The use of machine learning in official statistics

UNECE Machine Learning Team, November 2018
Wesley Yung (Canada), Jukka Karkimaa (Finland), Monica Scannapieco (Italy), Giulio Barcarolli (Italy), Diego Zardetto (Italy), José Alejandro Ruiz Sanchez (Mexico), Barteld Braaksma (Netherlands), Bart Buelens (Netherlands), Joep Burger (Netherlands)

Abstract

This paper is written for managers and policy makers to inform them about the possibilities to use machine learning (ML) in the production of official statistics and to demystify ML for those official statisticians unfamiliar with it. After providing some background on official statistics and ML, the paper explains why ML is becoming relevant for official statistics. The possible uses of ML in official statistics are presented for primary data (probability samples) using the GSBPM process model, and for secondary data (registers, big data/ non-probability samples and mixed sources). Finally, issues about quality are discussed, which we believe should be addressed with high priority.

1. Introduction

Computers have learned to paint in the style of Rembrandt (www.nextrembrandt.com) and to compose music in the style of Bach (Hadjeres et al. 2017). Chess computer Deep Blue already defeated the world champion over twenty years ago, question-answering computer Watson won the quiz show *Jeopardy!* in 2011 and last year the computer program AlphaZero learned from scratch to play three board games at superhuman level by playing against itself (Silver et al. 2017). But the fact that even creative jobs such as painting or composing music can be learned by computers without being explicitly programmed is the strongest illustration of the power of artificial intelligence.

These awe-inspiring and perhaps disturbing developments in Artificial Intelligence (AI) are driven by machine learning (ML) techniques in combination with the availability of an unprecedented amount of digital data and powerful IT infrastructures. Digital data often are merely a noisy by-product of non-statistical processes, but they contain signals of human activity that could be mined for the production of official statistics. Using big data to produce honest, precise and unbiased information that can be used for evidence-based policy making is a major challenge for producers of official statistics. ML is an indispensable tool to tackle this problem (Jordan and Mitchell 2015). In addition, the ability of computers to learn statistical tasks such as classification, regression and clustering calls for a review of the current statistical processes to see where ML can be of assistance. Chu and Poirier (2015) listed some ML applications already in use or in consideration at statistical offices. The current paper has a broader scope and aims to identify further opportunities.

This paper is written for managers and policy makers to inform them about the possibilities that we see to use machine learning in the production of official statistics. We also try to remove suspicion among some skeptical official statisticians by taking a balanced position. Statisticians can learn from computer science to solve a wider range of problems (Breiman 2001). We focus on applied or narrow AI rather than artificial general intelligence. Whether or not computers can actually create something new is outside the scope of this paper. The ability of computers to implicitly learn specific tasks suffices to qualify as a useful tool for the production of official statistics.

The structure of the paper is as follows. We start with the main conclusions and some recommendations for follow-up activities (Section 2). In the core of this paper we first give some background on official statistics and machine learning (Section 3). In the next sections, we discuss the use of machine learning in official statistics based on primary data (Section 4) and secondary data (Section 5). Primary data are collected for statistical purposes, usually by sending a questionnaire to a probability sample of the target population. Secondary data such as administrative registers and big data are not collected for statistical purposes but may contain statistical information. We close with some considerations about quality (Section 6).

2. Conclusions and recommendations

National Statistics Organizations (NSOs) are now looking more and more at secondary data sources and with it come many opportunities as well as some challenges. The processing of secondary data is steering NSOs to the use of ML techniques which include algorithmic models and other big data tools. These tools are not just in the secondary data domain: some of the techniques can, and should, be used in the processing of primary data as well. However, these tools do not lend themselves well to the traditional quality frameworks, so work is required to develop a framework that can handle both primary and secondary data sources. The predictive approach discussed here seems to have some promise but generalizing the performance in non-probability samples to statistical populations remains an unsolved challenge.

A key recommendation is hence to develop a quality framework tailored to the use of ML techniques. Traditional statistical quality frameworks assume that the data-generating process and further data processing steps are explicitly known. When applying ML methods, especially to 'found' big data or in multisource statistics, these assumptions are usually not valid. To guarantee quality, reproducibility and transparency, which are core values of official statistics, it is important to identify suitable quality indicators and performance metrics. In addition, it seems useful to design guidelines for reproducibility/transparency/causal inference of ML-based statistics.

The application of ML methods by nature implies interdisciplinary work. Modelers (methodologists), programmers (computer scientists) and subject matter specialists must work together. At the NSO level, teams should be formed that combine these different skills to achieve optimal results. An international ML project should similarly ensure a balanced composition.

