Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Automation of E&I processes

Prepared by Kerstin Lange, Federal Statistical Office, Germany

I. Introduction

1. A well-known problem official statisticians have to deal with is the fact that collected data are not clean. Thus, an editing or imputation step is required before the actual analysis is performed. Mostly, this step is done manually. The arising problem is that these manual steps bring high costs with regard both to time and effort. To improve the corresponding processes, an automation of the editing and imputation (E & I) process is strived in the Federal Statistical Office of Germany.

2. In Germany, currently in many statistics the editing and imputation step is still performed manually for all microdata, independent of the effect on the aggregated results. Furthermore, macro editing is not used generally. To improve the efficiency without any loss of quality, , we launched a project whose aim is the establishment of an automated editing and imputation in several statistics.

3. This paper gives an insight into the measures initiated to achieve the automation of editing and imputation processes. Our motivation to engage with the topic is described in Section 2. Section 3 outlines the first steps we initiated to approach to the topic and contains a description of the tools applied to edit and impute. A case study, where some of the tools were tested, is illustrated in Section 4. The paper concludes with an outlook given in Section 5.

II. Motivation

4. In Germany, the Federal Statistical Office and the Statistical Offices of the States have the order to generate official statistics by law. At the same time, they are encouraged to improve and enhance the quality of their published statistics continually. This requirement refers to the collection, preparation, analysis and publication of the data. The automation and standardization of all process steps are important and an ongoing tasks. The digitalization offers new perspectives and potential for further improvement. This potential was recognised by the Statistical Offices. For this reason they developed a so-called Digital Agenda. The Digital Agenda describes the status quo, formulates the reasons why the Digital Agenda is required, names its goals and contains the measures which will be conducted to achieve these goals.

5. One of these goals is to increase the transparency and standardization of the corresponding processes. This will both enhance the data quality and make the statistical production more efficient. One of the measures related to this target concerns the process of editing and imputation. The aim is to apply automated and, if appropriate, self-learning machine learning methods for editing and imputation. Subsequent to a successful editing, an analysis of the data and, if necessary, a manual control is aspired.

6. The Statistical Offices detected a requirement to take action in this area because currently often microdata are checked manually without checking the aggregated results. Existing methods for editing

and imputation are not self-learning. A validation of the results is usually not executed. The measure is planned to finish at the end of the year 2020. After this period, automated methods should be implemented in selected statistics. They should result in a faster and more efficient E & I process without a loss of quality.

7. An automated editing and imputation process has many advantages. In the case of manual edits, the editor has to check huge amounts of data. If a suspicious value appears, the editor has to contact the respondent and to correct the value. That is a time consuming task. An automated process is, thus, a much faster way to correct the data. Manual editing and imputation would not be feasible in situations with millions of data values. Besides, it is easier to implement and ensure the consistency between different variables and their values. Furthermore, automated editing reduces the respondent's burden because no interaction with the respondents is necessary. Finally, correcting values with an automated process are by no means subjective. An automated editing process would give reliable results, independent of the editor conducting the process.

III. First steps and E&I tools

8. We approach the task in two ways: from the structural and from the methodical point of view. Introducing an automated E&I process is a common measure from the Federal Statistical Office and the Offices of the States. Thus, we gathered a working group who engages with the topic of editing and imputation. The working group is guided by the Federal Statistical Office. All Statistical Offices of the States were asked to send representatives to join the working group. The main focus of their work is on the development and evaluation on methods rather than on the IT framework and software development of new tools.

9. From a methodical perspective, the working group started with conceptual works on the definition of editing. There are many different terms in the German and English literature which describe the process. Sometimes terms are used synonymous, whereas there are differences with regard to the content on other cases. We decided that the minimum requirement for an automated E&I process is that the effort for manual editing is reduced by a substantial size. This does not necessarily mean that machine learning methods have to be introduced.

10. Subsequently, we investigated existing methods and tools for editing and imputation with a focus on machine learning methods. We generated a list and examined briefly if there is potential for the use in our statistical production process. After that we picked three tools we want to test within the framework of the measure. The three tools described below are stand-alone solutions and are especially built for editing and imputation.

A. CANCEIS

11. CANCEIS (Canadian Census Edit and Imputation System) is a tool for editing and imputation. It was developed by Statistics Canada for the use in the Canadian Census. The capabilities of the system have expanded over time so that CANCEIS is now also used in other surveys as well as in other statistical agencies around the world. The system is targeted on numerical, categorical and alphanumerical variables.

