Machine learning for imputation

Organisation: Federal Statistical Office of Germany (Destatis)

Author(s): Florian Dumpert[*1]

Date: 27.07.2020

Version: 3.0

1. Background and why and how this study was initiated

Background for this study is the need to use statistical methods for imputation to achieve results from incomplete data which then can be analysed. Pioneer work has been done by Rubin (1987) und Little & Rubin (1987). A short overview on motivation and approaches is for example given by Little (2011). The omnipresent success of machine learning in the field of regression and classification leads to the question whether these methods are suitable for imputation tasks as well. It is possible to argue that imputation is prediction because missing values are replaced by estimates. It is also possible to explicitly see a difference, see, e. g., Bertsimas et al. (2017). It might be only a question of semantics. Nevertheless,

> *"[i]mputation is a method that allows incomplete cases to be included in the analysis. 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. If there is little information to be recovered in the cases with missing values, as for example cases in regression for which the outcome variable is missing, then imputation is not very useful."*
>
> – Little (2011), p. 651

Prima facie, it looks as if machine learning of all things is unsuitable as it often has the goal to provide good (point) predictions (Breiman 2001) which is – referring to the quotation – explicitly not the aim of imputation. On the other hand, if the true values would be found (which is obviously only checkable in the perfect world of simulations) the data set would be as perfect as if there would never have been a problem with missing values. Consequentially, the EUREDIT project (Chambers 2001) mentions different goals of imputation in official statistics: Predictive accuracy[2], ranking accuracy[3], distributional accuracy[4], estimation accuracy[5], and, where required, an overall goal in imputation tasks: plausibility[6].

Is machine learning suitable for imputation tasks? In order to find an answer, the section for machine learning and imputation in the division for mathematical-statistical methods at Destatis was tasked to run a first small simulation study (Dumpert et al. 2018). This report summarizes the study and the developments since then. The study was – more or less – a "neighbour" and follow-up of a proof of concept Destatis ran in 2018 (Beck et al. 2018). The overall basis for working on machine learning is the "digital agenda" of Destatis, a strategy paper concerning the digital transformation including electronic data management, digital workflows, automation of process steps, and

[*] Federal Statistical Office of Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany

development of new digital services. The “political” support for the topic machine learning in general is given by the senior management of Destatis.

Beck M., Dumpert F., & Feuerhake J. (2018). Proof of Concept Machine Learning – Abschlussbericht. Online available on: https://www.destatis.de/GPStatistik/receive/DEMonografie_monografie_00004835 (in German)
Shorter English version available on arXiv: https://arxiv.org/abs/1812.10422

Bertsimas D., Pawlowski C., & Zhuo Y. D. (2017). From predictive methods to missing data imputation: an optimization approach. The Journal of Machine Learning Research, 18(1), 7133–7171.

Breiman L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199–231.

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

Dumpert F., Hansen M., Peters F., & Spies L. (2018). Bericht zur Maßnahme Machine Learning Methodik. Internal Paper, yet unpublished, in German.

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

Little R. J. & Rubin D. B. (1987; 2002). Statistical analysis with missing data. Wiley.

Rubin D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley.

2. Data

2.1 Input Data (short description)

To get an impression on how machine learning works in an imputation task, a small simulation study was designed. The study has been run on the German cost structure survey of enterprises in manufacturing, mining and quarrying[7] which provided 15,646 records. All of these records were used.
Missing values were artificially created in two[8] of its 54 substantial variables by two different missing mechanisms: missing completely at random (MCAR) and missing at random (MAR). Furthermore different missing rates were induced (5 %, 10 %, 20 %). Both variables are continuous and refer to internal research and development of the enterprises.

2.2 Data Preparation

We ran the simulation study without any special data preparation step. The data set was interpreted as “gold standard”.

2.3 Feature Selection

No explicit feature selection was done before the simulation study started. We used all available variables.

2.4 Output data

Complete datasets (imputed data).

3. Machine Learning Solution

3.1 Models tried

The investigation included k-nearest-neighbours (weighted and unweighted)[9], Bayesian Networks[10], Random Forest[11] and Support Vector Machines[12]. These methods were tried because of their already shown success in imputation tasks outside official statistics (see, e. g., van Buuren 2018, Richman et al. 2009, Mikchi et al. 2016, Honghau et al. 2005, Yang et al. 2013).

The models were trained on the 52 complete explaining variables in order to estimate the missing values of the two target variables. After having created the artificial missing values (cf. Section 2.1), one of the target variables was treated as Y-variable and the 52 complete explaining independent variables were treated as X-variables. The goal now was to train several models Y~(X1, ..., X52), i. e. models where Y depends on X1, X2, ..., and X52. In this step, i. e. for learning the models, we only used those records where Y was not missing. After that, the same was done for the second target variable. Note that the investigations for the two target variables were conducted disjointly, so no interactions between the two target variables have been considered. The trained models were then used to predict the missing values. As the missing values were created artificially, it was possible to compare the real and the imputed version of the data set, in particular the entries of the two target variables (see Section 4).

