Theme Report of the Editing & Imputation Group

by Florian Dumpert (Germany)

based on sound pilot study reports and fruitful discussions, and with a lot of help from Fabiana Rocci and Roberta Varriale (Italy), Bart Buelens and Anneleen Goyens (Belgium), Romina Filippini, Diego Zardetto, Marco Di Zio, Simona Toti, and Fabrizio De Fausti (Italy), Sebastian Wójcik (Poland), Claus Sthamer (United Kingdom), Christian Ruiz (Switzerland), and Philip Bell, Sean Buttsworth, and Jenny Pocknee (Australia).

September 2020

Preliminaries

It is surely uncontroversial that there is a need for NSIs to identify and deal with suspicious and missing values in datasets. It is truly also uncontroversial that there are several ways to do this. For example, this can be done in a rule based way where data items are checked whether they fulfil restrictions on their values. Another approach is to work with the distribution of (parts of) the data. Data that is not plausible should – in some sense – not be belonging to the rest of data at hand. Domain knowledge is usually a necessary component of editing. Several editing procedures are suggested (for example the GSDEM by UNECE which is internationally agreed), mostly based on the identification of the most proper statistical method to detect and treat specific type of errors. Analogously, there are several ways of dealing with missing values and a lot of approaches and goals to impute values.

This theme report provides a summary of the activities that took place in and the experiences that have been made by the members of the editing and imputation group of the UNECE HLG-MOS Machine Learning Project. Essentially, it covers the time from May 2019 to May 2020. Main goal of the editing and imputation group is to show to which extent machine learning algorithms can be used to efficiently improve editing and imputation processes in NSIs (by replacing, improving or complementing methods used so far). Some of the paragraphs mentioned below are literally or paraphrased already part of the cited pilot study reports or of a paper which has been presented at the UNECE Statistical Data Editing Virtual Workshop 2020 (Dumpert 2020b).

To make clear what the two parts (editing on the one hand side, imputation on the other) are of, it is necessary to introduce the following differentiation: for the machine learning project, we treated

- editing as the task to identify missing and problematic data (i. e. implausible values, contradictions in records, …) in data sets and
- imputation as altering values that have been classified as incorrect and inserting missing values.

Other definitions (which are not used here) treat the process of altering incorrect values as part of the editing.

The editing and imputation group has members from from Belgium (imputation; Goyens & Buelens 2020), Germany (imputation; Dumpert 2020a), Italy (editing and imputation; Rocci 2020, Rocci & Varriale 2020, De Fausti et al 2020), Poland (imputation; Wójcik 2020), UK (editing; Sthamer 2020), and co-workers from Switzerland (editing; Ruiz 2018) and Australia (imputation; Buttsworth 2020).

Motivation

This paragraph describes the motivation of the participating NSIs to look into machine learning. Often, machine learning is not used exclusively but in addition to or at least compared to an already existing process, i. e. to both other specific statistical methods and (in some cases) also to human interactive work. This is true for exploratory phases as well as for the production of official statistics and it is the case for survey and register based work in

NSIs. One of the goals is often to increase the proportion of records in a statistic that can be treated in a more automated way. Sometimes, the need for the usage of statistical methods also comes from the special situation that data from more than one source has to become combined. Another goal is an improvement in the process of reporting official numbers by delivering better (e. g., more accurate) or faster (shorter production process for the final report and its contents) predictions. Basic considerations and an embedding of the work into models like the Generic Statistical Data Editing Model (GSDEM) are provided by Rocci (2020).

Expectations on machine learning

Editing

As mentioned, editing is meant to detect "problematic" cells or items in the data, to be treated more carefully. Broadly speaking, those methods can be classified according to two main criteria: (i) Whether they are based on edit rules that data are expected to respect. They can be either hard or soft, i. e. they represent constraints on data or only expected values or relationships between variables. (ii) Explorative methods aimed at identifying anomalous data or with respect to some models thought to represent properly the data. Hence, the following aspects were mentioned as possible value added of the usage of machine learning (ML) for editing:

