of the organization. This benefits the communication with the users of the statistics compiled as well as those who have to pay for them or make sacrifices, including respondents.

What "cultural" features or elements would hamper an open and enthusiastic investigation of the potential of new data sources, possibly using new approaches? Some papers offer clues. For instance, Hsieh and Murphy (2017), discussing the possible use of Twitter data, state: "As survey researchers, we [...] are not content with using a data source not well understood at the level of the individual." Another examples is given by Baker (2017): "The refined process of designing, conducing, and analyzing surveys offers a level of control (and perhaps comfort) that we lose with big data." And Karr (2017) writes: "Statisticians see themselves facilitating a path from data to information to knowledge, while data scientists see the path as data to insight to action."

The standard for NSIs to produce official statistics is probability sampling in combination with inference methods. This approach is considered low risk (as long as response rates remain high) because with sample surveys, an NSI has full control over the availability of the data, as well as the quality and frequency of the statistical output. Improving the level of detail, frequency and timeliness of statistical information, without increasing sample sizes and thus data collection costs, requires model based inference procedures.

Replacing sample surveys for registers or non-probability data sources, is often perceived as a substantially increased risk due to the loss of control of data collection and its quality. In addition model-based procedures based on strong ignorability assumptions are required to correct for selection bias. The low-risk appetite of NSIs hampers the use of new data sources and inference methods (Van den Brakel 2019). However, whilst focusing on the risks of unstructured data sources, survey methodologists tend to perhaps underestimate the risks of falling response rates and downplay the inherent weaknesses of surveys using questionnaires as a measurement tool (measurement error and selection bias).

The wish to understand the data used and to have some control over inference to the population and the process seems quite reasonable, and there is nothing wrong with how statisticians see themselves. However, at an emotional level, resistance may occur if some promising new approach implies that what is taken as self-evident requirements is challenged or seen as a risk. There may actually be quite a few of such barriers, some having to do with laudable professional concerns, other more related to matters of personal expertise, interest and the consequences, such as:

- The opinion that a data source should not be used if the units (population) and the data on the units (variables) are not well understood.
- The notion that official statistics should not be based on observed correlations if the causality is not understood.
- The view that the desired output (user needs) should be leading in designing statistical processes and that data-driven approaches pose a risk to objectivity, impartiality and scientific principles.
- The feeling that with new approaches one loses control over quality, for instance if part of the process is a black box (e.g. based on machine learning).
- The fear of disrupting valuable time series if changes in the way of compiling statistics are allowed or possibly volatile data sources are used.
- The perceived risk of the introduction of non-professional criteria for the production of statistics. This could arise from opportunities to generate income from exploiting new data sources, or from the desire to fulfil user needs for quick results at the cost of quality.
- Lack of knowledge of the new tools and techniques of the future, the need to learn them and the associated apprehension.
- The threat to the comfort zone to which one has become used, for instance if methodologists with a more theoretical or problem-solving orientation are asked to develop a more entrepreneurial attitude with a more external orientation.
- Some human beings have an innate resistance to change in general.

All these concerns are important to take into account when assessing new approaches and contribute to a qualitatively good assessment. However, they may become obstacles to the development of new approaches and optimizing their integration with already existing approaches reflecting cultures. Being aware of the possible cultural barriers listed above and making them discussable, helps the rational, merit-based assessment of new opportunities. Changes should not be implemented recklessly, but undue delay in implementation of improvements also has its costs.

4 Towards a common quality framework

In order to implement experimental statistics based on new data sources and methods into the official statistical process, one needs to go through a validation process where quality requirements are relevant (De Broe et al. 2019). There are issues of the quality of the unstructured data, interpretability of algorithms, the data generating process of machine generated data, issues of different concepts as measured in different data sources, just to mention some. In the following section we elaborate on existing quality frameworks (Braaksma et al. 2019) in an attempt to address these new quality requirements.

In a general sense, quality is *fitness for purpose*, i.e. whether a product satisfies the requirements of its user. In statistics, quality is sometimes equated to accuracy, which is then often measured by the mean square error. However, this is too limited, and we must first identify the intended use. When it comes to big data or other new data sources in official statistics, there are various purposes and uses that may be distinguished. First we may look at new data for research purposes, for example to see whether a big data source might be at all useful or whether a particular statistical method might be interesting enough for further investigation; we call this called *research quality*. Secondly we may be interested in how good certain methods are, i.e. we are interested in *methodological quality*. Thirdly, we may be interested in the potential use of new data sources as input in the statistical process, for example to see whether they may be linked to other data sources or whether their concepts and measurement standards are identifiable; this may be called *data* or *input quality*. Fourthly, we may be interested in the actual use of new data sources in statistical processes; this may be called *process quality*. And finally, we may be interested in statistical products (statistical outputs) that are wholly or partly based on new data sources; this may be called *output* or *product quality*.

