Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Implementation of artificial intelligence and machine learning methods within the Federal Statistical Office of Germany

Prepared by Lydia Spies, Kerstin Lange; Federal Statistical Office, Germany

I. Introduction

1. The use of machine learning algorithms in the production of official statistics is a relatively new but omnipresent trend in statistical agencies all over the world. In 2015 the High-Level Group for the Modernisation of Official Statistics (HLG) published an overview of several applications of machine learning techniques currently in use or in consideration at statistical agencies. Most of the examples are in the areas of automatic coding, editing and imputation, and record linkage but of course the range for possible implementations is much wider, especially when you think about artificial intelligence in general and not only machine learning.

2. In the Federal Statistical Office of Germany (FSO), we have only limited experiences with machine learning applications so far. First concrete classification tasks in earnings and craft statistics and in the business register have already been done but there had not been an overall implementation strategy yet. This is about to change now. In 2017 we developed a digital agenda in order to advance the holistic digitalization of the organization. One of its main objectives is a faster and more agile provision of information. For this purpose, we are working on an automation of our processes. The focus thereby lies on an integration of artificial intelligence and machine learning methods into our statistical production processes.

3. The paper will give an overview of the measures taken in order to establish artificial intelligence and machine learning methods within the FSO. Our motivation regarding the implementation of artificial intelligence and machine learning methods is outlined in Section 2. In Section 3 some of the measures taken so far are introduced. Section 4 presents a first test application of machine learning methods for imputation. The paper finishes with an outlook.

II. Motivation

4. Our interest in artificial intelligence and machine learning methods at this very moment is not random. There are several parallel developments that are responsible for our efforts. In this section two of the main triggers are explained.

A. Digitalization

5. The need for a digital transformation of the German Statistical System results from a constantly proceeding digitalization of our society. First of all the amount of digital produced data is growing rapidly. While so called big data allows us to carry out completely new evaluations, which were not possible before, their analysis and interpretation require much faster processes and the application of new technologies. Second of all the expectations of our users are rising too. In a highly technical and

digitalized world, they expect a tailor-made supply at the push of a button. Currently, however, the development of new statistics or even just the modification of already existing statistics often takes several years, which makes it difficult to meet these expectations and necessary to adapt our processes.

6. Digitalization is not a new topic for the FSO and many measures have already been implemented. While a few years ago many businesses still reported their data on paper, today we receive their reports only electronically. To this end, an online reporting obligation was introduced and consistently enforced for business statistics. Another example is the development of new innovative presentation formats for our users, such as interactive maps for a variety of topics. So far, however, the measures have usually been limited to individual areas or process steps. An overall ambitious target and a comprehensive plan were yet missing.

7. Therefore in 2017 a digital agenda for the FSO was developed. The digital agenda is the strategic framework for the FSO's digital transformation which has basically 5 aspects:

(a) Electronic data management:
All statistics and documents (also from support processes) are available electronically and data are collected electronically.
(b) Digital workflows:
Processes are end-to-end digital and all process steps are supported by IT systems. No manual import/export and no media discontinuities between systems.
(c) Automation of process steps:
Individual process steps are as automated as possible. Examples are automatic quality control or data collection through intelligent IT systems.
(d) Development of new digital services:
New user-specific analyses are possible through the use of new analytical methods and the integration of data sources and registers.
(e) Setup of a platform:
A platform is created that connects other authorities, companies and additional external partners. The FSO is thus developing a pioneering and creative role in the field of digitalization.

8. One of the concrete goals till 2020 is a faster and more agile delivery of complex new information and statistics, with the target that for 90% of the statistics the results should be available in less than three months after data collection. This can only be achieved by highly automated processes. Therefore several measures have been started in the FSO in order to achieve a complete automation of the data collection, the data processing and the data analyzing phase.

9. In context of data editing and imputation there are of course plenty of “traditional” methods that are suitable for automation too and we will continue to use and consider them as well. But we are also very interested in exploring the advantages of artificial intelligence and machine learning methods and we will focus on this area for now. On the one hand, these methods are very promising in terms of automation. In theory a self-learning system which is able to learn from previous surveys and improve over time will make it even unnecessary to define the edit rules manually or to test and adopt the imputation model every year. On the other hand, our interest is also strongly politically motivated.