The hyperparameters (e. g. k in k-nearest-neighbours, mtry in Random Forest, gamma and C in Gaussian kernel based Support Vector Machines) were determined by either simply learning and evaluating models for different possible values of the hyperparameters (as it was the case for k-nearest-neighbours) or with a grid search in combination with five bootstrap evaluations on the training data (as it was the case for SVMs). As we performed only regression imputation, hyperparameters yielding small RMSEs were selected in the tuning step.

Honghai F., Guoshun C., Cheng Y., Bingru Y., & Yumei C. (2005). A SVM regression based approach to filling in missing values. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 581–587).

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.

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).

van Buuren S. (2018). Flexible Imputation of Missing Data. 2nd edition. CRC.

Yang B., Janssens D., Ruan D., Bellemans T. & Wets G. (2013). A data imputation method with support vector machines for activity-based transportation models. In Computational Intelligence for Traffic and Mobility (pp. 159–171).

3.2 Model(s) finally selected and the criterion

In the study, no model had to be chosen because it was a first comparative study in order to see whether machine learning yields sufficient results also in imputation tasks. From our point of view, weighted k-nearest-neighbours and Random Forest performed best in terms of conserving the distribution or at least the first moments of the original data which was the minimal goal that should be achieved in this study. However, the number of situations where implausible values were imputed is considerably higher for Random Forest compared to

weighted k-nearest-neighbours. This fact has to be addressed before using Random Forest in production. Some details are shown in Section 4.

3.3 Hardware used

Intel Core i5-6500, 3.2 GHz, 8 GB RAM.

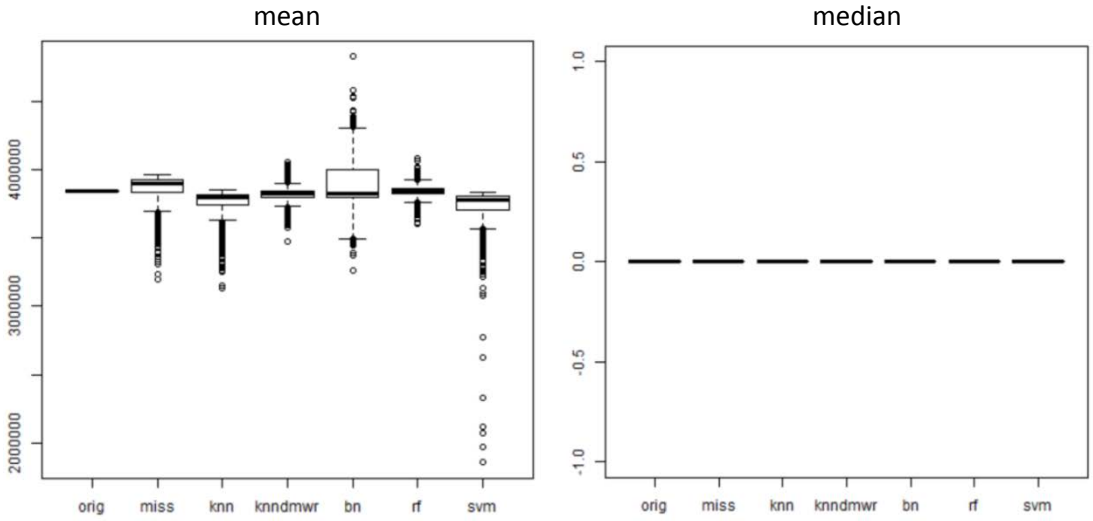
3.4 Runtime to train the model

We dealt with around 15,500 samples and 52 independent features. Per one of the 1,000 runs, per method, per missing mechanism, and per proportion of missing data, it took between 25 seconds (Random Forest), 58 seconds (weighted k-nearest-neighbours), and 144 seconds (SVMs) to learn the model and to calculate the imputed data set.