As Chu and Poirier (2015) have already shown, a lot of work has already gone on and since their paper the activity has increased further. It seems worthwhile to develop and maintain an inventory of ML projects carried out in the statistical community to learn from each other and stimulate further joint work, either in methods development or in practical applications.

Finally, there appears to be sufficient interest to launch a concrete international ML project under the HLG-MOS[1] umbrella. Deliverables of such a project could cover generic aspects like design of a quality framework and guidelines or developing an inventory of ML projects, as well as specific aspects like concrete ML applications of mutual interest in areas like image recognition or automated coding.

To summarize: while it is felt that ML has potential uses in Official Statistics, there are some issues that still need to be considered. The need to develop a quality framework and the potential loss of transparency from the use of 'black box' methods are two issues that immediately stand out.

3. Background

3.1 What are official statistics?

According to the Organization for Economic Cooperation and Development "official statistics are statistics disseminated by the national statistical system, excepting those that are explicitly stated not to be official" (http://stats.oecd.org/glossary/). They provide qualitative and quantitative information on all major areas of citizens' lives, such as economic and social development, health, education and the environment. The importance of these statistics is underlined in the first principle of the Fundamental Principles of Official Statistics (UN Statistical Commission) where it states that "Official statistics provide an indispensable element in the information system of a society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation" (https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf). For the most part, official statistics are produced primarily by NSOs which are responsible for producing "high quality official statistical information for analysis and informed policy decision-making in support of sustainable development, peace and security". These official statistics are typically used by all levels of government and private citizens alike.

3.2 What is machine learning?

Machine Learning (ML) is the science of getting computers to automatically learn from experience instead of relying on explicitly programmed rules, and generalize the acquired knowledge to new settings. When there are both auxiliary information and a variable of interest, or labeled input, the machine can learn supervised because its performance can be tested; when there is only auxiliary information, or unlabeled input, the machine has to learn unsupervised, i.e. without feedback. In supervised ML, when the variable of interest is qualitative/categorical, the machine learns to solve a classification problem; when the variable of interest is quantitative/numeric, the machine learns to solve a regression problem. In unsupervised ML, the machine learns to solve a clustering problem.

Figure 1 gives a schematic overview of supervised learning which is also applicable in the context of official statistics. In official statistics, we want to accurately estimate a variable of interest y in a target population A. For instance, the unemployment rate in young people or CO_2 emissions by the aviation industry. Auxiliary information x is known for all elements in target population A but the variable of interest y is only observed for the elements in sample B. ML might be used to **make inference** from sample B to population A.

[1] High-Level Group for the Modernisation of Official Statistics

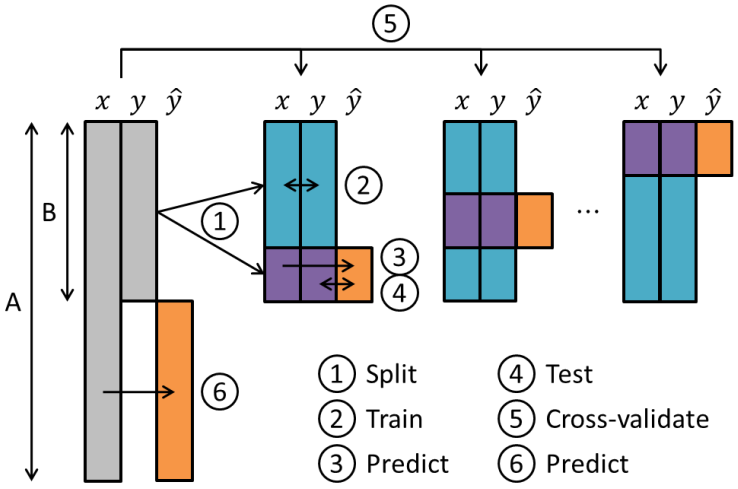

Figure 1 Supervised machine learning

Many other uses are possible. For instance, one could consider A a probability sample, B the set of respondents and use ML to **correct for unit-nonresponse**. In another scenario, A is the set of respondents, B the set with responses on both item x and y and ML is used to **impute item-nonresponse** on item y. A fourth option could be that A contains observations of a proxy variable x, B contains observations of the variable of interest y and ML is used to **model measurement error**. Finally, A could be a time series, B historic data and ML is used for **nowcasting**.

We discern six steps in supervised learning:

1. The labeled input (B) is randomly split into a training set (blue) and a test set (purple).
2. The model or algorithm learns the relationship between x and y in the training set.
3. The model or algorithm is used to predict y, $\hat{y}$ (orange) from x in the test set.
4. Predicted values $\hat{y}$ are evaluated against observed values y in the test set. Steps 2 through 4 are repeated with different (hyper)parameters until the prediction error is minimal.
5. Steps 1 through 4 are repeated for different splits to prevent overfitting. Nested cross validation may be performed to tune (hyper)parameters in an inner loop and estimate generalization error in an outer loop.
6. Unobserved values are predicted from x outside the sample using the model or algorithm that gives, on average, the smallest prediction error in the splits.