B. Political interests

10. After the federal elections in 2017 a new government was introduced. This new government determined the focus of their collaboration in the next legislative period in a coalition contract. One common agreement was that research on artificial intelligence will be intensely promoted including accompanying social and humanities research. In particular, it is planned to make Germany one of the world's leading location in the study of artificial intelligence.

11. Subsequently in July the key points for a strategy with regard to artificial intelligence were published by the Federal Government. One of the main goals is for Germany to become the world's leading location for artificial intelligence in particular through a comprehensive and fast transfer of research results into applications and the modernization of administration.

12. Following this strategic direction it was decided by the management that the FSO is going to contribute to this focal topic as well. Several short and medium-term measures have been initiated in order to introduce artificial intelligence and machine learning methods within the FSO. The following section will highlight some of them.

III. Implementation strategy

13. So far our experiences with artificial intelligence or machine learning methods are very limited. Only in business statistics several applications of classification with machine learning techniques have been tested and concrete classification tasks in earnings and craft statistics and in the business register have already been implemented. But in order to extensively and permanently introduce artificial intelligence and machine learning into our statistical production processes or even our administration processes profound expertise and much more experience is needed. Since a profound expertise and experience is not gained over night several short and medium-term measures have been initiated that will also include external assistance from the scientific sector.

14. In order to identify on what to focus our efforts first a Proof of Concept about machine learning has been started end of last year. The purpose is to conduct a feasibility study on the application of machine learning methods in official statistics. Part of this project is to make an inventory about internal and external applications. Therefore a questionnaire has been sent to various national and international statistical agencies. Afterwards an internal questionnaire was sent to all subject matter departments in order to identify conceivable applications within the statistical production process. The results will be used to determine the most promising applications and to decide which realizations should be tested first.

15. Another internal project deals with the evaluation of machine learning procedures as a possibility for the implementation of automated data preparation. Since a lot of the data editing is still done manually the potential for an efficiency gain in this process step is considered quite big. The aim is to examine different machine learning methods for data editing and imputation and respective software implementations in R and Python. Furthermore a software developed by researchers from Stanford University and the University of Waterloo called HoloClean, a statistical inference engine to impute, clean, and enrich data, will be tested within this project. The result of this project should be a better understanding of the pros and cons of the different machine learning methods and the respective application requirements.

16. Although these two internal projects are not finished yet one result became apparent already in the very beginning. For a quick and overall introduction of artificial intelligence and machine learning methods within the FSO all the knowledge, expertise and experience need to be gathered and concentrated in a centre of excellence within the methodology department. Additional, in order to be successful more staff has to be allocated to this topic and/or new staff has to be hired. A concept for the implementation of such a centre of excellence and the required staffing is currently getting evaluated by our management.

17. Parallel to the implementation of an internal centre of excellence and in order to accelerate its build-up, we are going to purchase external knowledge as well. The plan is to commission a consortium of researchers with the compilation of a report about various topics with respect to artificial intelligence and machine learning methods. The contents of the report are currently being discussed with the statistical offices of the federal states and our IT service provider. It will include topics about the current state of research in the field of artificial intelligence and machine learning, available software, required IT infrastructure, legal and ethical issues and recommendations for applications in official statistics. This report is intended to create a common knowledge base and serve as a starting point for further developments and applications.

18. Even though there are a lot of activities in progress right now and we have ambitious goals, right now we are still at the very beginning. In the methodology department we are currently looking into different machine learning methods for imputation and their respective software solutions in R. In the spirit of a learning by doing approach we used an imputation task we are anyway dealing with right now as our test application. First results are presented in the next section.

IV. Application

A. Data characteristics

19. The microcensus is a large household survey with a sampling fraction of 1% of the persons and households in Germany. Annually around 830000 persons in about 370000 private households are interviewed to get representative statistics about the German population. Since 1957 the microcensus supplies statistical information about the population structure, the economic and social situation of the population, families, the housing situation, health, employment, job search and education. For a majority of the questions the sampled households are obligated by law to answer.