Of course it may be argued, in line with the ISO view of quality, that output quality is what matters for an NSI, and that the other views of quality are irrelevant for an organization as a whole. But there are various reasons why this is not so strict. First, output quality has several dimensions (see below) and it is not always possible or easy to measure these dimensions; and so we must turn to indicators of methodological, process and data quality to get more or less implicitly a picture of output quality. Secondly, methodological and process quality are part of most quality frameworks in official statistics as the dimensions of sound methodology and appropriate statistical procedures. Thirdly, indicators for methodological quality, process quality and partly also data quality can be used as an early warning system to trigger adjustments to ensure a sufficient output quality. Therefore, methodological quality, process quality and data quality each have a value on their own. Finally, NSIs are publicly financed and often have restricted budgets. They have to be cost effective and cannot afford to spend large amounts of public money on research and development projects with highly uncertain outcomes. Therefore they have to follow a staged approach, with at the end of each stage a thorough decision whether to proceed to the next stage; and it is here that the various other quality views have a role.

In the next subsections we will look at these quality views in turn, always with respect to new data sources. Since the topic of this paper is on methodological paradigms, most attention will be given to methodological quality. We will also touch upon the other aspects of quality and there we will focus on typical issues that occur with new data sources. In the remainder of

this section we will first discuss research quality, followed by methodological quality. Then we will discuss a number of typical error types in the area of input and process quality. Finally, we discuss output quality and whether there is a need to add new quality dimensions to the established quality dimensions.

Throughout the subsections below, we will assess these quality views by using the quality principles from the quality framework of the European Statistical System (ESS) (Eurostat 2014b):

- (output) *quality principles*:
  - relevance
  - accuracy and reliability
  - timeliness and punctuality
  - coherence and comparability
  - accessibility and clarity
- *statistical (process) principles*:
  - sound methodology
  - appropriate statistical procedures
  - non-excessive burden on respondents
  - cost effectiveness
- *institutional principles*:[1)]
  - mandate for data collection and access to data
  - statistical confidentiality and data protection
  - impartiality and objectivity.

Of course in practice, these various quality dimensions have to be weighed against each other. Also, note that if one cannot find out whether a data source fulfills a requirement, this poses a serious barrier to using that data source in official statistics; we will see below that some of the quality requirements may indeed act as a binding restriction.

4.1 Research quality

When one obtains access to a new data source, one will first explore whether the new source is suitable at all to make statistical outputs. This requires access to the data as well as its metadata, such as a description of variables and of data editing; it is also important to assess the response burden and the costs. As we have seen in the previous sections, we may then explore the data to see whether the source is suitable for making official statistics or not.

Access to the data

The principle of appropriate statistical procedures requires NSIs to consult and cooperate with providers of administrative and new data about the concepts and contents of these data. Similar to administrative data, many of the new data sources have a data owner. A new data source can only be used by statistical institute after the data holder has given access to those data. Preferably there is a law in which free access to those new data sources is arranged. In the

[1)] We leave out three more general institutional principles (professional independence, adequacy of resources and commitment to quality) since they do not apply here directly.

absence of such a law, the statistical office may be asked to pay a certain contribution for access to the data source. One then has to weigh whether this contribution is worth its money, relative to the output that can be created by using this source.

When a statistical office is granted access to the data, it is important to arrange that the data holder notifies the office in advance of any changes which are going to occur (Daas et al., 2009). With new data sources however, this is not so easy to accomplish. Therefore after a source is accepted as input for official statistics, it is important to monitor whether there are changes in the input data or not. This is for instance done for the scraping of prices from fixed websites.

Non-excessive burden on respondents

Administrative and many new data sources do not directly burden households or enterprises, and so have a clear advantage over data collected directly from respondents. However, secondary use of data may be time consuming for the data provider, namely when the statistical office has many questions about the data, or when special processing steps are needed for the supply of the data to the statistical office.

Cost effectiveness

Administrative and new data sources save NSIs the cost of primary data collection, and so have a clear advantage over survey data. In some cases however, processing may be more costly, in particular when data are more volatile and consistent statistical time series are required.

Metadata

In case of administrative data and in case of structured, high-volume big data, it is important to obtain documented metadata of the population, the units, and the variables and of the time dimension; see Daas et al. (2009). That documentation is useful to determine whether the data can be used (as one of the sources) to make relevant output for official statistics. In case of unstructured data, for instance Twitter messages and enterprise website texts, there is no documented metadata, since the sources do not involve a well and predefined set of units, variables and time frequencies. In order to use those data, one can try to 'relate' the unstructured data to more structured data that fulfill the definitions (metadata) of the intended statistical output. For instance, concerning the population to which the data refer, one may try to link the unstructured data to a population frame. This latter approach can be followed for business websites, but there are also unstructured sources that lack any form of identifying information. Concerning the variables, one might try to derive a well-defined variable from the source (text, images) by using a supervised learning approach with carefully labelled data.