The crucial step is learning the relationship between x and y in the training set (step 2). Many algorithms exist for this purpose, such as random forests, neural networks and support vector machines. They differ from classical regression techniques by focusing on minimizing prediction error rather than causal explanation (Shmueli 2010) and are potentially better suited for modeling non-linear relationships in high-dimensional space. Moreover, x can be anything from classical numerical information to natural language text, audio recordings, images or video—just as long as the data x are in digital format. Several performance measures are available to quantify prediction error, most notably mean squared error in regression and Matthews correlation coefficient in classification.

The success of supervised ML largely depends on the predictive power of x and the size of the training data. Sample survey data are limited in the size of the training data (small n) but are typically rich in the number of auxiliary variables (large p). Big data on the other hand are named after the huge amounts of data (large n) but are often limited in the number of variables (small p). We therefore expect the biggest gain in register-based statistics with both large n and large p.

3.3 Why machine learning is becoming relevant for official statistics

NSOs are currently facing unprecedented pressure to evaluate how they operate. Years of declining response rates to primary data collection efforts and the proliferation of readily accessible data, which has made it easier for private companies to produce statistics, is putting into question the role of NSOs. In response, many NSOs are looking to tap into these alternative data sources to supplement, or even replace, data collected by traditional means.

One of these alternative data sources being investigated as a potential source for official statistics is big data (Hassani et al. 2014; Daas et al. 2015). Typically, volume, velocity and variety, the three V's of big data, are used to characterize the key properties of big data:

- 'Volume' refers to the size of the dataset.
- 'Velocity' refers to the data-provisioning rate and to the time in which it is necessary to act on them.
- 'Variety' refers to the heterogeneity of data acquisition, data representation, and semantic interpretation.

In projects implemented by NSOs to date, the three Vs do not necessarily characterize a big data source in a simultaneous way. For instance, if looking at the pilots of the recent ESSnet on big data (https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page): web scraping of job vacancies is mainly affected by 'Variety', Automatic Identification System (AIS) vessel tracking data are mainly characterized by 'Volume' and 'Velocity', Twitter data for social mood detection by 'Variety' and 'Volume'.

Big data do not naturally fit within the established quality framework of official statistics. The main reason for this is that the data generating mechanism of big data sources does not fall under the NSO's control and is typically unknown. This is evidently at odds with statistical surveys, whose data generating mechanism is *designed* by the NSO through probability sampling, but also with administrative data sources, whose data generating mechanism is at least *known* to the NSO. As a consequence, the development of sound methodologies to extract valid statistical information from big data is still fairly embryonic: how to guarantee the quality of inferences drawn from big data is a matter of current research.

While the *"found data"* nature of big data definitely calls for out-of-the-box statistical thinking and novel inferential approaches, even only the 'Volume' and the 'Variety' dimensions of big data can wreck the traditional computation toolbox of official statistics. These two dimensions are further discussed in Section 4.1.

If indeed NSOs move towards using big data to produce official statistics, ML techniques will almost surely become an indispensable tool. However, we feel that ML techniques do not have to be restricted to the domain of big data and many of them can be used in the current environment. As previously mentioned, depending on the format of the variable of interest (qualitative versus quantitative) supervised ML techniques lead to classification or regression and unsupervised learning leads to clustering. In the traditional framework, supervised learning could be used for imputation for missing data, prediction of response propensities, construction of groups for imputation, reweighting or calibration or coding to standard classifications. Unsupervised methods can be used for outlier or error detection.

These situations are more fully developed in Section 4 where we discuss the use of ML in primary data. Uses of ML in secondary data, including big data, are discussed in Section 5.

4. ML in primary data

We first consider ML in a primary data processing framework. To be more specific in which processes could benefit from ML, we use the Generic Statistical Business Process Model (GSBPM) as an underlying structure (https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0). The GSBPM was developed under the auspices of the United Nations Economic Commission for Europe. It describes and defines the set of business processes needed to produce official statistics. The GSBPM places its focus on producing official statistics using primary data sources such as sample surveys and as such does not touch much on secondary data sources such as register or big data. The use of ML in these situations will be discussed later in the document. Version 5.0 of the GSBPM is given in Figure 2.