1. ML may discover rules that have only been "known" by intuition at first, trained in previous experience on the same process mainly through interaction by the subject matter experts. This may help
   - to conserve knowledge over time and changes in editing teams;
   - to formalize the knowledge and to improve the automated detection of "problematic cells" in data sets;
   - human editing staff to focus on validating "important", or in some sense "influential" records.
2. More concretely: A supervised machine learning model could learn from former editing results which units (records or even cells) in a data set are problematic. This means:
   - The eventual goal is to learn a model that classifies every unit of an incoming data set as "plausible" or "not plausible".
   - If such a model is sufficiently interpretable, rules that represent one possible way to classify a unit as "plausible" or "not plausible" can be extracted from it.
3. ML (as well as model based approaches) may offer a valid and efficient new instrument for the not rule based perspective on editing. This would help
   - to detect "problematic cells" which can hardly get found by intuition or rules;
   - to use not only logical but also statistical aspects in the editing process.
4. More concretely: An unsupervised machine learning model could be used to analyse data with respect to its "hidden structure" with a less need of a priori model for the data. At first glance, it means that it can help to gain efficiency to
   - find outlier candidates in and to find typical subgroups of an incoming data set;
   - identify possible (soft) edit rules to classify specific group of data as being problematic, to be further analysed.

It has also been expected that machine learning has the capacity to exploit a huge amount of information to support the design and the maintenance of the editing process features. However, this obviously requires the availability of a suitable amount of data.

Imputation

What follows shows the expectations on machine learning on imputation at the beginning of the project.

1. ML may improve prediction tasks within already existing imputation schemes (like – possibly stochastic – regression imputation or predictive mean matching). This would possibly lead to better imputation results.
2. ML may be faster in doing imputation compared to other methods once the model is learnt.

The first aspect here directly leads to the question when an imputation job is done satisfactorily. At this point, there is an intersection to work package 2 of the machine learning project that deals with quality aspects. However, there are different goals of imputation which can be summarized by citing the EUREDIT project (Chambers 2001):

1. Predictive accuracy: The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are “close” as possible to the true values.
2. Ranking accuracy: The imputation procedure should maximise preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.
3. Distributional accuracy: The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.
4. Estimation accuracy: The imputation procedure should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).
5. Imputation plausibility: The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

Although mentioned as number 5, imputation plausibility is a criterion which should be applied in addition to 1.–4. Obviously, these different goals have to use different metrics to measure their success. Machine learning may offer an additional value when there is (a) either a regression or a classification (i. e. a prediction) step within the imputation process. If the focus is on predictive or ranking accuracy, this is obvious because machine learning is known to yield good predictions. If the focus is on distributional or estimation accuracy, very often a “prediction step” is involved like in (stochastic) regression imputation or predictive mean matching. There may also value added by machine learning on the task of building imputation classes. Clustering algorithms might be useful in this situation.

Exploration

The expectations above have been checked against the results of several pilot studies. The results of the pilot studies are written down as stand-alone papers separately. On this account, this theme report only provides a short summary of the pilot studies. The first table offers insights into the motivation why the pilot studies have been conducted. The tables are ordered: first the pilot studies on editing, then the pilot studies on imputation (in alphabetical order of the countries that ran the pilot studies).

Country	E/I	Title	Legacy System and Aims
Italy (Rocci & Varriale 2020)	E	Machine Learning tool for editing in the Italian Register of the Public Administration	No legacy system, the task is new. Edit rules are of the main focus, but there are also investigations whether the application of machine learning can add value to the traditional editing process.
UK (Sthamer 2020)	E	Classification of records of LCF (Living Cost and Food) survey income data that need editing	So far, there is only manual detection of spurious records. The goal was to replace the need for manual detection by learning a supervised model from former editing steps.
Belgium (Goyens & Buelens 2020)	I	Early estimates of energy balance statistics using machine learning	Old-fashioned working methods, such as large and complex Excel sheets should be replaced.
Germany (Dumpert 2020a)	I	Machine learning methods for imputation	No legacy system. The study should show the principal behaviour of several ML methods in an imputation task. The aim was to investigate whether ML can replace other approaches in regression imputation.
Italy (De Fausti et al 2020)	I	Imputation of the variable "Attained Level of Education" in Base Register of Individuals	No legacy system. The task is new. Goal of the investigation was to determine how and where ML can give greater benefits in solving the imputation problems compared with classic statistical models.
Poland (Wójcik 2020)	I	Imputation in the sample survey on participation of Polish residents in trips	No legacy system. The goal was to achieve high predictive accuracy by imputation to avoid additional surveys.

The second table gives an overview on used data, important conducted steps, and compared algorithms.