20. Since 2008 additional to the core programme information on childbirth is collected every four years. Questions concerning childbirth are addressed to women from 15 to 75 years. The women are asked whether they gave birth to at least one child and if yes, to how many children in total. Both questions concerning childbirth are optional. Due to this fact there is a nonresponse rate of nearly 9% for the two variables. To get valid inferences about the whole population missing values have therefore always been imputed, but it is planned to improve the imputation procedure for the next round.

21. For this purpose we first of all examined the missing mechanism in the 2016 data. It became apparent that the nonresponse is not missing completely at random (MCAR). From the considered variables the type of collection method had the highest relationship to the nonresponse.The two most commonly used types were laptop interviewing and filling out a paper questionnaire. Among the women between 15 and 75 years who answered by paper questionnaire, 19% did not answer the questions concerning childbirth in 2016. In comparison the data from laptop interviews only contained 3% missing values.

22. Furthermore there are differences between these two groups of women concerning their living and family situations and their level of education. Women who answered the microcensus questions by paper questionnaire are two years younger on average, have a higher level of education, live alone more frequently and are more often single in comparison to women who took part in the laptop interview. At the same time in the observed data women with those characteristics have a lower probability of being a mother. That leads to the assumption that among the women who did not answer the questions on childbirth, there are more childless women. As a consequence the percentage of mothers could decrease after imputation.

23. The central publication estimates are the fraction of childless women and the average number of children per mother for different age groups. For those estimates also the standard errors are stated. Further publications without the specification of standard errors are the fraction of childless women and the average number of children per mother by region, level of education, marital status, country of origin, nationality, employment and income.

B. Evaluation of traditional imputation methods

24. First of all we were testing “traditional” imputation methods. A logit model for the imputation of maternity and predictive mean matching for the imputation of the number of children seemed to be a promising approach (Method A).
As an alternative the application of only predictive mean matching was tested as well (Method B). Here predictive mean matching was done to estimate the number of children including 0. Afterwards a deterministic imputation of the maternity was realized where women with an imputed number of 0 children were set to “childless” whereas all others were set to “mother”. For implementations the R package MICE 3.0.0 was used.

25. In order to evaluate the imputation model and to compare the two options a small simulation study was done with the data from 2016. Based on the complete data 10% of the values for maternity and the number of children were deleted at random. The generated missing values were imputed according to methods A and B. This way it was possible to compare the imputed values with the true values and the

population parameters for the imputed data with the original data respectively. For the comparison mainly the parameter estimates named above which appear in publications were used.

26. To reduce the chance that extreme results were produced by random the deletion of values and their imputation was performed 10 times. This small number already resulted in clear differences between the two methods and gave us a helpful impression on the quality of the results. Furthermore for each of the two methods the variation of the results between the 10 replications was very little, so we did not expect different results if we raised the number of replications.

27. First we included as predictor variables all variables relevant for publications as well as the type of collection method because of its strong relationship to the missing indicator (model 1). The appliance of method B resulted in a nearly identical fraction of mothers for the imputed values and the observed values. For method A however, the fraction was around 5% higher for the imputed values. Concerning the number of children, method B again provided the better results with regard to the distributions. With method A less childless women were imputed and much more mothers with two children. Because of those results we favoured method B and applied it to the real data to impute the real missing values.

28. With method B and predictor variables according to model 1 63.4% of the missing values were imputed as mothers. Among the observed women there were 66.1% mothers. Keeping in mind the assumption that there are less mothers in the missing data because of a missing mechanism which is not completely at random, the results seem quite plausible. Nevertheless when we looked at the imputed values in more detail there were quite some implausible effects.