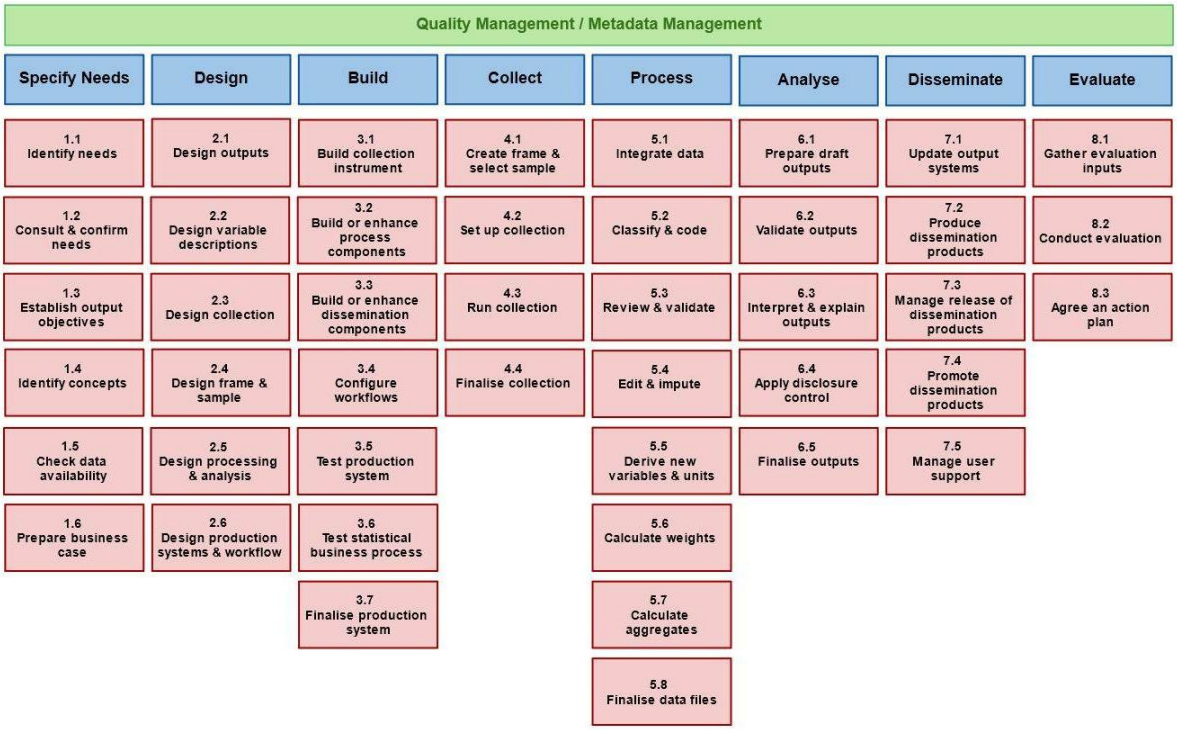

Figure 2 Generic Statistical Business Process Model

As one can see, the GSBPM consists of eight phases with sub-processes contained with each phase. The first phase (Specify Needs) covers activities associated with engaging data users to identify their data requirements and, as such, we do not see many opportunities for ML in this phase.

In the second phase (Design), the statistical outputs, concepts, methodologies, collection instrument and operational processes are defined. The sub-process 2.4 (Design frame and sample) is an area where it is felt that ML could be applied. This sub-process, which only applies to statistical processes involving collection of data from a sample, includes the identification of a sampling frame. Sources of sampling frames include registers, administrative data, censuses and other surveys. These sources may be combined through record linkage processes where clustering algorithms could be used. When preparing sampling frames, the quality of the design information (industry, geography, activities, occupations, etc.) is of the utmost importance. These 'coding activities' are prime candidates for classification algorithms.

In addition to coding type activities, ML can be used to validate the quality of the design information on the frame. For example, clustering methods can be used to identify outlying values in the design information.

In the third phase (Build), there does not appear to be any opportunities to use ML as this phase mostly involves the creation and testing of production solutions.

The fourth phase (Collect) includes sub-processes to create the frame, draw the sample and perform the collection. ML can be used in several of these sub-processes. Many surveys will stratify their frame in order to improve the efficiency of the resulting estimates. Classification techniques could be used to stratify the population in sub-phase 4.1.

The next sub-process where ML could be used is 4.3 (Run Collection). While ML may not be helpful in the actual collection operation, it may be used in the management of the collection activity. Many NSOs are employing active collection management strategies to make the collection process more efficient. Many of these strategies use response probabilities that need to be estimated or predicted. Both traditional methods (e.g. logistic regression) and ML methods (e.g. regression algorithms) can be used to predict the probability of response for individual units using information available for the entire sample. These predicted response probabilities can then be used to manage collection activities in the most efficient manner. In order to improve the fit of the response propensity models, which are used to estimate the probabilities of response, they are commonly estimated for groups of units which exhibit similar behaviors. Classification methods can be used to 'optimally' define these groups before estimating the propensity models within each of them.