4.2 Methodological quality

The principle of *Sound methodology* requires "that scientific criteria are used for the selection of sources, methods and procedures." As described in the previous sections, statistical-learning techniques are often used in the analysis of new data sources. For methodological quality, the most important aspects are: interpretability, robustness, stability, impartiality and objectivity, and accuracy and validity.

Interpretability

The core purpose of statistical learning models is their generalizability: prediction of new, unseen, cases. However, it may not always be clear how the model comes to its decision. Some models are interpretable models 'by nature'. Examples are decision trees, decision rules and linear regression models. For models which are not interpretable 'by nature', techniques have been developed, and are under development, that aim to give insight in how the model takes its decision. A general approach is to synthetically make a minimal change to the input of a record such that the decision of the model is changed (Krause et al. 2017). An example of an algorithm to do so is LIME, which builds a local interpretable model (Ribeiro et al. 2016). The interpretability of the model is an important quality characteristic for two reasons. Firstly, it is important for the trust of the general public in the model. Secondly, interpretability of models is also important in official statistics, when the predictions of models for output purposes are fine-tuned. When checking the quality of a model one may find that the model quality is insufficient for certain parts of the target population. In that case it can be very useful to check how the model came to its decision for those subpopulations. That can be a starting point for making improvements to the model.

Robustness

In official statistics we would like statistical learning models to take consistent decisions that are not affected by small (measurement) errors in the data source. The reason is that one does not want a model to lead to a break in the trend when there are (limited) changes in the source data. This property is also referred to as 'robustness to noise' of a model. The typical approach to achieve more robust models is to use data augmentation techniques. Augmentation means that new, synthetic, training examples are created that have the same label as the original examples but the example itself is modified. For instance, imagine a classifier of pictures that is trained to distinguish cats from dogs, but its performance is sensitive to the background of the pictures. Starting with pictures of cats and of dogs, one could create artificial pictures by changing the background of the original pictures (Bloice et al. 2018). The model will then be retrained with these new additional training examples. Hopefully, afterwards the model is better able to distinguish cats from dogs, irrespective of the background of the picture. For text mining classifiers similar methods are under development to create new artificial training examples from the original ones (Mueller and Thyagarajan 2016; Zhang and Yang 2018).

Stability

A term closely related to robustness is the 'stability of a statistical learning model', which refers to the extent to which a model produces consistent predictions with small perturbations in the training set. So, when the algorithm would have been offered a slightly different training set, how would this affect its model performance? This latter term is also used to quantify generalization of the model.

Different metrics have been developed to quantify stability. These metrics are based on comparing two outcomes of the loss function of the model with each other: one when the algorithm is trained on the full training set and one where subsequently 'one unit at a time' is left out of the training set. Depending on the exact metric different forms of stability have been defined: hypothesis stability, error stability, uniform stability, leave-one-out-stability, see Bousquet and Elisseeff (2002), Kearns and Ron (1999) and Mukherjee et al. (2006). Recently, Sun (2015) defined 'classification instability' as the probability that a model classifies the same instance in a different class for two samples that are independent and identically distributed as the actual available sample. Sun proposes an approach where a model is not only selected on the basis of its accuracy but also on the basis of its stability. Stable classifiers support trust of the general public in the model but it is also important for reproducibility of the outcomes of the model.

The stability of a machine learning model is the result of the sensitivity of the model parameters to their input values. (A change in model parameters may subsequently lead to a different classification for part of the units.) This effect also occurs with imputation models. The specific input values determine the estimated parameters of the imputation model, which in turn determine the imputed values. The method of multiple-imputation has been developed to analyze the effect of uncertainty in the model parameters (because they are based on a sample of the population) on the imputation outcomes.

Model accuracy

For statistical learning in official statistics the quality dimension 'accuracy' may refers to the quality of the prediction of new, unseen cases. This model quality concerns the level of association between the true labels and the predicted labels. In case of a binary variable this concerns four values. Numerous summary measures have been developed that summarize part or all values in the confusion matrix in a single value. The advantage of a single value is that it can directly be used to compare the performance of statistical learning models. Examples of such single measures are 'accuracy', 'precision', 'recall' , 'F1', 'Cohen's Kappa' and 'Matthews Correlation Coefficient', see Powers (2011) for an overview.