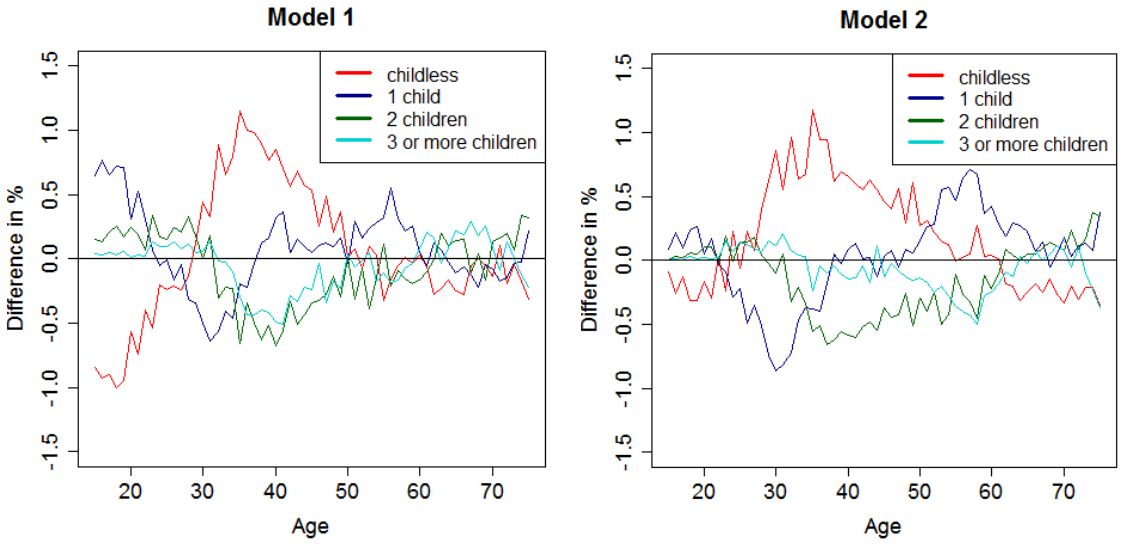

Figure 1: Differences of the fractions for the number of children by age before and after imputation in percent

29. Figure 1 shows the percentages of how much the fractions for “0”, “1”, “2” and “3 or more” children in the observed data differ from the fractions in the data after imputation. A positive difference stands for a higher fraction in the data after imputation than in the observed data. As we see the fraction of childless women between 35 and 45-years is about 1% higher in the imputed data while the fraction of childless young women under 25 is about 1% lower. If we make us aware of the fact that for 15 to 20-year old women the fraction of mothers in the observed data is only around 1% in total, this fraction would be about twice as high after imputation. This change seems really unrealistic from an expert’s point of view.

30. In further examinations we concentrated on the choice of predictor variables. We included variables which indicated that a woman belongs to the group of 15 to 20-year old (e.g. variables in the context of school) and variables which indicated that a woman is a mother with a high probability (e.g. person gets child allowance or works less because of child care). The selection of predictor variables was made on the basis of expert knowledge, an analysis of the relationship between the dependent variables

and possible predictor variables as well as a stepwise search. This resulted in 23 predictor variables in model 2.

31. As we can see in figure 1 the difference produced by imputation decreased for the young women in model 2 compared to model 1. Changes after imputation around 0.3% seem reasonable and can be explained on the basis of the missing mechanism and the data. Nevertheless the differences for the group of the 35 to 45-year old women remained nearly unchanged. If we look just at the imputed values alone 32% of the women are childless in this age group. In comparison for the observed data there are only 24% childless women in this age group. This difference between the observed data and the imputations cannot be explained by the missing mechanism alone.

32. In order to get an impression variance due to imputation we conducted a multiple imputation with m=10 for model 2 and method B. Thus the imputation variance could be included in the computation of the standard errors. For the published standard errors mentioned above there is nearly no difference between the standard errors yield by the observed values and the standard errors after multiple imputation. This shows that the conducted method did not introduce much variance in the estimation of the population parameters.

33. Nevertheless we are still not completely satisfied with our imputation results. Other sets of predictor variables were not leading to more plausible results either. In the microcensus dataset there are no better predictors which separate mothers from childless women. The appliance of method A on the real data did not improve the results as well. Hence we tried to use machine learning methods to approach our missing data problem.

C. Machine learning methods for imputation

34. In our tests we have just been dealing with the imputation of the maternity status so far. Of course later on we will also have to look into the imputation of nominal or ordinal variables with more than two categories, but for the beginning a dichotomous variable is the easiest to start with. As software we were using R because Python is not available in the FSO yet. We used 3 different machine learning techniques for classification with two different settings each in order to perform the imputation. We trained them on a 90% random sample of the observed data and tested them on the remaining 10%. Then the procedures were used for the imputation of the real missing values. In the following we will present the results based on the test dataset and the results of the final imputation.