In adaptive collection designs, features such as the survey mode and the use of incentives may be tailored to subpopulations. Auxiliary information or paradata could be used in combination with clustering, classification or regression algorithms to classify units into the different subpopulations for improved response rates.

Also during collection, in particular with electronic collection, data are verified as they are reported. Clustering methods could be used to identify outlying or erroneous data points during collection so that they can be corrected by the respondent or supervised learning could be used to predict the correct value. Finally many NSOs offer respondents the opportunity to record comments or questions. Natural language processing tools could be used to process these items and identify any of them which need to be responded to immediately.

Moving to the fifth phase (Process), there are many opportunities for ML methods. Sub-process 5.1 (Integrate data) integrates data from possibly multiple sources, so probabilistic record linkage techniques could be used if a common unique identifier is not available on all data sources. Once integrated, ML methods could be used to cleanse the data (identify outliers, errors, inconsistent records, etc.). Sub-process 5.2 (Classify and code) classifies and codes data to standard classification (industry, geography, commodities, etc.) and is a classic use of ML.

Sub-process 5.4 (Edit and impute) deals with replacing data which are considered as incorrect, missing or unreliable. Supervised ML can be used to impute for missing or incorrect data. Since imputation is usually improved when homogeneous units are grouped into imputation classes, ML methods similar to those used for stratification can be used to define these classes.

Sub-process 5.6 (Calculate weights) is another area where classes are formed in a similar fashion to imputation classes, however for slightly different purposes. In the presence of unit non-response reweighting classes are often formed of units with similar probabilities to respond. It is common to use logistic regression models to estimate these response probabilities and then homogeneous groups are formed based on these response propensities. ML methods could also be used to define these groups. If auxiliary data are available, many surveys use a calibration estimator with multiple calibration groups which can be defined through the same ML techniques mentioned above. For both imputation classes and calibration group definition, ML techniques can be used to choose the best variables to be used in defining the class or group (subset selection).

The sixth phase (Analysis) deals with preparing the statistical outputs and involves validation and interpretation of outputs and disclosure control. The validation and interpretation work is typically done by analysts but ML methods could possibly be used to identify outlying estimates. In addition, the classification error of a ML algorithm applied to privatized data might be used to strike a balance between privacy and utility during disclosure control (Mivule and Turner 2013).

While the possibilities mentioned in this section may not be exhaustive, it is clear that there are many opportunities for ML in the processing of primary data. Table 1 summarizes opportunities for ML by task.

Table 1 Possible use of machine learning in tasks encountered with the use of primary data

Task	Family of ML techniques	GSBPM phase
Record linkage	Clustering	2.4, 5.1
Coding	Classification	2.4, 4.3, 5.2
Outlier detection	Clustering	2.4, 4.3, 5.1, 6.2
Stratification	Classification	4.1, 4.3, 5.4, 5.6
Estimation	Regression/classification	4.3
Imputation	Regression/classification	5.4
Calibration	Regression/classification	5.6
Disclosure control	Regression/classification	6.4

5. ML in secondary data

As previously mentioned, secondary data are not collected for statistical purposes but may contain statistical information of interest to NSOs. These data include big data, register data and the combination of these data with other sources, including possibly primary data. In this section, we discuss the use of ML techniques for big data, register data and multisource data.

5.1 Big data

As introduced in Section 2.3, the 'Volume' and the 'Variety' dimensions of big data challenge the traditional computation toolbox of Official Statistics. Here we discuss issues arising from 'Volume' and 'Variety' and provide arguments in favor of ML as a possible way to address such issues.

Volume

Classical methods in Official Statistics – like design-based and model-assisted survey sampling theory, regression estimation and so on – have been devised to process probability samples. As a consequence, mainstream algorithms implementing these methods perform well on small amounts

of high quality data. However, these algorithms typically exhibit high computational complexity; a feature that hinders their ability to be used on huge amounts of data.

Consider linear regression for example: on a dataset with n observations and p variables, least squares estimation has $O(p^3 + n\,p^2)$ time complexity. Things are even worse in the case of iterative methods such as iteratively reweighted least squares or when solving calibration tasks by means of the Newton-Raphson algorithm. Not only are these algorithms so computationally demanding to become impractical for big data (especially when either p, or both n and p are very large), what is worse is that they are difficult to optimize by means of *divide-and-conquer*, or parallelization, approaches. As a result, many mainstream algorithms in computational statistics cannot easily take advantage of successful big data processing models and big data software frameworks like MapReduce and Spark/Hadoop.