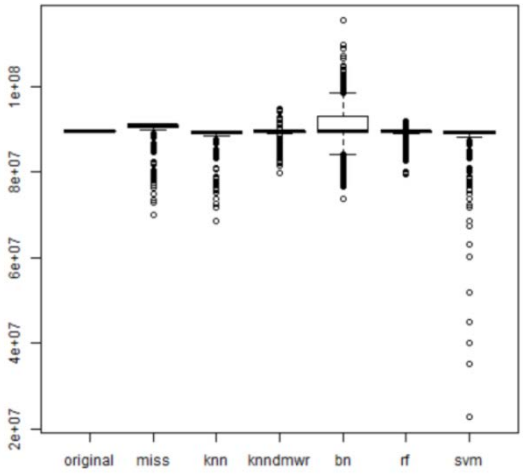
4. Results

To evaluate the success of a certain method, we compared the distribution of the imputed data set with the distribution of the original complete data set in terms of univariate statistics like mean, standard deviation, skewness, kurtosis, minimum, maximum, 25 % quantile, median, 75 % quantile; and in terms of bivariate statistics (correlations). The following plots show several results for one of the variables to impute (expenditures for internal research and development) based on 1,000 runs per scenario. We show the results for the "best" and the "worst" situation (missing pattern, proportion of missing values) we have had a look at. In each run, missing values were created randomly according to a missing mechanism. In the following plots, *orig* stands for the original complete data set, *miss* stands for the data set with simulated missing values, *knn*, *knndmwr*, *bn*, *rf*, and *svm* stands for the data set after imputation with k-nearest-neighbours, weighted k-nearest-neighbours, Bayesian Networks, Random Forest, and Support Vector Machines, respectively.

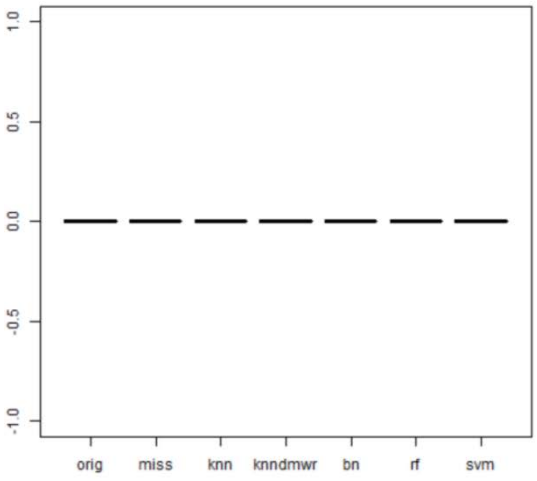
(a) Results for the situation that we have MCAR and only 5 % missing values