Impartiality and objectivity

To safeguard objectivity, NSIs have mainly relied on direct observations and have tried to avoid the use of models, in particular behavioral models.[2)] With the advent of new data sources it seems that the use of models becomes more necessary and more important. When this will indeed be the case, NSIs should develop guidelines so that the statistical results can be regarded as objective. Some examples have already been discussed above, and based on that, some more general rules may be stated (Buelens et al. 2014; Braaksma and Zeelenberg 2015; Braaksma et al. 2019, section 4.3; McLaren and Drew 2015):

1. The model should be used only for estimating missing data. That is, the model should only be used for the time period for which the data are available. This precludes forecasting and analysis of expected effects of policy measures. The use of models to analyze the effectiveness of policy measures that have already been taken in the past and for which the data have been collected already is acceptable.

2) This is also reflected in the methodological discussion on design-based versus model-based statistics; this has led to the compromise of model-assisted methods, which are meant to be more robust against model failure. It is important to note that in some areas, e.g. non-response adjustment, small-area statistics, and editing, NSIs already rely on complex models.

2. The variables in the model should be directly related to the statistical topic for which the model is used. That is, both the entities and the populations related to the model should reflect those of the statistical phenomenon in question. This precludes the use of behavioral variables, since these are indirectly related. Note that the question is how strict we wish to apply this rule. In deep learning models for instance there is hardly any feature engineering. As a result model predictions will also depend on variables that are indirectly related to the topic of interest. This is also very common in classical machine learning models. It can only be avoided when much effort is put into feature selections procedures, possibly at the cost of model accuracy.

The model specification should be extensively tested against alternative specifications, the model should be robust against outliers and breaks in the data, and the model should be stable over time. Some of the aspects have been discussed already.

Validity

Generally speaking, validity refers to the confidence that one puts in a set of results. For confidence in the predictions of a statistical learning model, it is important that the training and test procedure is appropriate. For instance, an appropriate approach to estimate the accuracy metrics is to use a test set which has been set aside and which has not been used in model and hyperparameter selection procedures. Furthermore, the test set should be representative for the target distribution that one aims to predict. Validation procedures also involve human judgement. For instance, it is important to assess whether features that strongly influence the model predictions "make sense." That is for instance important to ensure that the model predictions are generalizable.

When statistical learning models are used to make population estimates over time one has to make sure that the model is kept up-to-date. It is therefore important to have a valid procedure to keep the model up-to-date. One aspect of that will be to refresh the training and test set regularly and to retrain the model. Every now and then a more complete check of the model performance, including judgement of the features should be made.

In classical modeling, it is good practice to validate whether the model assumptions hold in practice. This is also a very important aspect in statistical learning. In doing so, one should keep the purpose for which the model is applied in mind. For instance, the conditional independence assumption underlying the Naïve Bayes model in text mining is known to be violated often in practice. As a consequence the estimated probability for a case to belong to a certain class is estimated less accurately. Still, there are other conditions which explain why a Naïve Bayes model often has a good accuracy when it is only applied to determine the zero-one membership of a category (Domingos and Pazzani 1997). However, it is important to check validity in every case, since NSIs cannot afford to base statistics on models that might break down.

4.3 Input and process quality

Once the statistical office has decided that a new data source, in combination with a certain means of estimation is a promising one to make official statistics, more structural information about the data source and its metadata are needed.

Once that has been arranged, important aspects of quality with new data sources are determined by the occurrence of all kinds of errors in the data and the processing of the data. Some examples of error types occurring with new data sources are given below.

Coverage errors

Big data sources sometimes cover only a select part of the population. For instance mobile phone data have an under coverage of elderly people. Sometimes it is unclear whether a new data source is selective or not. With Twitter data, for instance, the population of units is often not so clear. Another example is the use of website texts to derive enterprise characteristics. Not all enterprises have a website, but whether this is selective or not might depend on the variable of interest. Research on how to correct for selectivity is still ongoing.

Measurement errors

The raw input values of big data sources may have measurement errors that need to be corrected for during processing of the data. Typically, measurement errors in big data sources occur in data based on sensors. The sensors in mobile phones for instance differ in quality and they are not always correctly calibrated. That may lead to measurement errors. Furthermore, mobile phones may be stolen and as a result the obtained measurements may refer to another person than the official registered owner. Sensors of AIS system for boats may be inaccurate, leading to positions of boats on land rather than in the water. Another typical kind of measurement error in big data are missing sensor data. For instance in road sensor data 98% of the sensors lacked at least one minute of data (Puts et al. 2019).

Processing errors

An example of where the correction of processing errors is different in big data than in survey and administrative data is with high-volume data that are streamed rather than stored. In case of processing errors in such a source we cannot directly correct the microdata and compute an updated release. In case of a serious processing error with streaming data one can either decide not to publish the data or one can try to correct for the erroneous process at macro level. In practice it is very important to invest in a good monitoring system in case of streaming data that checks whether the data are of the expected quality. For instance one might count the number of valid cases in the streaming data.