Besides being computationally intensive and hard to parallelize, many classical methods in official statistics are extremely sensitive to outliers and erroneous values, a circumstance that leads to tremendous efforts being made by NSOs in editing and imputing survey data. This challenge is exacerbated with big data as they are often very noisy and poorly curated and thus contain numerous outliers and erroneous values. In addition, the size of big data simply makes it impractical to perform complete and thorough data checking and cleaning.

Modern ML approaches are better positioned than traditional statistical methods to **enable scalability** on big data sources. This may seem paradoxical at first, since successful ML methods like Random Forest or Deep Learning are well known to be computationally expensive and under certain circumstances, more expensive than traditional methods. However, many ML methods are **easily parallelizable** which make them prime candidates for processing big data. For instance, the Random Forest method is easily parallelizable because the trees composing the forest are built and trained independently on different samples. As a consequence, many fully distributed off-the-shelf implementations of Random Forest exist, which allow for very efficient processing of big data. Similarly, Deep Learning models, despite being computationally intensive, are inherently parallelizable, since neurons within each layer of the neural network are processed independently. Therefore Deep Learning applications can take advantage from specialized hardware architectures which are optimized for massively parallel computing. This way, tremendous reductions in computing time can be obtained, making it practical to train very complex Deep Learning models on big data.

Moreover, many ML methods generally show **less sensitivity to outliers and erroneous data** than most classical statistical methods do. This follows from the fact that state-of-the-art implementations of these methods rely on *subsampling* approaches. As already mentioned, each tree within a Random Forest is fit on a subsample of the original data, which in turn involves only a random subset of the original explanatory variables. After fitting, the predictions of all the trees are aggregated or averaged. Subsampling (during training) and averaging (in prediction) implicitly *smooth* the input data, largely mitigating the effects of outliers and errors. Indeed, outliers and errors, being *relatively* rare combinations of values, will contaminate and bias just a *minority* of the subsamples/trees, therefore leaving the *overall* model almost unaffected. A very similar mechanism of sub-sampling makes Deep Learning applications robust against anomalous data.

As a last (and perhaps conclusive) reason that a switch to ML is in order, we note that state-of-the-art big data analytics frameworks adopt ML techniques that cover simple classical models such as linear regression.

Variety

In Official Statistics, both survey data and administrative data are very much structured. The data model that is typically used within a traditional processing pipeline is the "case by variable" matrix, where cases are represented by records, variables are represented by records' fields, and records have a fixed number of fields of definite type.

The 'Variety' dimension of big data deals with loosely structured or even unstructured data. Traditional methods like generalized linear models work with "case by variable" data and hence cannot easily cope with data where observed variables can change from one case to another. If data are entirely unstructured, for example natural language texts, the challenges with traditional methods are even bigger, because a natural notion of "variable" no longer exists. Instead, meaningful features have to be somehow *extracted* from raw data. When performed by human analysts, this data-preparation step is called *feature engineering*. On the contrary, some ML techniques (notably Deep Learning) have the ability to *automatically* extract features from raw data that are useful for the task at hand.

5.2 Register data

Register data are often referred to as big data because enumerating a population results in a big dataset. The distinction is not based on size but structure: a register is a complete list of identifiable objects in a population (Wallgren and Wallgren 2007). By this definition, sensor data may or may not be (repeated measures) register data, depending on whether or not they can be linked to identifiable population units. Administrative registers are secondary data to statistical offices because they are maintained by other organizations for administrative rather than statistical purposes. National legal frameworks may provide statistical offices access to register data. Statistical registers are a goldmine for statistical analyses, especially to study small domains, rare events and longitudinal processes (Connelly et al. 2016). Given the large number of cases (large n) and rich set of auxiliary variables (large p), statistical registers also provide ample opportunities for ML (Thompson 2018). These are not yet fully appreciated.

Social scientists are interested in understanding social phenomena but often without the ability to conduct experiments. For instance, what distinguishes people that move to a new address from people that stay? What auxiliary variables relate to people finding or losing a job? What kind of company trades internationally? What are the profiles of companies which have gone bankrupt? How do social participation and trust differ between subpopulations (CBS 2015)? These questions can be answered using the observational data in statistical registers. They focus on association rather than causation due to the non-experimental nature of the data (Hand 2018). At first sight, there seem to be a dichotomy between young and old, lowly and highly educated, rich and poor, native and migrant, religious and secular, flex and permanent employees, urban and rural. Such potential determinants are, however, highly correlated and confounded. To isolate the effect of one, it would be necessary but practically impossible to correct for all the others. That is, what is the effect of age within lowly educated, rich, native, religious, flex employees in urban areas?