Country	E/I	Data	Steps	Algorithms
Italy (Rocci & Varriale 2020)	E	Public Administration Database (BDAP) and the Information System on the Operations of Public Bodies (SIOPE)	comparing several variables from the two sources, identifying different types of inconsistent data, list of units regarded as important to be analysed deeper delivered by subject matter experts, identifying edit rules behind such units	decision trees and random forests
UK (Sthamer 2020)	E	pre- and post-edited LCF data for one year	data preparation, calculation of the change vector, learning models to predict the change vector	decision trees, random forests, neural network
Belgium (Goyens & Buelens 2020)	I	quarterly data, ranging from Q1 2000 through Q1 2019	z-standardization of the data, feature selection for linear regression, calculating and comparing predictions	linear regression, ridge regression, lasso, random forest, neural network, ensemble prediction
Germany (Dumpert 2020a)	I	German cost structure survey of enterprises in manufacturing, mining and quarrying	creating missing values (several proportions, several missing mechanisms), calculating and comparing predictions	k-nearest-neighbours (weighted and non-weighted), Bayesian networks, random forests and support vector machines
Italy (De Fausti et al 2020)	I	administrative information from the ministry of education, university and research, 2011 census data, sample survey data	focussing on one region and on incomplete records, some manual feature selection, calculating and comparing predictions	multi-layer perceptron, random forests, log-linear model
Poland (Wójcik 2020)	I	quarterly sample survey on participation of Polish residents in trips for 2016 to 2018 and some big data sources	learning different models for estimation and comparing their predictions by several measures	different kinds of (generalized) linear models, regression tree, random forest, nearest neighbour, different kinds of support vector machines

The following third table shows some details on the used software and hardware as well as on the metrics used to assess the performance of the compared algorithms.

Country	E/I	Software / Hardware	Measures
Italy (Rocci & Varriale 2020)	E	R no special hardware	usefulness of the results indicating whether a variable determines the presence of a dangerous error in data, accuracy for model selection
UK (Sthamer 2020)	E	Python Intel Core i5-8365U, 1.60GHz, 8 GB RAM	recall, precision, F1
Belgium (Goyens & Buelens 2020)	I	Python Intel i7 CPU with 6 cores, and 32 GB of RAM	RMSE, ME, MAE, MAPE
Germany (Dumpert 2020a)	I	R Intel Core i5-6500, 3.2 GHz, 8 GB RAM	mean, standard deviation, skewness, kurtosis, minimum, maximum, 25 %-quantile, median, 75 %-quantile of the imputed variables, correlations between the variables
Italy (De Fausti et al 2020)	I	Python Azure cloud platform with Tesla V100-PCIE-16GB GPU	micro-level accuracy, macro-level accuracy
Poland (Wójcik 2020)	I	R Intel Core i7-4770, 2x3.40 GHz, 64bit, 16 GB RAM	MAE, MAPE, RMSE, R^2

A fourth table eventually contains the most important aspects of the individual conclusions from the projects.

Country	E/I	Conclusion
Italy (Rocci & Varriale 2020)	E	• the first application of ML methods in this context has shown the possibility to use ML to support the design of an E&I scheme to make it more efficient • exploring hidden patterns in the data with ML tools can help to understand how to classify units in a more efficient way in erroneous/not erroneous in terms of different error types and, therefore, how to combine the different E&I process steps
UK (Sthamer 2020)	E	ML can be used for editing, but some points have to be borne in mind: • a ground truth/gold standard data set for retraining the model has to be created periodically • ML expertise should be within the survey team to monitor and retrain the model when required • editing will be far more efficient and faster with the ML solution compared to existing processes • survey data will be available sooner for further processing and this will allow for more timely data and faster release • it remains open if ML can save cost here, because clerical editing resources have to be maintained as well as technical expertise to build, analyse and keep the ML solution in operation
Belgium (Goyens & Buelens 2020)	I	• think of a baseline method that is simple, common-sensical and reasonably performing • no single ML method worked best • in this study the ensemble method, averaging results from several ML methods, seems promising • manage expectations well; some people expect great results without effort or investment • substantial effort is needed to conduct a proper investigation into the usability of ML methods • making data and code publicly available has been well received by the community and can stimulate future joint work
Germany (Dumpert 2020a)	I	• it is too early to give a general (not survey specific) advice to use one of the investigated methods for imputation • random forest does the imputation faster than the other tested methods in the study • the usage of weighted k-nearest-neighbours and random forest lead to more stable and "correct" estimations of the moments and quantiles; furthermore, the boxplots of these two methods are more symmetric than the other ones