12. CANCEIS performs editing and imputation simultaneously. First, the user defines edits. Then the system uses the edits to check the data for plausibility. Finally, the data are imputed while taking the edits into account. CANCEIS consists of two independent components: The first component performs a deterministic imputation. Conditions and conditional actions are specified by the user. The second component performs a donor imputation. Imputed values are obtained from a donor who is as similar as possible to the recipient with missing or implausible values. In the process, similarity is determined by a calculation of distance functions and weights. Donor Imputation in CANCEIS is based on the Nearest Neighbour Imputation Methodology by Michael Bankier (Bankier, 2012). Hence, the performed imputations are data driven. It is also possible to use outlier edits in this step.

13. We integrated CANCEIS in our toolbox because the E&I system has many advantages. Due to the implemented algorithm, imputations by CANCEIS are always plausible. There is an efficient searching algorithm for the determination of donors, and CANCEIS automatically shows quality indicators and many useful additional reports concerning the imputation process. The user has the possibility to alter the imputation process as well as the output information by parameter adjustments. Other advantages are that the system is free of cost and is running without dependence on other software. Due to that, we test CANCEIS on further surveys within the scope of the measure for the automation of E&I processes.

B. BANFF

14. Another tool for editing and imputation developed by Statistics Canada is the Banff system. It is based on SAS and provides nine procedures. The procedures can be run independently due to a modular structure. Because of that, it is easily possible to adapt the process flow to the requirements of a survey. Banff is targeted to surveys with numerical and continuous data.

15. With the Banff procedures the user can specify and analyze edits and apply these edits on the data. An outlier detection based on two different methods and an error localization can be performed to identify the faulty records. To correct these errors, Banff offers the possibility to conduct a deterministic imputation first. After that, both donor imputation and the imputation based on estimators are feasible. If the user wants to perform the estimator imputation, there is a choice between twenty pre-defined algorithms and a custom-defined algorithm. Subsequent to the imputation itself, a procedure for prorating is available to ensure the components of a sum to add up to a desired total. As a last step, one Banff procedure can perform mass imputation in the case that a block of information should be imputed for the nonsampled units of a survey.

C. HoloClean

16. HoloClean is an integrated system for error detection and data repairing. The software was developed during the DAWN-project at Stanford University in 2017. The aim of this project is to make machine learning implementations accessible to subject matter specialists who are not familiar with software developing. Furthermore, it aims to integrate information from different resources. The system is targeted on categorical and alphanumerical variables and discretizes numerical variables to treat them. It is written in Python.

17. The general idea of HoloClean is to consider the E&I process as a supervised machine learning problem, i.e., the system uses training data to create a model. The model is then used to make predictions concerning the location of errors in the data and to derive corrections for the faulty values. In the current version of HoloClean, the error detection is deterministic. The module for data repairing follows the approach of a model-based imputation. The process of model-building resembles the one of a neural network.

18. The advantage of HoloClean is an automated process to detect and correct data on the basis of machine learning. Moreover, it is free-to-use and available for download with the source code freely accessible for everyone. The system has disadvantages as well, as it is a new tool and still under construction. For this reason, it does not include many functionalities yet. Furthermore, a missing documentation makes it difficult to understand the details of the method.

IV. Case Study

19. To evaluate the tools mentioned and to apply them on data, we picked a survey as a first use case which confronts us with new challenges. This survey is a new concept of the survey of earnings, named the new digital structure of earnings survey. The new approach is necessary because the committee for minimum pay demands a new request. This committee is a new data user and requires structural data on employees to evaluate the effect of minimum pay. As a second trigger, we are facing a request from the conference on ministers of women and gender equality. They ask for an annual computation of the adjusted gender pay gap. Until now this index was available just every four years.

20. To satisfy these requests, we will combine two existing surveys: the survey of earnings and the structure of earnings survey. The data for the survey of earnings are collected quarterly, but only contain aggregated data about a subset of the population and just around 40000 holdings. The data for the structure of earnings survey contain individual data for all groups of employees and for all relevant sectors. However, this survey with approximately one million sampled records is only conducted every four years.

21. The combination of the two surveys leads to a larger amount of data. The new digital structure of earnings survey will include individual data from nearly seven millions records. These data will be provided monthly. The data is collected by the Statistical Offices of the States who did a lot of manual editing so far before the junction to a nationwide dataset. This restructuring leads to a challenge which cannot be managed by utilizing the current editing and imputation process which heavily relies on manual data cleaning. Without an automated E&I process the realization of this new survey would not be possible.