Modeling errors (from raw to statistical data)

Some of the new data sources cannot directly be used as input for statistics, but need to be transformed first. For instance, by analyzing contacts of mobile phones with phone masts, one 'translates' mobile phone data to data about who (person) was where (location) at which moment. For AIS signal data a similar translation is made for vessels. Another example is translation of traffic loop measurements along a trajectory into traffic intensity. Often this transformation involves changes from input object types (mobile phone signals) into statistical object

types (persons), or a translation from events (traffic loop measurements) into totals or rates (traffic intensity). This transformation from raw to statistical data usually involves a model-based approach and hence model assumptions. With this, model errors may occur. A special set of models are machine learning models. These were already discussed under "methodological quality".

4.4 Output quality

When statistics based on new data sources are being released as official statistics, we have to take care of output quality, that is, the quality they have for the external user. We then have to consider the quality aspects relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and accessibility and clarity as well as statistical confidentiality and data protection.

Relevance

Since statistical output is a public good and publicly financed, it should always be relevant, i.e. there must be definite or prospective users and the society must be potentially interested in the output topic. This means that it is not the newness of the data source or the statistical technique that determines whether a statistical output will be produced and, by extension, whether R&D will be carried out. Research without a direct route to relevant statistical output is only justified when the aim is to investigate new methodology or new applications of existing methods.

Accuracy

We are usually interested in the accuracy of population estimates per category. New data sources often make use of models to make estimates, for instance machine learning classifiers. For that group of models one estimates the accuracy of aggregates by using the methodology presented in Van Delden et al. (2016). Meertens et al. (2019) applied that method to the situation of statistical learning and showed that in many cases the errors that are made by statistical learning algorithms to predict different categories, the false positives and false negatives, do not cancel out and lead to considerable bias in estimates of population totals when this is not corrected for. In practice, there can be another source of bias for population estimates with big data, namely when those data are available for a selective part of the population that one is interested in. Some methods to try to correct for this selection bias can be found in Elliot and Valliant (2017).

Timeliness and punctuality

Clearly, many new data are much more timely, at least with the owner or provider. If the data can be provided quickly to the NSI, and processing is standardized and is also done quickly, new data have a definite advantage over survey and administrative data. If, for commercial reasons or because of confidentiality or data size, data cannot be provided to the NSI, an alternative may be to do the processing at the data provider.

Coherence and comparability

Measuring changes in official statistics is often more important than measuring levels in a single period; for example economic growth is much more interesting than the GDP level. With volatile data sources, measuring changes is however difficult and sometimes impossible or only possible at considerable cost, and comparability across time may be severely hampered.

Statistical integration, where a phenomenon is viewed from different angles and datasets, and comparability across groups, such as household and industrial groups, are strategic priorities for NSIs. Therefore it is often necessary to link several datasets. However, as mentioned above, new data sources sometimes lack variables that allow linking with other datasets. It is important to develop new methodology to deal as far as possible with such data.

Accessibility and clarity

Official statistics should be presented transparently, i.e. "in a clear and understandable form". Together with the principle of appropriate statistical procedures, this implies that we should be able to describe how the statistics have been created, and thus how the data have been collected and processed. With new data sources it is essential to have information about data collection, initial data processing and quality control. If a commercial data provider is unwilling to provide such information or to allow publication of it, NSIs will in general have to refrain from using such data sources. Another aspect of accessibility and clarity is that statistics are archived. Besides the statistical output, it is good practice to store the micro data. The latter is also useful in terms of data sharing (principle 9 of the Code of Practice). For certain big data sources however, the volume of data may become so large that it is no longer feasible to store the microdata.

Statistical confidentiality and data protection

The abundance of new data poses new problems for statistical disclosure control, both strategic and ethical problems as well as methodological problems, which need to be addressed but for which there are at present no clear-cut solutions. For example, Recital 26 of the European General Data Protection Regulation (GDPR) implies that NSIs, when protecting tables or data files against disclosure, take into account the possibility that government agencies and private companies will use the NSI data to enrich their own databases and thereby get to know more about their citizens or customers. And what if these enriched data are used to profile citizens so that they may come under suspicion or are denied access to certain services such as loans? Should we use another methodological paradigm than the one we have used so far? Should we consider different disclosure scenarios? Does the changing attitude towards privacy influence the way we should treat our published data? How much existing as well as future data should we take into account when assessing disclosure risks?

4.5 A need for additional quality dimensions?

How do the above mentioned methodological quality aspects relate to the ESS quality dimensions (Eurostat 2014a) and its broadening into the European Statistics Code of Practice (Eurostat 2017) that were in place before the use of statistical learning models and use of big data became to spread over the community of official statistics? To what extent do they concern new elements that were not considered before?