Italy (De Fausti et al 2020)	I	• the results of estimation with the two approaches (MLP vs. log-linear model) are completely comparable • for particular sub-populations, such as extreme items (PhD), log-linear imputation is better • MLP micro accuracy is a bit better respect the log-linear model • MLP approach does not require variables pre-treatment
Poland (Wójcik 2020)	I	• machine learning is much more powerful than traditional models and can easily overfit the data • estimating the out-of-bag error is important to compare various methods by bootstrapping or cross validation • when k-fold cross validation was run several times, it lead to confusion about that which model is the optimal model; bootstrapping seems to be a more reliable method for model selection but at the same time it is more time-consuming • model selection cannot be based just on the accuracy measures like MAPE, RMSE etc. without checking distributional accuracy including biasedness • when data is imputed, it is hard to expect to impute data perfectly on the individual level; it may be expected to retrieve a true mean level of imputed data with respect to some strata; then, on average, totals can be calculated correctly

Retrospective and lessons learnt

Editing

Editing, i. e. the task to find missing and problematic data (e. g. unplausible values, contradictions in records, and so on), is obviously very important in official statistics. Traditionally, rule based comparisons of observed (or transmitted) values with (weak or strong) plausibility constraints, distributional investigations (e. g. for outlier detections), and comparisons with external and/or former data sets are applied. Every editing procedure can be designed in different flows, according to the process features. Several steps are usually considered, in which both automation (through edit rules) and subject matter experts (through interactive editing) play an important role in detecting problematic data. The degree of automation usually depends on the type of errors identified to be most common and from the possibility to detect edit rules that characterize them. However, complete automation should not be the most important goal of the use of machine learning in editing; and it should never be the only goal. Mainly UK's editing pilot study (Sthamer 2020), supplemented by the editing pilot study from Italy (Rocci & Varriale 2020), delivers first insights. The first study from this project (Sthamer 2020) where the aim has been to analyse the capacity of the use of ML to increase the automation of the editing phase as much as possible, i. e. to reduce interactive editing in favour of automation, showed:

1. Learning from former editing results is possible: It is possible to predict whether a unit needs special attention.
2. The extraction of rules suffers from the trade-off that good predictions are only achievable with very detailed (i. e. long and complex) rules.

A second experiment (Rocci & Varriale 2020) has been started (but not yet finished), to assess the use of ML to design a complete new editing process.

According to the study so far, with machine learning the editing process can be completed much faster and more consistently (compared to manual editing). It may possibly even lead to higher quality of the data and allow for much sooner publication, but the effort required maintaining training data, the machine learning model and the analysis of the results in a short term might not proof to be a cost saver. Hence, the gain until now seems to be not so much in efficiency of the results but in the efficiency of the statistical process: Machine learning allows using huge amount of data with much less a priori knowledge, hypotheses and data preparation (general underlying structure of the data, stratification, etc.).

Imputation

Imputation, i. e. the task of altering incorrect values and inserting missing values, is obviously very important in official statistics. From the pilot studies from Poland, Italy, Belgium, and Germany, we have seen:

1. Machine learning delivers comparable (compared to traditional methods) results in a more automated way (e. g., De Fausti et al 2020).
2. Machine learning methods produced often plausible predictions. Nevertheless, in some cases, unplausible predictions appeared (e. g., Wójcik 2020).
3. Machine learning can produce more timely statistics by skipping some pre-treatment (e. g. being aware of correlations, statistical transformations (like logarithm) of the values, technical aspects like dealing with empty cells meaning 0 vs. empty cells meaning that there is a missing, grouping variables, treatment of ordinal and nominal variables, and so on) of variables but there is also the experience that a successful use of machine learning in production is possible only after a lot of (successful) experimentation on the topic (e. g., De Fausti et al 2020, Goyens & Buelens 2020).
4. Machine learning can reduce human intervention (e. g., when it is doing variable selection automatically; e. g. Goyens & Buelens 2020).
5. Imputation projects with time dependencies in the data (like in time series) can be successful (e. g., Goyens & Buelens 2020).
6. It may happen that no single machine learning method works best for a given problem (e. g., Wójcik 2020).
7. Some machine learning methods (or approaches within them) perform better in terms of distributional aspects than other ones (e. g., De Fausti et al 2020, Dumpert 2020a).

