Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(intended: Geneva, Switzerland, 15–17 April 2020; actually: virtual, 31st August to 4th September 2020)

The UNECE High-Level-Group for the Modernization of Official Statistics Machine Learning Project: A report of the Editing & Imputation Group

Prepared by Florian Dumpert, Federal Statistical Office, Germany

I. Introduction

1. At the end of 2018, the UNECE High-Level-Group for the Modernization of Official Statistics (UNECE HLG-MOS) decided to launch a project on machine learning. One of its work packages has the goal to investigate the value added of machine learning approaches in, among other themes, editing and imputation. Several countries take part in the editing & imputation group, some of them present their work in more detail on this workshop. Two main tasks are considered: editing as identifying suspicious values in datasets, and imputation as filling in missing or during the editing process deleted values. For imputation, some machine learning approaches are well studied in the literature. In contrast, only very few approaches are known for editing.

2. National statistical organisations (NSOs) of the following countries are involved in the editing & imputation group: Australia (ABS), Belgium (VITO), Italy (ISTAT), Poland (Statistics Poland), Switzerland (BFS), United Kingdom (ONS), and Germany (DESTATIS). Besides editing and imputation, there are groups working on classification and coding, on imagery, on web analysis, on quality aspects, and on integration of machine learning into the statistical processes. The team of the whole machine learning project consists (at the moment, however still increasing) of 39 participants from 14 countries and 19 organisations. They collaborate on the research and application of machine learning techniques to improve the production of official statistics. The project also includes sharing of knowledge and experience in machine learning techniques.

3. Since the machine learning project being launched in 2019, over 40 documents (presentations, working papers, machine learning scripts, software documentation) have been shared within the team. Most of the work was discussed at a sprint held in September 2019 and now continues to be shared and discussed at monthly meetings.

4. A report on the work of the project in 2019 is available online (Julien 2019). Furthermore, presentations that were given at a virtual sprint meeting in April 2020 are available on https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home, a preliminary version of the project documentation webpage.

II. Machine learning for editing and imputation

A. Preliminary considerations

5. It is uncontroversial that there is a need for NSOs to identify and deal with suspicious values in datasets. It is 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 not be – in some sense – belonging to the rest of data at hand. Domain knowledge is usually a necessary component of editing.

6. 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 NSOs (by replacing, improving or complementing methods used so far).

B. Machine learning

7. Machine learning in this paper should not be seen as a collection of methods. Most of the methods commonly “said to be machine learning” were invented decades ago. To our best knowledge, there is no widely accepted definition of *machine learning* at all. We refer to Breiman (2001), Shmueli (2010), and the recent paper by Efron (2020) when we define machine learning by the goal we want to achieve: If the focus of our work is on the prediction of output values (given input values) rather than on explaining which relationships lead to these outputs in nature, we do machine learning. So machine learning is statistics but with a special focus. In particular, the basic statistical reflections everyone is asked to do when working with data should not be forgotten.

8. Historically, there are complementing ways to describe machine learning. For example, Samuel (1959) wrote: *The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning. [...] Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.*

9. Motivated by the frequently profitable use of machine learning approaches in questions of classification (including coding) and regression also in official statistics (see e.g. Beck et al. 2018), it should also be investigated if they might be usefully used as a component of editing and imputation steps.

10. While there is only little scientific literature for machine learning for editing, machine learning for imputation has already been studied by the scientific community (e.g. for an overview Richman et al. 2009, Mikhchi et al. 2016). More specific literature for certain methods (for example for tree-based methods) is available, too.

C. Pilot studies

11. Currently, there are two pilot studies on machine learning for editing and four pilot studies on machine learning for imputation run within the HLG-MOS Machine Learning Project. These are for editing:

(a) United Kingdom (ONS): Editing of LCF (Living Cost and Food) survey data.
(b) Italy (ISTAT): Validation of administrative data sources during the integration phase of the statistical register.

For imputation:

(a) Belgium (VITO): Imputation in the energy balance of Flanders.
(b) Italy (ISTAT): Imputation of the variable “attained level of education” in the base register of individuals.
(c) Poland (Statistics Poland): Imputation of tourist expenditures and imputation of sports clubs.
(d) Germany (DESTATIS): Simulation study on machine learning for imputation.