35. The machine learning methods and respective parameters used for the imputation were the following:

(a) k-nearest neighbour classification by majority vote: (knn mode)
- The value of the majority of the nearest neighbours is used for imputation.
- k=10
- R package {class}, function *knn*

(b) k-nearest neighbour classification by probability: (knn probability)
- The value for imputation is drawn from the distribution of the the nearest neighbours.
- K=10
- R package {class}, function *knn*

(c) Classification with a decision tree by majority vote: (tree mode)
- The value of the majority of the observations in the respective final node is used for imputation.
- R package {tree}, function *tree*

(d) Classification with a decision tree by probability: (tree probability)
- The value for imputation is drawn from the distribution of the observations in the respective final node.
- R package {tree}, function *tree*

(e) Classification with a Support Vector Machine with a linear kernel: (svm linear)
- cost=0.001
- R package {e1071}, function *svm*

(f) Classification with a Support Vector Machine with a radial basis function kernel: (svm radial)
- cost=1
- gamma=0.8
- R package {e1071}, function *svm*

36. The results for the test dataset are shown in table 1.

	prevalence	accuracy	sensitivity	specificity	balanced accuracy	positive predicted value	negative predicted value	detection rate	detection prevalence
knn mode	0,6618	0,9110	0,9620	0,8113	0,8867	0,9089	0,9161	0,6366	0,7005
knn probability	0,6618	0,8669	0,9017	0,7986	0,8502	0,8975	0,8060	0,5967	0,6649
tree mode	0,6618	0,9214	0,9698	0,8268	0,8983	0,9164	0,9333	0,6418	0,7004
tree probability	0,6618	0,8684	0,8994	0,8078	0,8536	0,9015	0,8041	0,5952	0,6602
svm linear	0,6618	0,8937	0,9819	0,7210	0,8515	0,8732	0,9533	0,6498	0,7442
svm radial	0,6618	0,9208	0,9767	0,8115	0,8941	0,9102	0,9468	0,6463	0,7101

Table 1: Results for the test dataset

37. All quantities are based on the confusion matrix which is the number of true values in the test data plotted against the number of predicted values.

		Predicted	
		0	1
True	0	a	b
	1	c	d
			n

The considered quanities are:
- prevalence = (c + d) / n
- accuracy = (a + d) / n
- sensitivity = d / (c + d)
- specificity = a / (a + b)
- balanced Accuracy = (Sensitivity + Specificity) / 2
- positive predicted value = d / (b + d)
- negative predicted value = a / (a + c)
- detection rate = d / n
- detection prevalence = (b + d) / n

38. As you can see the accuracy is the lowest for the knn and tree approach where the imputations are drawn from the distributions of the respective observed values. At the same time the difference between the detection prevalence (=predicted fraction of mothers for the test dataset) and the prevalence (=true fraction of mothers in the test dataset) is the lowest for those two methods. This means that those two methods do not work that well in predicting the true values but do work best in predicting the true distribution. It is important to keep this result in mind because imputation is not about predicting the true values but to obtain statistically valid inferences from incomplete data. Thus we cannot evaluate imputation methods by their ability to optimize classification accuracy (van Buuren, 2012).

39. Unfortunately due to computational issues we were not able to set up a proper simulation study so far in order to see which machine learning method can predict the true distribution best over a number of replications. But since this result would only be valid for this particular data situation and in practice we do not have the time to conduct a simulation study for every task, we need to determine general

performance quantities anyway. This will be part of our internal project for the implementation of automated data preparation.

40. After the training and testing of the machine learning methods we applied them on the full data to impute the missing values of the maternity status. As mentioned before our assumption was that the percentage of mothers would decrease after imputation. This effect occurred for all methods but the support vector machine with a linear kernel as you can see in table 2.

	Mothers	Childless women
observed	0,6613	0,3387
knn mode	0,6592	0,3408
knn probability	0,6565	0,3435
tree mode	0,6595	0,3405
tree probability	0,6572	0,3428
svm linear	0,6646	0,3354
svm radial	0,6606	0,3394

Table 2: Distribution of maternity status before and after imputation

41. Figure 2 shows the percentages of how much the fraction for childless women in the observed data differ from the fraction in the data after imputation for the six applied machine learning methods.