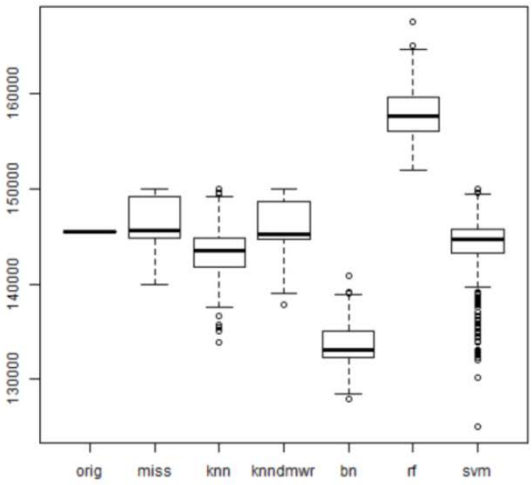
standard deviation



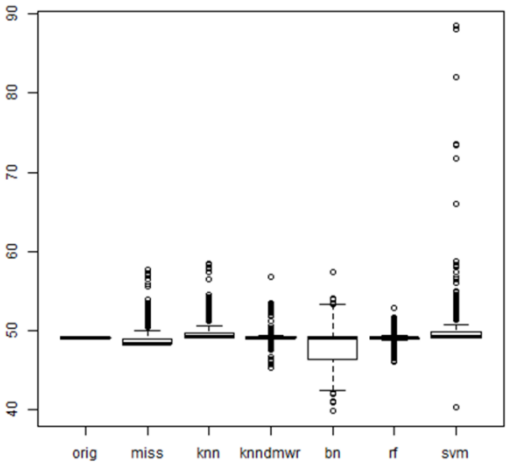
25 % quantile



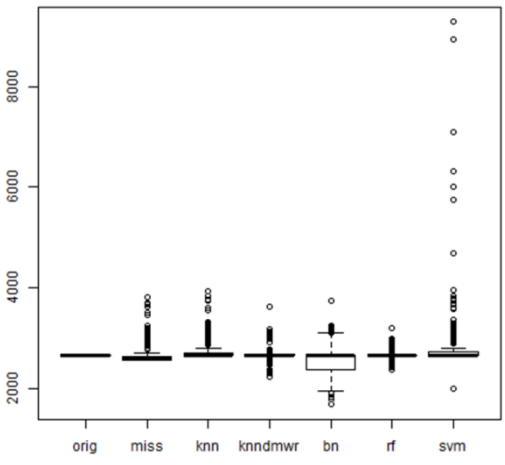
75 % quantile



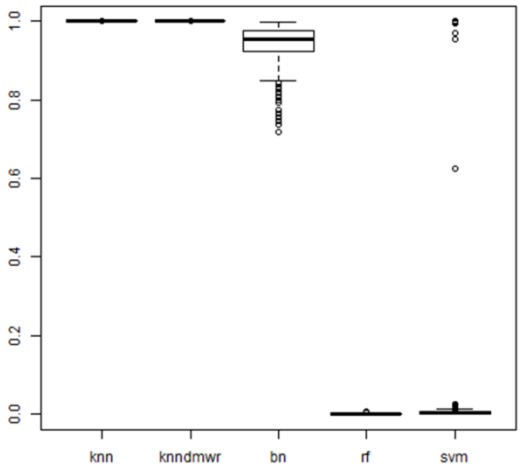
skewness



kurtosis



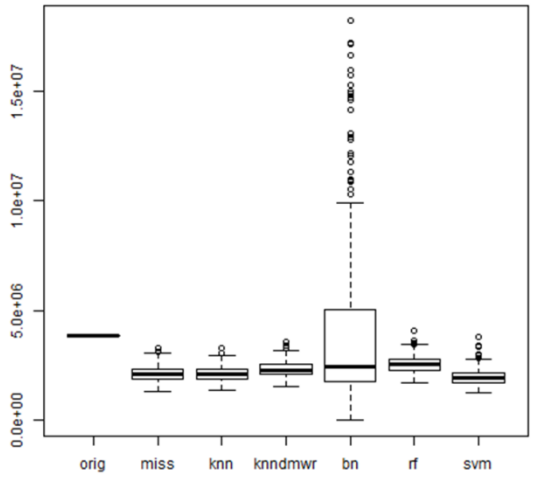
p-value of the Kolmogorov Smirnov test[13]



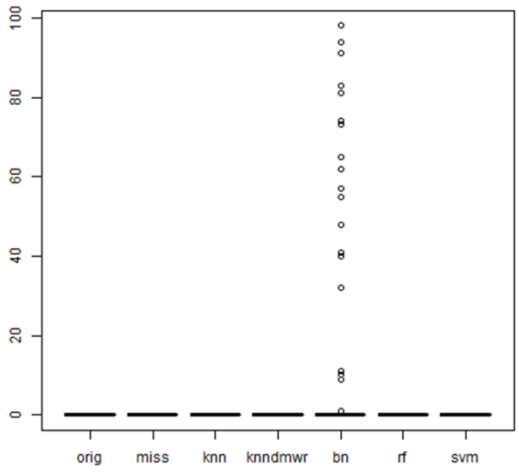
The graphics show that the usage of weighted k-nearest-neighbours (knndmwr) and Random Forest (rf) lead to more stable and "correct" estimations of the moments and quantiles. Furthermore, the boxplots of these two methods are more symmetric than the other ones. Only the third quartile is overestimated by Random Forest which leads to a p-value of 0 of the Kolmogorov Smirnov test.