From this it was possible to learn that machine learning belongs to the class of methods which are more powerful because of their property that fewer assumptions are needed (in comparison with the fully parametric models); on the other hand, by this, they are flexible enough to be perfect on the training set, but often to perform poorly on unseen data. To avoid this, it is highly recommended to assess the performance of a machine learning model on a separate test set, for example to estimate population parameters based on completed test set. To use machine learning successfully in production is possible only after a lot of (successful) experimentation on the topic of interest; substantial effort is needed to conduct a proper investigation into the usability of machine learning methods. Parametric models are always the best, from every point of view, if the hypothesis is good. Unfortunately, often mistakes in specifying the underlying hypothesis are made, i. e. in modelling the phenomena; hence the parametric model is not able to provide good predictions. Non-parametric models run less risk from this point of view but fit (in the finite data situation) less well than the "true" parametric model. Furthermore, there is a need to shift the interest of stakeholders to accuracy and timeliness of results rather than to the interpretation of the parameters. There are no obvious quick wins to be made, and the uptake of machine learning methods in standard procedures requires substantial and continued effort and commitment. One should also always consider and check against a baseline method that is simple, well accepted, and reasonably performing; this is to avoid drowning in complexities with only marginal effects.

For both, editing and imputation, we have learnt that to apply machine learning methods to statistical processes needs *data science* skills in terms both, programming/coding and statistical training/testing principles. It is also important that subject matter experts are involved. Programmers, statisticians, subject matter experts have to work together intensively, and all of them need some data wrangling skills. This has already been expressed, for example by Cao (2017), who wrote: *"data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including*

domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology."

Conclusion and further recommendations

1. Machine learning and statistical methods have always a serving role in the processes in official statistics. They can assist the subject matter experts and the management in their decisions. For example:
    a. Machine learning and statistical methods may flag an observation as suspicious. The decision whether it has actually to be corrected has to be made and to be accounted for by a subject matter expert.
    b. The choice of the threshold that should be used in a certain classification task has to be made and to be accounted for by a subject matter expert.
2. Applying machine learning needs a bit more data science skills (programming, coding, training/testing principles) than using traditional statistical methods (that are taught at the university in statistics courses).
3. Subject matter experts should always get involved. Usually both sides learn from each other.
4. The usage of machine learning is only useful if it is better (for quality dimensions see work package 2 of the machine learning project) than the currently used baseline method and more simple statistical methods.

References

Buttsworth S. (2020). Imputation of Dwelling Occupancy for Census 2021. https://statswiki.unece.org/download/attachments/266142512/WP1_E%26I_Australia.pdf (last access: 10th Sept. 2020)

Cao L. (2017). Data science: a comprehensive overview. ACM Computing Surveys, 50(3), 1–42.

Chambers R. (2001). Evaluation Criteria for Statistical Editing and Imputation.

De Fausti F., Di Zio M., Filippini R., Toti S., & Zardetto D. (2020). Imputation of the variable "Attained Level of Education" in Base Register of Individuals. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Dumpert F. (2020a). Machine learning for imputation. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Dumpert F. (2020b). The UNECE High-Level-Group for the Modernization of Official Statistics Machine Learning Project: A report of the Editing & Imputation Group. Paper presented at the UNECE Statistical Data Editing Virtual Workshop 2020. https://statswiki.unece.org/download/attachments/282329136/SDE2020_T1-B_Germany_Dumpert_Paper.pdf (last access: 10th Sept. 2020)

Goyens A. & Buelens B. (2020). Early estimates of energy balance statistics using machine learning. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Rocci F. (2020). Machine Learning for Data Editing Cleaning in NSI (Editing & Imputation): Some ideas and hints. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Rocci F. & Varriale R. (2020). Machine Learning tool for editing in the Italian Register of the Public Administration, a proposal. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Ruiz C. (2018). Improving Data Validation using Machine Learning. Paper presented at the UNECE Statistical Data Editing Workshop 2018. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUIZ_Paper.pdf (last access: 10th Sept. 2020)

Sthamer C. (2020). Editing of LCF (Living Cost and Food) Survey Income data with Machine Learning. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Wójcik S. (2020). Imputation in the sample survey on participation of Polish residents in trips. UNECE HLG-MOS Machine Learning Project, https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home