In situations like this, stepwise regression is routinely applied to select the model that optimally balances goodness of fit and parsimony. Dimension reduction techniques and mixed-effects models may be used to compress the number of parameters to be estimated. Lasso and ridge regression additionally penalize the magnitude of the regression coefficients to prevent overfitting. Model averaging may be applied when support for the best model is strong but not unequivocal (Symonds and Moussalli 2011).

Although parametric model-based methods are powerful and insightful, they suffer from two main issues. First, when the number of potential predictors is high, the number of possible models quickly becomes unwieldy. Second, both nonlinear relationships and higher-order interactions need to be defined explicitly. Within official statistics, ML methods such as neural networks are an underexplored alternative to **model nonlinear relationships and complex interactions**. Unsupervised, objects can be clustered in the high-dimensional space spanned by the rich set of features. Supervised, the clustering in high-dimensional space can be used to impute missing observations (de Waal et al. 2011) or to extrapolate relationships to unobserved subpopulations.

5.3 Multisource Statistics

Multisource statistics are based on multiple data sources such as combinations of one or more surveys, administrative registers or big datasets. We do not discuss techniques and methods to integrate, link or match data sources, but more the role of ML in analyses that can be conducted on linked data. Readers interested in the topic of data linking as such can consult, for example, Christen (2012) or Harron et al. (2015). Authors considering data linkage or matching specifically in relation to administrative or big data sources include, for example, Harron et al. (2017) and Lohr and Raghunathan (2017). An important determinant of which ML approaches are applicable in multisource statistics is the degree to which data sources can be integrated. We distinguish three levels of linkage and discuss the possibilities for use of ML in each. We pay particular attention to big datasets in relation to the more traditional survey and administrative sources for official statistics.

Micro-level linkage – micro integration

Micro-level linkage is achieved when individual units in multiple datasets can be associated with each other which often requires the presence of unique identifiers. When units observed in a big dataset can be linked to units present in one or more administrative registers, the big data records can be enriched with administrative data by providing auxiliary variables that can be used in estimation and prediction models to predict variables only observed in the big data. At the same time, the administrative data can be considered as the population frame which may help explain the data generating mechanism of the big data source, and hence remove or reduce bias potentially present in population estimates based on big data. An example is Buelens et al. (2018), where different ML methods are compared to predict annual mileage of cars.

Linking big data to survey data is another approach that could be used to remove selection bias in big data caused by non-random selection. In this setting, measurements for the surveyed units would be obtained from the big data source. When exact linking is not possible, sample matching can be applied to seek for similar, representative units which are not necessarily identical (Baker et al. 2013). The goal again is to achieve better representativity than that of the big data set alone.

Macro-level linkage – macro integration

With macro-level linkage we refer to data linking situations in which units from multiple sources cannot be linked or matched individually, but where they can be associated at some aggregate level. Examples include people in the same municipality, or businesses in the same industry or size class. Tennekes and Offermans (2014) use mobile phone metadata that can be geo-located accurately but that cannot be linked to individuals. They propose bias corrections at aggregated levels by linking the big mobile phone data to administrative registers at the municipal level.

Explicit modeling of survey based estimates using big data sources at aggregated levels as covariates is proposed by Marchetti et al. (2015) and applied by, for example, Pappalardo et al. (2015). These

are applications of area-level models within the small area estimation framework (Rao and Molina, 2015) which do not require unit-level linkage but at the same time exploit correlations that may exist between sources. Extensions of such approaches are possible where the traditional type of regression models would be replaced by ML predictive models. Literature on such approaches seems to be lacking.

No linkage

When no linkage is possible, data from multiple sources can be used for confrontation purposes. If multiple instances of the dataset are available through time, there are still possibilities to combine the data sources through a time series approach. Temporal correlation between the series can be used to improve nowcasting or forecasting accuracy.

Van den Brakel et al. (2017) improve the accuracy of survey based estimates through a structural time series modeling approach in which a big data time series is used as an independent covariate series. The big data time series is derived from social media messages and reflects the sentiment in the text of the messages. While the messages (such as from Twitter) cannot be linked to individuals, and cannot be aggregated at a sufficiently detailed level, this approach allows for exploiting temporal correlation.

6. Quality Considerations

Point estimates are hard to interpret without information about their quality, including accuracy, precision and reliability aspects. Many NSOs have policies about informing their users of the quality of their official statistics. The most common traditional indicator of quality of a point estimate is the mean squared error, combining bias and (sampling) variance. Other, indirect quality indicators include response rates, imputation rates and coverage rates. With the move towards increased use of secondary data, the traditional use of a probability sample and an unbiased estimator minimizing sampling variance will not be applicable as the data-generating mechanism is typically not known. In addition, with the transformation from data scarcity to data deluge, the contribution of non-sampling errors to the total error will increase. Thus different measures of quality are required when secondary data and/or ML are being used.