(b) Results for the situation with MAR and 20 % missing values

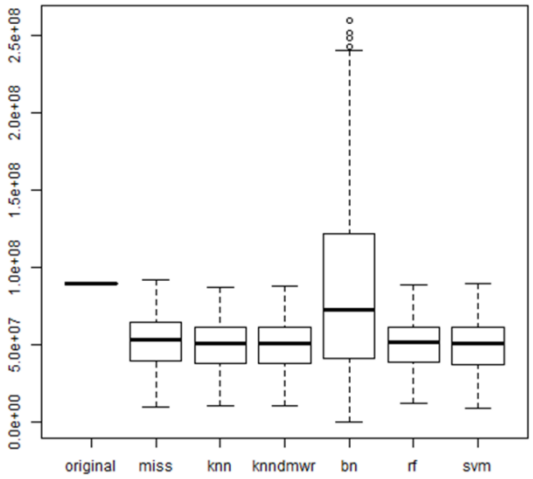
mean



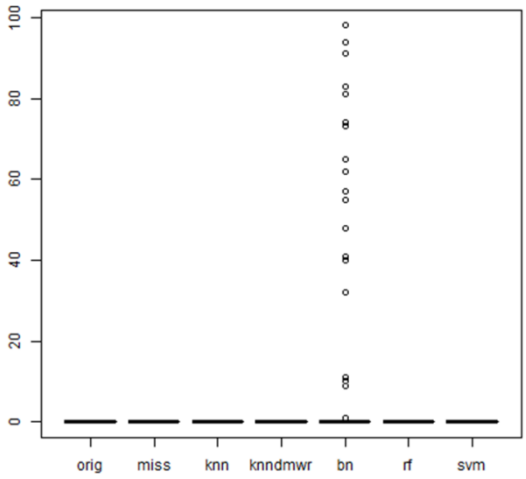
median



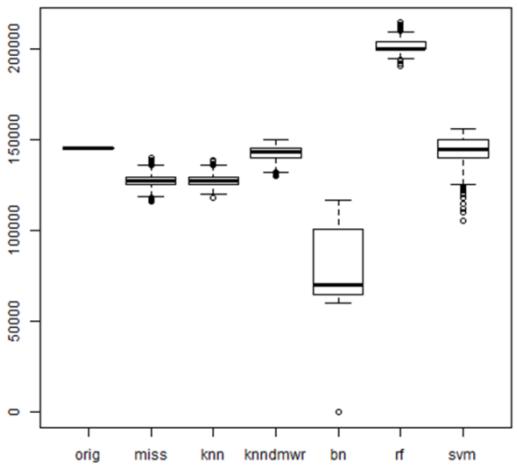
standard deviation



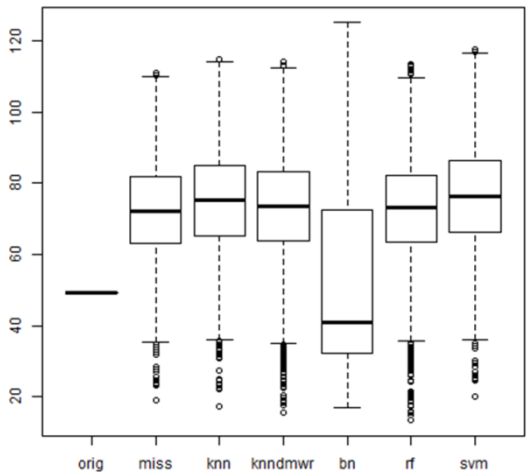
25 % quantile



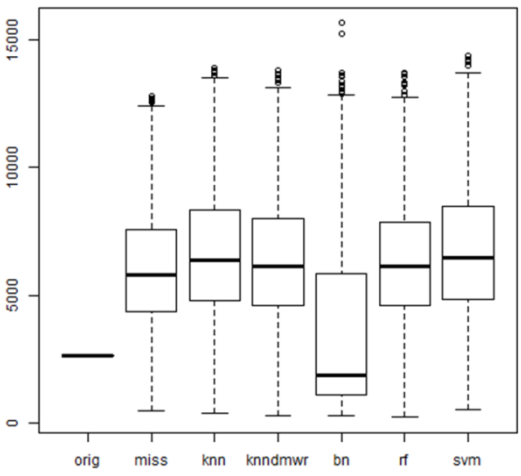
75 % quantile

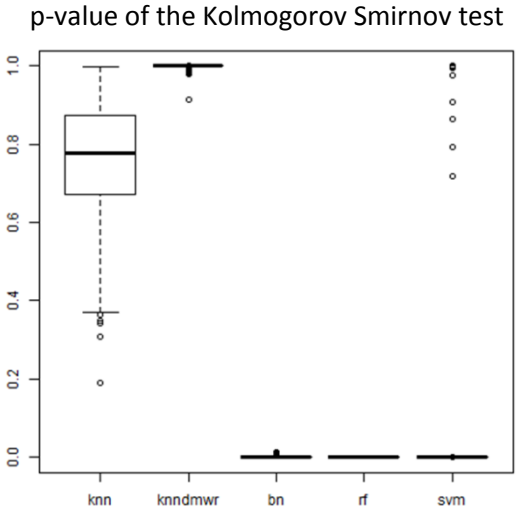


skewness



kurtosis





In the situation of the worst investigated situation (MAR and 20 % missing values), we observed that many methods have problems to "retain" the original distribution. Mean and standard deviation are usually underestimated; skewness and kurtosis are often overestimated. The results for the quantiles are ambiguous. Again, weighted k-nearest-neighbours leads to p-values close to one, Random Forest and the other methods to p-values close to 0.

5. Code/programming language

For the study, it was not necessary to implement specific methods. Instead, four for-loops (missing mechanism, missing rate, method, run per method) and standard implementations in the R packages yaImpute (for k-nearest-neighbours; Crookston & Finley 2007), dmwr (weighted k-nearest-neighbours; Torgo 2010), ranger (Random Forest; Wright & Ziegler 2017), bnlearn (Bayesian Networks; Scutari 2010), and liquidSVM (SVMs; Steinwart & Thomann 2017) were used. In addition, the R packages mice (creating the missing values; van Buuren & Groothuis-Oudshoorn 2011), Metrics (evaluation metrics; Hamner et al. 2018), and stats (some basic evaluation criteria) were applied. All these packages are available on https://cran.r-project.org/.