A. Data Characteristics

22. Since the new digital structure of earnings survey will be conducted in 2021 for the first time, we cannot test any methods and tools with data from this survey. As the current structure of earnings survey is similar to the new survey to a high degree, we use the structure of earnings survey to perform tests. We selected the features which will be part of the new survey as well. These are the gross monthly earnings, weekly working hours, paid hours, level of education, occupational code number and length of service in the enterprise as well as demographic feature like age, sex and nationality.

23. The used test data contain individual data and are obtained from a sample of around 500,000 records. We will use the raw, that is unedited, data as test data. The collection of data and the editing and imputation is a task of the states. Error detection was done with user specified edit rules which were executed with a standardized tool of the Federal Statistical Office and the Statistical Offices of the States. The data cleaning of those values which were detected as faulty was done by the Statistical Offices of the States. They changed or imputed values on the base of subject matter expertise or contacted the enterprises again to make corrections. For the tests of the automated tools, we will use this sample with the manually corrected values for comparisons.

24. The focus of the following analysis is on the examination of the gross monthly earnings. This information is directly provided by the enterprises, whereas other data about the earnings (e.g. gross hourly earnings and gross annual earnings) will be derived from this measure in the new survey. Concerning the gross monthly earnings, in 3% there is a difference between the raw data and the manually edited data.

B. Editing and Imputation with HoloClean

25. At the beginning, we use HoloClean as a tool to conduct our tests. The system can be split into two parts: one of them for error detection and the other one for data repairing. First, we aim to see how many values will be detected as errors by the algorithm. The error detection in HoloClean is based on so-

called denial constraints. These are specified by the user. Since HoloClean is still under active development, it is not possible to transform every edit into denial constraints.

26. Nevertheless, compare the rates for detected errors. To do so, we use a sub-sample of the test data. The sub-sample contains of about 10000 records. From the dataset with nineteen variables 12.3% of the values were detected as erroneous. The differentiation between missing values and values which fail edit rules is not implemented yet. HoloClean flags all values as erroneous which are included in the failure of an edit rule. For the same dataset 6.9% of the values were changed after the complete editing process. That means that HoloClean detects more potentially erroneous values than detected by manual editing, although the number of edit rules was less. This difference can be explained as follows: While the manual editing process only changes a minimum required number of variables in case of a failed record, HoloClean flags all variables as erroneous in the first step.

27. Regarding the data repairing step, we were confronted with some problems. The problems concern functionalities which were not implemented yet, the runtime and methodologically unmet assumptions. Missing functionalities could partly be inserted by the software developers of HoloClean during our test phase, but it is still not possible to transform every edit. However, running HoloClean is possible at the moment. The problem is that inside our test environment, we faced restrictions with the working storage. Because of that and because of the algorithm, which treats all variables as categorical, it is only possible to run datasets with a number of records which is much less than the expected number of records in the new digital structure of earnings survey.

28. Another problem is HoloClean's assumption that all variables are categorical. As a default HoloClean conducts a discretization of all numeric variables by generating one category for every observed value. Naturally, this procedure implies negative methodological consequences as well as a negative effect on the runtime. To prevent a building of one class per observed value, bins can be created. Binning brings a loss of information. However, our first aim is to get HoloClean run with more than a hundred records.

29. For the gross monthly earnings, we used binning with quantiles. We chose to use ten bins so that the data were arranged into the deciles. With these changes we could run HoloClean with a sub-sample of 10,000 records. Afterwards, the imputed classes had to be transformed back into numerical values. As a start we inserted the median of the imputed class. For the evaluation of the run we used the average of the gross monthly earnings. The results are shown in Table 1.

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with HoloClean in Euro	Difference in %
Overall	2340	2435	4.1
Men	2848	2509	-11.9
Women	1739	2347	35.0

Table 1: Comparison of the mean gross monthly earnings for manually versus HoloClean edited dataset in a sub-sample of 10,000 records

30. When comparing the HoloClean results with the manually edited data, we see that the mean gross monthly earnings differ by 4.1%. For the subgroups by sex the differences are even bigger and yield unsatisfying results. This test shows that dealing with numerical data is a problem for HoloClean. Because of the discretization we inserted the median of the imputed classes which leads to biased estimators of the mean gross monthly earnings. Maybe the use of an alternative instead of inserting the median would provide slightly better results, but we omitted that because it does not seem promising enough.