An approach commonly used in machine learning is one which aims to minimize the prediction error in a test set. This approach is particularly useful when algorithmic models are used as they do not assume a stochastic process. Under this approach, many ML techniques perform much better than traditional approaches in terms of predictive accuracy in certain situations. Examples include face and speech recognition which call for highly complex models to reach high predictive accuracy. In other situations where the traditional methods fit the problem well from a theoretical perspective and produces high predictive accuracy, there may be no need to consider ML methods.

The predictive accuracy of supervised methods can be estimated via (nested) cross-validation; applying the method to a test set and comparing the results to the truth. However, in case the training data is a non-probability sample of the target population, the actual accuracy of the method may well be worse when applied to previously unseen data in the future and the predictions should not be trusted alone (Elliott and Valliant 2017; Buelens et al. 2018).

Predictive accuracy of a method can also be improved by using feedback information from the model to increase the training set. For instance, suppose a model is used to identify units with a high

probability of being in error. These units can be resolved manually, added back to the training set and new model parameters can be calculated, thus improving the predictive accuracy.

As NSOs continue to investigate the use of secondary data and/or ML to produce official statistics, the inferential framework based on primary data only needs to be re-evaluated and expanded to cover secondary data sources. The predictive approach discussed here seems to have promise but it needs to be fully thought out and how it can be combined with the traditional one needs to be established.

All in all, ML requires a different approach to quality than the one official statisticians are familiar with. We should be aware of possible algorithmic bias and lack of fairness when applying ML. Carryover effects may appear in supervised methods, so posterior analysis are necessary to diminish any risk and to assess the use of a specific ML technique. And last but not least, an important issue often is how to obtain sufficient transparency of ML-based results.

7. References

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

Breiman, L. (2001). Statistical modeling: the two cultures. Statistical Science, 16, 199–231.

Brakel, J. van den, Söhler, E., Daas, P. and Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Survey Methodology, 43, 183-210.

Buelens B., Burger, J. and van den Brakel, J. (2018). Comparing inference methods for non-probability samples. International Statistical Review, doi: 10.1111/insr.12253.

CBS (2015). Sociale samenhang: wat ons bindt en verdeelt. Centraal Bureau voor de Statistiek, Den Haag.

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer, Canberra.

Chu, K. and Poirier, C. (2015). Machine Learning Documentation Initiative. UNECE Workshop on the Modernisation of Statistical Production Meeting.

Connelly, R., Playford, C.J., Gayle, V. and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social Science Research, 59, 1–12.

Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P. (2015). Big data as a source for official statistics. Journal of Official Statistics, 31, 249–262.

Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. Statistical Science, 32, 249–264, doi: 10.1214/16-STS598.

Hadjeres, G., Pachet, F. and Nielsen, F. (2017). DeepBach: a steerable model for Bach chorales generation. Proceedings of the 34th International Conference on Machine Learning, Sydney.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society Series A, 181, 1–24.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. and Goldstein, H. (2017). Challenges in administrative data linkage for research. Big Data & Society, 4, doi: 10.1177/2053951717745678.

Harron, K., Goldstein, H. and Dibben, C. (2015). Methodological developments in data linkage. Wiley, Chichester.

Hassani, H., Saporta, G. and Silva, E.S. (2014). Data mining and official statistics: the past, the present and the future. Big Data, 1, 34–43.

Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects. Science, 349, 255–260.

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

Mivule, K. and Turner, C. (2013). A comparative analysis of data privacy and utility parameter adjustment, using machine learning classification as a gauge. Procedia Computer Science, 20, 414-419.

Pappalardo, L., Smoreda, Z., Pedreschi, D. and Gianotti, F. (2015). Using Big Data to study the link between human mobility and socio-economic development. Paper presented at the IEEE International Conference on Big Data, Santa Clara, CA.

Rao, J.N.K. and Molina, I. (2015). Small Area Estimation Second Edition, Wiley, Hoboken.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, J., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. and Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv:1712.01815.

Shmueli, G. (2010). To Explain or to Predict? Statistical Science, 25, 289–310.

Symonds, M.R.E. and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology, 65, 13–21.

Tennekes, M. and Offermans, M.P.W. (2014). Daytime population estimations based on mobile phone metadata. Paper prepared for the Joint Statistical Meetings, Boston, MA.

Thompson, M.E. (2018). Dynamic data science and official statistics. The Canadian Journal of Official Statistics, 46, 10–23.

Waal, T.de, Pannekoek, J. and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. Wiley, Hoboken.