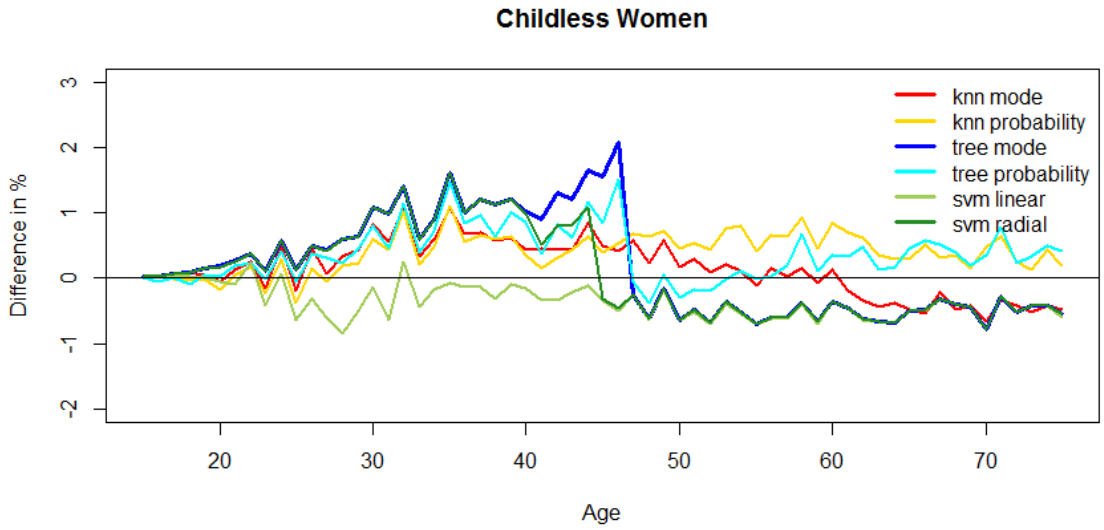

Figure 2: Differences of the fractions of childless women by age before and after imputation in percent for the six machine learning methods

42. As you can see none of the methods is producing a noticeable difference in the fraction of childless women for the 15 to 20-year old women as we observed it with model 1. On the other hand besides the support vector machine with a linear kernel they all show the same effect as already seen with model 1 and model 2for the 35 to 45-year old women. The level of the difference varies over the methods but without knowing the truth we cannot decide which is the right amount. But since all those different approaches showed all more or less the same effect this might be an indication for the effect being correct.

V. Outlook

43. As already mentioned we have just started to look into machine learning methods for imputation. Before we can actually apply them in our production process we still need to learn much more about their respective advantages and disadvantages. But one of the first problems that have to be solved is the computational limitation we were faced with in our test application. Therefore we are already working very closely with our IT service provider ITZBund in order to find a solution that preferably will not only be used by the FSO but also other authorities.

44. Furthermore we are very interested in cooperating with other users of artificial intelligence and machine learning methods especially in context of statistical production. We have already started collaborations with other national statistical agencies and researchers. But additional to that we are particularly interested in international cooperation with other NSI in order to learn from the more experienced ones and to explore and develop together the rising potential of those new techniques.

References

CDU, CSU and SPD (2018): Koalitionsvertrag 2018, URL: https://www.bundesregierung.de/Content/DE/StatischeSeiten/Breg/koalitionsvertrag-inhaltsverzeichnis.html

Chu, K.; Poirier C. (2015): Machine Learning Documentation Initiative

Die Bundesregierung (2018): Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz, URL: https://www.bmbf.de/de/eckpunkte-der-bundesregierung-fuer-eine-strategie-kuenstliche-intelligenz-6578.html

Lantz, B. (2015): Machine Learning with R (Second Edition), Birmingham: Packt Publishing Ltd.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Rekatsinas, T.; Chu, X.; Ilyas, I.; Rè, C. (2017): HoloClean: Holistic Data Repairs with Probabilistic Inference, URL: https://arxiv.org/pdf/1702.00820.pdf

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

Van Buuren, S.; Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, December 2011, Volume 45, Issue 3, URL: http://www.jstatsoft.org/v45/i03/

Van Buuren, S. (2012). Flexible imputation of Missing Data. Boca Raton: CRC Press