In addition, Switzerland (BFS) made their project results of Plausi++ (Ruiz 2018) available to the editing and imputation group; Australia (ABS) also presented additional work during several work sessions of the editing and imputation group. Some of the pilot studies present their work in separate papers at this workshop.

12. Goal of these pilot studies is to show whether machine learning offers a value added to the processes in NSOs. The final reports of the pilot studies as well as a final theme report for editing and imputation are currently under revision by the project team members and the project manager. Nevertheless, first insights can already be communicated. At the very beginning of the project, the experiences in machine learning for editing or imputation were not sound, in particular in machine learning for editing. Some experiments had already been done in order to get first impressions. Frequently, standard techniques from "traditional statistics" were –often successfully – used so far for the editing and imputation tasks of interest. Some months later, promising as well as ambiguous as well as surprising results could be observed. As an omnipresent and on-going process, the participating NSOs learn from each other by exchanging ideas, knowledge, presentations, and reference papers.

13. Machine learning showed some positive aspects in the pilot studies: It is more powerful than other statistical approaches because of its property that there are fewer assumptions (e. g. in comparison with the fully parametric models) and – as a consequence – machine learning can deliver comparable (compared to other statistical approaches) results in a more automated way and can reduce formerly necessary human interaction. Doing this, machine learning can help to produce more timely statistics. Often, machine learning produced plausible predictions (i. e. it imputed values that did not break plausibility rules). It has also been shown that machine learning also works when time dependencies are included in the data (like in time series). Furthermore, it is possible to learn from former editing results (and then, e. g., to predict whether a new observation coming in needs special attention). Hence there is the potential to complete the editing process much faster and more consistently.

14. There were also some issues that appeared during the work on the pilot studies. From a methodological point of view, questions about the identification of key predictor features, text written in non-latin alphabets, unbalanced data, low quality of training and test data, and incomplete data sets were challenging aspects. IT infrastructure is a bottleneck for a lot of machine learning applications in NSOs. Legal aspects prohibit in some of the participating countries the use of cloud computing or other external solutions. From an integration perspective, the situation that it is often not clear what machine learning is, what can be done with it and what cannot, the fear of losing control on the statistical processes and the quality of the results, the often low interpretability of machine learning approaches (and the resulting mistrust) but also the sometimes lengthy process of building up the necessary methodological knowledge became visible during the project.

15. The preliminary overall assessment is that machine learning is suitable for editing and imputation in official statistics in principle but further investigations on the quality of the processes and the results have to be conducted. However, a positive tendency can be stated.

D. Example: Simulation study on machine learning for imputation

16. *Imputation is a method for the analysis of data with missing values, where missing values are replaced by estimates and the filled-in data are analysed by complete-data methods. […] In fact, the main reason for imputation is not to recover the information in the missing values, which is lost and usually not recoverable, but rather to allow the information in observed values in the incomplete cases to be retained.* (Little, 2011) Who would dare to contradict R.J. Little? Affirmatively, one of the authors of the missForest imputation package in R, D. J. Stekhoven, writes in a paper on his package: *However, it should always be kept in mind that imputing data with missing values does not increase the information contained within this data. It is only a way to have completeness for further data analysis.* (Stekhoven 2012).

17. Imputation is well established in traditional statistics, in particular if the downstream tasks are already known, see Little & Rubin (2002). Keeping this in mind on the one hand side, and knowing that it is often needed to impute without knowing the downstream task on the other hand side, DESTATIS

started to have a closer look at the question whether machine learning approaches are suitable for imputation tasks.

18. According to the Euredit project, there is a list of imputation goals (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.

19. There are different imputation approaches shown in literature and practice which should not be discussed in this short paper but they are mentioned to give an overview.

- feature-wise imputation of location parameters
- feature-wise imputation of conditional location parameters
- regression based imputation
- stochastic regression based imputation (prediction + random error)
- similarity based imputation (e.g. via nearest neighbour, CART, predictive mean matching)

It is well known that the first three approaches tend – by construction of the approach – to reduce or to underestimate the variance of a feature. Therefore, as a compensation, one may, e. g., add random errors (found e.g. via sampling from the residuals of the regression or from a data dependent distribution).