Crookston N. L. & Finley A. O. (2007). yaImpute: An R Package for kNN Imputation. Journal of Statistical Software, 23(10), 1–16.

Hamner B., Frasco M., & LeDell E. (2018). Metrics: Evaluation Metrics for Machine Learning. Online: https://CRAN.R-project.org/package=Metrics.

Scutari M. (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1–22.

Steinwart I. & Thomann P. (2017). liquidSVM: A Fast and Versatile SVM package. Online: https://arxiv.org/abs/1702.06899.

Torgo L. (2010). Data Mining with R, learning with case studies Chapman and Hall/CRC. Online: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

van Buuren S. & Groothuis-Oudshoorn K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67.

Wright M. N. & Ziegler A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17.

6. Evolution of this study inside the organisation

The study resulted in further investigations which in turn led to the result that the package missForest (Stekhoven & Buehlmann 2012), a Random Forest implementation for imputation, or a more efficient one based on missForest, like e. g. missRanger (Mayer 2019), is going to get implemented parallel to CANCEIS (an editing and imputation system developed by Statistics Canada which is based on k-nearest-neighbours; Bankier et al. 2000) in the editing and imputation step of the new version of the German earnings survey. Starting in 2022, the new version of the German earnings survey will be carried out on a monthly basis. It contains individual employee data (such as earnings or number of hours worked, but also personal data like age or education) and links this data to information on the company.[14] The need for a more automated editing and imputation step German earnings survey results from the high number of records (around 7 million) that will be submitted to the statistical offices each month. For the next years, it is planned to use both approaches in a parallel manner and to compare them in order to be able to decide[15] which one is more suitable in this particular survey. From this point of view, this study broadened potential use cases for ML within Destatis.

Bankier M., Lachance M., & Poirier P. (2000). 2001 Canadian census minimum change donor imputation methodology. UNECE Work Session on Statistical Data Editing 2000, Working Paper No. 17. Online: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2000/10/sde/17.e.pdf

Mayer M. (2019). missRanger: Fast Imputation of Missing Values. Online: https://cran.r-project.org/web/packages/missRanger/index.html

Stekhoven D. J. & Buehlmann P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118.

7. Is it a proof of concept or is it already used in production?

This study was a proof of concept. It was successful because the general usability of machine learning for imputation has been shown. For further developments and future usage see Section 6.

7.1 What is now doable which was not doable before?

Random Forest does the imputation faster than the other tested methods in the study (cf. runtimes in Section 3.4). Further investigations during the “probation period” mentioned in Section 6 will detail the different behaviour of Random Forest and CANCEIS.

7.2 Is there already a roadmap/service journey available how to implement this?

See Section 6.

7.3 Who are the stakeholders?

Internal subject matter statistics (as the direct “user”), the IT division (that provides the technical infrastructure), the mathematical-statistical methods division (development and investigation of methods, methodological knowledge), controlling (embedding), the senior management of Destatis (support).

7.4 Fall Back

See Section 6. In that sense, CANCEIS is a fall back solution for the missForest-based approach and vice versa.

7.5 Robustness

The results of the automated imputation by CANCEIS and a missForest-based approach will be compared to each other in the afore mentioned "probation period" during the next years and – moreover – randomly checked regarding their plausibility (beyond the obligatory editing rules) by subject matter experts. The latter one has to be done in order to ensure the quality of the editing & imputation process because – obviously – there is no gold standard (comprising true values) available for new data. See also Sections 6 and 7.4.

8. Conclusions and lessons learned

A great limitation of the study presented in this paper is of course that only one survey has been used; this has to be extended in the future (see Sections 6 and 10). However, also if the result, that weighted k-nearest-neighbours and Random Forest perform successfully in naïve but very fast regression imputation (remember that we performed only regression imputation), is stable (and it seems to be stable from today's perspective), we do not have a theoretical justification yet. It is therefore too early to give the general (not survey specific) advice to use one of these methods for imputation.
Further investigations (see Sections 10 and 11) are mandatory.

9. Potential organisation risk if ML solution not implemented

None (cf. 7.4). However, to be able to accelerate the editing and imputation processes, to allow for more automation, and to have enough time to have a closer manual look at "the very important cases" in surveys, machine learning solutions (including CANCEIS) are needed.

10. Has there been collaboration with other NSIs, universities, etc?

Yes. (i) With Statistics Netherlands (CBS), we plan an extended simulation study on official data sets of at least The Netherlands and Germany. (ii) With the Technical University of Dortmund, first further investigations, also from a more theoretical point of view, are already ongoing.