Starting with research quality: aspects related to 'Access to the data' are part of SIMS (Eurostat 2014b) although it is limited there to contact information of the owners of the data. Presence and completeness of metadata can also be found in SIMS. What would be good to add to SIMS is monitoring of changes in the structure or in the source and of the metadata.

Concerning methodological quality, the aspect 'interpretability' of models can be seen as an extension of the quality dimension 'accessibility and clarity'; that dimension underlines that statistics should be presented in a clear and understandable form. We extend this by stating that the models used to produce the statistics should also be understandable. The 'robustness' of models can be seen as an extension of the quality dimension 'coherence and comparability'. Comparability is about measuring the impact of differences in concepts, procedures and measurement tools where statistics are compared over different geographical areas or over time. With robustness one is interested in the impact of differences in the inputs on statistics which are compared over time. The aspects 'stability' and 'prediction accuracy' can be seen as part of the quality dimension 'accuracy and reliability'. Where the quality dimension 'accuracy and reliability' relates to output statistics that are produced, stability and prediction accuracy concerns the model performance from which output can be constructed. We have explained already some of the relations between those two. Finally, 'validity' can be seen as covered by principle 8 from the Code of Practice: 'appropriate statistical procedures': appropriate procedures [...] underpin quality statistics.

In summary, we conclude that the methodological quality aspects related to statistical learning and big data clearly contain new elements compared to the well-established ESS output quality dimensions. Those new elements can be seen as extensions of the well-established dimensions rather than completely new quality dimensions of their own. It would therefore be desirable to include these aspects in the quality assessment frameworks based on the Code of Practice and in the quality guidelines of NSIs.

The error types that have been mentioned above concerning input and process quality and the bias and variance of output quality, have all been mentioned in Eurostat (2014a) under the heading of accuracy. So the error types and quality measures are not new, they are just specific instances for certain new data source.

5 Conclusion

In this paper we raise the paradigm question of how to optimally rebalance statistical quality dimensions and how to maximally exploit new data sources and methods for the production of official statistics. The discussion echoes earlier discussions when administrative registers were introduced as a data source half a century ago in the Nordic countries. We discussed the strengths and weaknesses of survey data, administrative data and new data sources. The new data sources and methods have the potential to speed up the production process, to improve the level of detail and to detect new phenomena. They also provide methodological, technical and cultural challenges, summarized below.

Arguably the largest methodological barrier, the unknown data generating mechanism, can be tackled by using information in the non-probability sample as covariates in small area or structural time series models, by predicting the information on missing units using a statistical model or machine learning technique, or by linking the non-probability sample to a probability reference sample. The low signal-to-noise ratio can be tackled by applying filtering and dimension reduction techniques. Profiling may be applied to convert events to units. Strong case studies will be needed for risk-averse NSIs to give up control and traditional robust methods.

Technical barriers are the need for a strongly connected but secure infrastructure, a computing environment that enables fast data processing and analysis, access to data science tools including the latest packages and pre-trained models, and adoption of new tools and workflows to shorten and improve statistical dissemination.

Cultural barriers can be plentiful but they stem from differences between people's goals and expectations. Stated boldly, output-driven survey methodologists seek knowledge through methods based on sound mathematical theory, whereas input-driven data scientists seek fit-for-purpose products through more pragmatic approaches.

Finally, we discussed to what extent the existing quality frameworks cover the quality issues that arise when new data sources and methods are considered for the production of official statistics. These were grouped into aspects of research quality, methodological quality and input/process/output quality. New quality aspects were identified that are associated with new data sources (e.g. supplier relationship management, metadata) and machine learning techniques (e.g. interpretability). These can be integrated into existing quality frameworks.

We conclude that new data sources and methods provide exciting opportunities for faster, cheaper and new statistics, but pose quality risks and seriously challenge the existing production process. The main (cultural) challenge is to stimulate a curious mindset, mutual respect for one's area of expertise and crossovers between computer science and statistics. The ESS quality principles, as laid down in the Code of Practice, as such need not be adapted to the emergence of new data sources. The statistical and institutional principles apply equally to these new data. However, it is important that at a lower level, at the indicator level of the Code of Practice and in the ESS Quality Assurance Framework, many of the topics discussed in this paper have to be introduced. Much preparatory work has already been done by the UNECE (2014) and the ESSnet Big Data (Consten et al. 2018a, b). One of the next steps could be a quality handbook for big data.

References

Baker, R. (2017). Big data, a survey research perspective. In: Biemer et al. (eds.), Total Survey Error in Practice. Wiley, New York.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile and R. Tourangeau (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1: 90–143.

Benedikt, L. (2019). Can social media data improve official statistics? March 9, weblog located at: https://blog.ons.gov.uk/2019/02/05/can-social-media-data-improve-official-statistics-not-yet-suggests-new-work-on-tourism/.