20. The simulation study included (so far) only regression based imputation of continuous variables with different methods like nearest neighbour regression, random forest regression, support vector machines and Bayesian networks. Different missing rates as well as different missing mechanisms have been simulated. The result was surprising: weighted nearest neighbour regression and random forests (in contrast to the other methods) did (at least for moderate missing rates) not suffer too much from the often observed problem in non-stochastic imputation that the distribution after imputation is not the same as in the complete original data (which is available in simulation studies). In particular, variance, skewness, and kurtosis of the imputed data set were close to the original ones. This observation motivated to start further investigations and a cross-border cooperation with CBS where – independently – similar results were found (Park et al. 2018).

21. A great limitation of the study mentioned above is of course that only one survey was used; this has to be extended in the future. However, also if the result, that weighted nearest neighbour and random forest perform successfully in naive but very fast regression imputation, is stable (and it seems to be stable from today's perspective), there is, so far, no theoretical justification for these experiences. It is therefore too early to give an unreserved advice to use one of these methods for (regression) imputation.
A more detailed report on this study will be available as part of the UNECE HLG-MOS Machine Learning Project Documentation.

III. Outlook

22. The HLG-MOS Machine Learning Project plans to deliver – among others – the following outputs in 2020:

(i) A report on each pilot study. These reports will describe the pilot study, its value proposition to the organisation, its technical details, its results in demonstrating the value-added of machine learning, accompanying actions or challenges in advancing to use of machine learning in the organisation, and future work.

(ii) A report on each pilot study theme. These reports will summarize the value-added and best practices in using machine learning in the theme (classification and coding, editing and imputation and imagery), advances in the implementation in organisations and recommend future work.

(iii) A report on the integration of machine learning in the production process. This report will summarize the value-added of machine learning for NSOs, and best practices in developing and maintaining them in production. The report would be accompanied by a draft quality framework on key aspects in the use of machine learning and a report on how different NSOs are organized to integrate machine learning in their production processes, sources of impediments and successful practices.

Besides these, established mutual collaborations are supposed to be continued even after the formal end of the HLG-MOS Machine Learning Project.

References

Beck, M., Dumpert, F., & Feuerhake, J. (2018). Machine Learning in Official Statistics. Online available: https://arxiv.org/abs/1812.10422

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

Chambers, R. (2001). Evaluation Criteria for Statistical Editing and Imputation. Online available on: https://www.cs.york.ac.uk/euredit/

Efron, B. (2020). Prediction, Estimation, and Attribution. Journal of the American Statistical Association, 115, 636–655.

Julien, C. (2019). Background document on the Machine Learning Project – Prepared for the 2019 Workshop on the Modernisation of Official Statistics. Online available: https://statswiki.unece.org/download/attachments/256970769/HLG-MOS%20November%20workshop%20ML%20project%20background%20document.docx?version=2&modificationDate=1573770390702&api=v2

Little, R. J. (2011). Imputation. In: Lovric, M. (2011). International Encyclopedia of Statistical Science. Springer.

Little, R. J. & Rubin, D. B. (2002). Statistical analysis with missing data, 2nd ed. Wiley.

Mikhchi A., Honarvar M., Kashan N. E. J., & Aminafshar, M. (2016). Assessing and comparison of different machine learning methods in parent-offspring trios for genotype imputation. Journal of theoretical biology, 399, 148–158.

Park, S., Pannekoek, J., & van der Loo, M. P. J. (2018). Imputation of Economic Data based on Random Forest. Technical Report. Available upon request.

Richman M. B., Trafalis T. B., & Adrianto I. (2009). Missing data imputation through machine learning algorithms. In Artificial Intelligence Methods in the Environmental Sciences (pp. 153–169). Springer.

Ruiz, C. (2018). Improving Data Validation using Machine Learning. Online available: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUIZ_Paper.pdf

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3, 210–229.

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

Stekhoven, D. J. (2012). Using the missForest Package. Online available on: https://stat.ethz.ch/education/semesters/ss2012/ams/paper/missForest_1.2.pdf