11. Next Steps

To get started, we wrote the R code for our setting and pure regression based imputation (i. e. without adding errors or doing predictive mean matching). However, the results were surprising: weighted k-nearest-neighbours and Random Forests seem to be able to do good imputation in terms of distributional accuracy (and also in terms of RMSE) although we did not compensate the expected underestimation of the variance explicitly. Actually, we do not know why this happened in this situation; we cannot explain this phenomenon from a theoretical point of view. Interestingly, CBS found comparable results for Random Forest independently (Park et al. 2018).

Motivated by these results we started to look for universities that would be interested in our findings and that would like to further investigate theoretical and further practical aspects of them. Furthermore, we plan to extend the simulation study in cooperation with CBS (cf. Section 10).

Challenges have mainly been based on IT bottlenecks. Some kinds of machine learning (e. g. SVMs) need a lot of memory and time to get computed. Though some methods speed up remarkably by using parallelization this is only useful if we could use computers with many kernels. Destatis is currently augmenting its IT infrastructure due to this finding.

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

Supplementary notes:

[1] The author would like to thank Lydia Spies, Fabian Peters, and Malte Hansen, who have contributed to the presented study to at least as great an extent as he himself. Furthermore, Jörg Feuerhake deserves great thanks for his critical review of the document and his valuable suggestions for improvement.

[2] The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are as "close" as possible to the true values.

[3] 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.

[4] 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.

[5] 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).

[6] 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.

[7] In terms of content, the cost structure survey provides comprehensive information on the production results, the production factors used and the value added in its various gradations; it is thus the most important starting point for all structural studies in the fields of politics, enterprises and their associations and economics. The main users of the cost structure survey are the federal ministries, in particular the Ministry of Economics and Technology, the European Commission, the national accounts of the Federation and the Länder. In addition, research institutes, trade associations and the respondents themselves are important stakeholders in the statistical results. To support the scientific work, the results of the cost structure survey are made available in anonymous form via the Research Data Centre of the Statistical Offices of Scientific Research for extended use, e. g. in the context of micro data analysis. The data is collected in electronic form and by means of a postal survey. There is an obligation to provide information. The owners or managers of the legal units involved are obliged to provide information. The obligation to provide information ensures a high response rate and thus increases the accuracy of the results. The sample is a single-stage stratified random sample.
The substantial variables contain information on where in Germany the enterprise is located, on the number of employees (full time, part time), legal form, economic activity, turnover, stock, energy consumption, direct and indirect wage costs, social-insurance contributions, costs for lease and rent, tax, depreciations, interests, the expenditures for internal research and development, and the number of employees working for internal research and development.
More detailed information on the German cost structure survey of enterprises in manufacturing, mining and quarrying can be found here: https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Industrie-Verarbeitendes-Gewerbe/kostenstruktur-verarbeitendes-gewerbe.pdf (in German).

[8] the expenditures for internal research and development and the number of employees working for internal research and development

[9] Some references for k-nearest-neighbour approaches are given by:

Beretta L. & Santaniello A. (2016). Nearest Neighbor Imputation Algorithms: A Critical Evalutation. Medical Informatics and Decision Making, 16, 197–208.

Cucala L., Marin J. M., Robert C. P., & Titterington D. M. (2009). A Bayesian Reassessment of Nearest-Neighbor Classification. Journal of the American Statistical Association, 104, 263–273.

Devroye L., Györfi L., & Lugosi G. (1996). A Probabilistic Theory of Pattern Recognition. Springer.

Liao S. G., Lin Y., Kang D. D., Chandra D., Bon J., Kaminski N., Sciurba F. C., & Tseng G. C. (2014). Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or not, and how? Bioinformatics, 15, 346.

Troyanskaya O., Cantor M., Sherlock G., Brown P. O., Hastie T., Tibshirani R., Botstein D., & Altman R. B. (2001). Missing Value Estimation Methods for DNA Microarrays. Bioinformatics, 17, 520–525.

[10] Some references for Bayesian Networks are given by:

Cheng J., Greiner R., Kelly J., Bell D. A., & Liu W. (2002). Learning Bayesian Networks from Data: An Information-Theory Based Approach. Artificial Intelligence, 137, 43–90.

Di Zio M., Sacco G., Scanu M., & Vicard P. (2004). Multivariate Techniques for Imputation Based on Bayesian Networks. Compstat 2004 Symposium.

Di Zio M., Scanu M., Coppola L., Luzi O., & Ponti A. (2004). Bayesian Networks for Imputation. Journal of the Royal Statistical Society Series A, 167(2), 309–322.

Kalisch M., Bühlmann P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research, 8, 613–636.

Jensen F. V. & Nielsen T. D. (2007). Bayesian Networks and Decision Graphs. Second edition. Springer.

Lauritzen S. L. (1995). The EM Algorithm for Graphical Association Models With Missing Data. Computational Statistics and Data Analysis, 19, 191–201.