Bloice, M.D., C. Stocker and A. Holzinger (2018). Augmentor: An image augmentation library for machine learning. arXiv:1708.04680v1.

Blumenstock, J., G. Cadamuro and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science* 350: 1073-1076.

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg and M. Smeets (2008). Towards small area estimation at Statistics Netherlands. *METRON International Journal of Statistics*, vol. LXVI: 21–49.

Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *Journal of Machine Learning Research* 2: 499–526.

Bosch, O. ten, D. Windmeijer, A. van Delden and G. van den Heuvel (2018). Web scraping meets survey design: combining forces. BigSurv18, Barcelona.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l' Institute International de Statistique* 22, Supplement to Book 1: 6–62.

Braaksma, B. and K. Zeelenberg (2015). "Re-make/Re-model": Should big data change the modelling paradigm in official statistics? *Statistical Journal of the IAOS* 31: 193–202. https://doi.org/10.3233/SJI-150892.

Braaksma, B., K. Zeelenberg and S. De Broe (2019). Big data in official statistics: a perspective from Statistics Netherlands. In: L. Japec, L. Lyberg et al. (eds.), Big Data Meets Survey Science. Wiley, New York.

Buelens, B., P.-P. de Wolf and K. Zeelenberg (2014). Model-based estimation at Statistics Netherlands. Report. Statistics Netherlands, The Hague/Heerlen.

Buelens, B., J. Burger and J.A. van den Brakel (2018). Comparing inference methods for non-probability samples. *International Statistical Review* 86: 322–343.

CBS (2018). Social tension indicator based on social media. Beta-product of Statistics Netherlands. Located at: https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media.

Cochran, W. (1977). *Sampling Theory*. Wiley, New York.

Consten, A., V. Chavdarov, P. Daas, V. Horvat, J. Maslankowski, S. Quaresma, M. Six and T. Tuoto (2018a). Report describing the methodology of using big data for official statistics and the most important questions for future studies. ESSnet Big Data, Deliverable 8.4.

Consten, A., V. Chavdarov, P. Daas, V. Horvat, J. Maślankowski, S. Quaresma, M. Six and T. Tuoto (2018b). Report describing the quality aspects of big data for official statistics. ESSnet Big Data, Deliverable 8.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP8_Reports,_milestones_and_deliverables1.

Daas, P.J.H., B. Braaksma, R. Aly, Y. Engelhardt, D. Hiemstra and R. Zurita Milla (2016). Big Data Masterclass and DataCamp 2015. Discussion paper 201615. Statistics Netherlands, The Hague/Heerlen.

Daas, P., C. Harmsen and M. Offermans (2019). Results of the quality study on Mezuro mobile phone data (in Dutch). CBS report, 22-03-2019.

Daas, P.J.H., S.J.L. Ossen, R.J.W.M. Vis-Visschers and J. Arends-Toth (2009). Checklist for the quality evaluation of administrative data sources. Discussion paper 09042, Statistics Netherlands, The Hague/Heerlen.

Daas, P.J.H. and M.J.H. Puts (2014). Big data as a source of statistical information. *The Survey Statistician* 69: 22–31.

De Broe, S., R. Meijers, O. ten Bosch, B. Buelens, B. Laevens, A. Priem, T. de Jong and M. Puts (2019). From experimental to official statistics: The case of solar energy. *Statistical Journal of the IAOS*, DOI:10.3233/SJI-180458.

De Waal, T., A. van Delden and S. Scholtus (2019). Multi-source statistics: Basic situations and methods. Accepted by the *International Statistical Review*.

Domingos, P. and M. Pazzani (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29: 103–130.

Elliott, M.R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science* 32(2): 249–264.

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

Eurostat (2014a). ESS handbook for quality reports. Eurostat manuals and guidelines.

Eurostat (2014b). Technical manual of the single integrated metadata structure (SIMS). Eurostat, Luxembourg.

Eurostat (2017). European Statistics Code of Practice, third edition 2017. Eurostat, Luxembourg. http://ec.europa.eu/eurostat/web/quality/overview and https://doi.org/10.2785/798269

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

Groen, J.A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics* 28: 173–198.

Grutter, M. (2019). Semiautomatic reports with R Markdown. 7th International Conference on the Use of R in Official Statistics (uRos2019). `https://www.r-project.ro/conference2019`.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society Series A* 181: 555–605.

Hansen, M.H. and W.N. Hurwitz (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14: 333–362.

Harvey, A.C. and C. Chung (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society Series* A 163: 303–339.

Hero, A. (2013). Correlation mining in massive data. Presentation for Electrical Engineering and Computer Science department meeting, University of Michigan. Available at: `http://www.eecs.umich.edu/eecs/pdfs/events/2711.pdf`.