Moore A. & Wong W. (2003). Optimal Reinsertion: A New Search Operator for Accelerated and More Accurate Bayesian Network Structure Learning. In Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), 552–559.

Rey del Castillo P. (2012). Use of Machine Learning Methods to Impute Categorical Data. Conference of European Statisticians WP. 37.

Riggelsen C. (2006). Learning parameters of Bayesian networks from incomplete data via importance sampling. International Journal of Approximate Reasoning, 42(1-2), 69–83.

Spirtes P., Glymour C., & Scheines R. (2000). Causation, prediction, and search. Second edition. MIT Press.

Tsamardinos I., Brown L. E., & Aliferis C. F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 65, 31–78.

[11] Some references for Random Forests are given by:

Athey S., Tibshirani J., & Wager S. (2019). Generalized Random Forests. The Annals of Statistics, 47(2), 1148–1178.

Biau G. & Scornet E. (2016). A random forest guided tour. Test, 25(2), 197–227.

Breiman L. (2001). Random forests. Machine learning, 45(1), 5–32.

Burgette L. F. & Reiter J. P. (2010). Multiple imputation for missing data via sequential regression trees. American journal of epidemiology, 172(9), 1070–1076.

Caiola G. & Reiter J. P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. Trans. Data Privacy, 3(1), 27-42.

Ding Y. & Simonoff J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. Journal of Machine Learning Research, 11, 131–170.

Doove L. L., Van Buuren S., & Dusseldorp E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. Computational Statistics & Data Analysis, 72, 92–104.

Feelders, A. (1999). Handling missing data in trees: surrogate splits or statistical imputation? In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 329–334).

Mentch L. & Hooker G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. Journal of Machine Learning Research, 17(1), 841–881.

Reiter J. P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21(3), 441–462.

Saar-Tsechansky M. & Provost F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1623–1657.

Wager S., Hastie T., & Efron B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. Journal of Machine Learning Research, 15(1), 1625–1651.

[12] Some references for Support Vector Machines are given by:

Boser B. E., Guyon I. M., & Vapnik V. N. (1992). A training algorithm for optimal margin classifiers. Fifth Annual ACM Workshop on Computational Learning Theory, 144–152.

Chechik G., Heitz G., Elidan G., Abbeel P., & Koller D. (2007). Max-margin classification of incomplete data. In Advances in Neural Information Processing Systems (pp. 233–240).

Cortes C. & Vapnik V. N. (1995). Support-vector networks. Machine Learning, 20, 273–297.

Drechsler J. (2010). Using support vector machines for generating synthetic datasets. In International Conference on Privacy in Statistical Databases (pp. 148–161).

Drechsler J. & Reiter J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Computational Statistics & Data Analysis, 55(12), 3232–3243.

Hable R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. Journal of Multivariate Analysis, 106, 92–117.

Honghai F., Guoshun C., Cheng Y., Bingru Y., & Yumei C. (2005). A SVM regression based approach to filling in missing values. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 581–587).

Pelckmans K., De Brabanter J., Suykens J. A., & De Moor B. (2005). Handling missing values in support vector machine classifiers. Neural Networks, 18(5-6), 684–692.

Rogers S. D. (2012). Support Vector Machines for Classification and Imputation. Master thesis. Brigham Young University.

Smola A. J., Vishwanathan S. V. N., & Hofmann T. (2005). Kernel Methods for Missing Variables. In AISTATS 2005 – Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (pp. 325–332).

Steinwart I. & Christmann A. (2008). Support Vector Machines. Springer.

Stewart T. G., Zeng D., & Wu M. C. (2018). Constructing support vector machines with missing data. Wiley Interdisciplinary Reviews: Computational Statistics, 10, 1–16.

Wen Z., Shi J., Li Q., He B., & Chen J. (2018). ThunderSVM: A fast SVM library on GPUs and CPUs. Journal of Machine Learning Research, 19(21), 1–5.

Yang B., Janssens D., Ruan D., Bellemans T., & Wets G. (2013). A data imputation method with support vector machines for activity-based transportation models. In Computational Intelligence for Traffic and Mobility (pp. 159-171).

Zhang Y. & Liu Y. (2009). Data imputation using least squares support vector machines in urban arterial streets. IEEE Signal Processing Letters, 16(5), 414–417.

[13] Note that we did not valid statistical testing. The results are just shown in order give an additional impression on distributional aspects.

[14] In general, companies remain in the sample for 6 years. This panel-like structure allows observing employment relationships over time (with certain limitations). The data is collected in electronic form and by means of a postal survey. There is an obligation to provide information. The sample is a one-stage stratified random sample.

[15] The final decision which method will be used after the “probation period” will be made in collaboration of methodologists and subject matter experts.