Hsieh, Y.P. and J. Murphy (2017). Total Twitter error, decomposing public opinion measurement on Twitter from a Total Survey Error perspective. In: Biemer et al. (eds.), Total Survey Error in Practice. Wiley, New York.

Isaksson, A. and G. Forsman (2003). A comparison between using the web and using the telephone to survey political opinions. In *Annual Meeting of the American Association for Public Opinion Research*, Nashville, TN, 100-.106.

Karr, A.F. (2017). The role of statistical disclosure limitation in Total Survey Error. In: Biemer et al. (eds.), Total Survey Error in Practice. Wiley, New York.

Kearns, M. and D. Ron (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation* 11: 1427–1453.

Krause, J., A. Dasgupta, J. Swartz, Y. Aphinyanaphongs and E. Berini (2017). A workflow for visual diagnostics of binary classifiers using instance-level explanation. IEEE Conference on Visual Analytics Science and Technology, arXiv:1705.01968.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

Lazer, D., R. Kennedy, G. King and A. Vespignani (2014). The parable of Google Flu: Traps in big data analysis. *Science* 343: 1203–1205.

Lesur, R. (2019). Reproducible corporate publications using pagedown. 7th International Conference on the Use of R in Official Statistics (uRos2019). `https://www.r-project.ro/conference2019`.

Loumaranta, H., M. Puts, G. Grygiel, A. Righi, P. Campos, C. Grahonja and T. Speh (2018). Early estimates of economic indicators. ESSnet Big Data, Deliverable 6.8.

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

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

McLaren, C. and S. Drew (2015). Practical guidance on modelling for the implementation of changes to National Accounts outputs. ONS Methodology Working Paper Series no 2. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries.

Meertens, Q., A. van Delden, S. Scholtus and F. Takes (2019). Bias correction for predicting election outcomes with social media data. 5th International Conference on Computational Social Science (IC2S2), July 17–20, Amsterdam.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics* 12: 685–726.

Mueller, J. and A. Thyagarajan (2016). Siamese recurrent architectures for learning sentence similarity. Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ.

Mukherjee, S., P. Niyogi, T. Poggio and R.M. Rifkin (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics* 25: 161–193.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97: 558–625.

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

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

Powers, D.M.W. (2011). Evaluation: from precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2: 37–63.

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

Rao, J. and I. Molina (2015). *Small Area Estimation*. Wiley-Interscience, New York.

Ribeiro, M.T., S. Singh and C. Guestrin (2016). "Why should I trust you?" Explaining the predictions of any classifier. arXiv:1602.04938v3.

Ricciato, F., M. Skaliotis, A. Wirthmann, K. Giannakouris and F. Reis (2018). Towards a reference architecture for trusted smart statistics. DGINS, Bucharest.

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 US national elections. Proceedings of the Joint Statistical Meetings, Washington, DC, 627–639.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* 57: 377–387.

Särndal, C.-E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

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

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

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 & 623–656, July & October.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail — but Some Don't*. Penguin, New York.

Struijs, P. and P.J.H. Daas (2014). Quality approaches to big data in official statistics. Paper presented at the Q2014 conference, Vienna.

Sun, W. (2015). *Stability of Machine Learning Algorithms*. PhD Thesis, Purdue University West Lafayette, IN.

Tanton, R. and K. Edwards (eds.) (2013). *Spatial Microsimulation: A Reference Guide for Users*. Springer, Dordrecht.

UNECE (2014). A suggested framework for the quality of big data. `https://statswiki.unece.org/download/attachments/108102944/BigDataQualityFramework-final-Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2`.

Valliant, R., J.A. Dever and F. Kreuter (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer-Verlag, New York.

Valliant, R., A.H. Dorfman and R.M. Royall (2000). *Finite Population Sampling and Inference, A Prediction Approach*. Wiley, New York.

Van Delden, A., S. Scholtus and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics* 32: 619–642.

Van den Brakel, J.A. (2012). Models in official statistics. Inaugural lecture 27-04-2012, Maastricht University, isbn: 978-90-5681-388-8.

Van den Brakel, J.A. (2019). New data sources and inference methods for official statistics. Discussion paper, July 2019. Statistics Netherlands, Heerlen.

Van den Brakel, J.A., E. Sohler, P. Daas and B. Buelens (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology* 43: 183–210.

Vavreck, L. and D. Rivers (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties* 18: 355–366.

Zhang, D. and Z. Yang (2018). Word embedding perturbation for sentence classification. arXiv:1804.08166v1.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands

Design
Edenspiekermann

Enquiries
Telephone: +31 88 570 70 70
Via contact form: www.cbs.nl/infoservice