31. The test run with the sub-sample of 10,000 records had a long runtime. On a machine with 16 GB RAM, 4 kernel CPU and 200 GB SWAP-space it took 3 hours. Keeping in mind that the expected amount of data will be around 7 million every month, this tool seems inappropriate in practice. Because of the worse results concerning performance and the quality of imputation for the gross monthly earnings, and because of a missing method documentation, we decided to reject the use of HoloClean for this

survey. We will observe the development and will probably examine a new version of the tool again for a different use case, if documentation and performance have been improved.

C Editing and Imputation with CANCEIS

32. As an alternative we tested CANCEIS. This edit and imputation system is also used so far in the structure of earnings survey. With CANCEIS it was possible to conduct all records from the sample of full time employees (almost 500 000) in one run. No data preparation step is needed in advance which changes the values (like discretization). The CANCEIS run could be conducted in less than five minutes. This suggests that we do not have to face problems even with regard to the expected amount of data within the new digital structure of earnings survey.

33. Table 2 shows the mean gross monthly earnings computed with the data imputed by CANCEIS in comparison to the manually edited data. The differences are very small. The biggest differences occur in the subgroup of over 65 year-old with 0.29% and for the managers with 0.28%. On the basis of these results we can state that CANCEIS is a proper tool for the imputation of the gross monthly earnings in the context of the new digital structure of earnings survey.

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with CANCEIS in Euro	Difference in %	N
Overall	3226	3223	-0.08	490 795
Men	3477	3473	-0.12	336 551
Women	2677	2678	0.04	154 208
≥ 65 years[1]	3882	3871	-0.29	2 610
Managers[1]	6646	6627	-0.28	45 981

Table 2: Comparison of the mean gross monthly earnings for manually versus CANCEIS edited dataset for all full time employees

D Editing and Imputation with the R package missForest

34. Due to the fact that we work on incorporating machine learning methods in the production processes in our NSI, we were searching for a machine learning algorithm which could fulfill the survey's requests. In earlier tests we found out that random forests provide good results for the structure of earnings survey. An algorithm which uses a random forest approach and which was developed for the imputation of missing values is included in the R package missForest (Stekhoven & Buehlmann). The algorithm can impute continuous and/or categorical data including interactions and nonlinear relations.

35. A disadvantage of the missForest function is that it does not take into account any edit rules. That means that we had to implement a shoot-out algorithm. The process starts with the imputation of missing and erroneous values by missForest. Next, we apply the edit rules on the datasets. The imputed values of the records which fail the edit rules were deleted and imputed by the missForest function again. This cycle is repeated until all records pass the edit rules or a stop criterion is fulfilled.

36. Because of the iteration between imputation and editing the execution needs some time for processing all records. For this reason, we again used a sub-sample of 10,000 records. The results of the imputation with missForest in comparison with the manually edited data are presented in Table 3. The first results achieved with this method show that we found a promising approach. Although the difference is larger than with CANCEIS, it is less than 1% compared to the manually edited data overall and for the subgroups by sex.

[1] Chosen subgroups with the highest differences

	Mean gross monthly earnings – manually edited material in Euro	Mean gross monthly earnings – edited with missForest in Euro	Difference in %
Overall	2502	2482	-0.80
Men	3094	3068	-0.84
Women	1808	1814	0.33

Table 3: Comparison of the mean gross monthly earnings for manually versus missForest edited dataset in a sub-sample of 10,000 records

E Further steps

37. We tested three tools for the editing and imputation process of the new digital structure of earnings survey in Germany. Among the tested tools CANCEIS provided the best results. The results of CANCEIS came closest to the manually edited data. For this reason we decided to incorporate CANCEIS in our production process. Before the final implementation we want to do some further tests. These will include the evaluation of different statistics for other variables and the search for the final CANCEIS parameters.

38. As an additional test solution we want to include the implementation with missForest in our future process which should run parallel to the productive CANCEIS solution. Thus, we can compare the edited results of two processes for the surveyed data and analyze the differences. Another idea is to test Banff as a tool for this survey. Although Banff doesn't seem to be appropriate for the imputation of the new survey, which contains continuous and categorical data, it is worth a try. It probably provides a good solution for the imputation of the numerical variables, for example the gross monthly earnings and, thus, allows a combination with other methods.

V. Outlook

39. The above mentioned new digital structure of earnings survey represents one case study where an automated E&I process is useful and necessary to fulfill the user's demands. After we have completed the tests on this survey, we will test further surveys. It also includes that we consult additional methods and tools. One focus of our investigations will be on Banff since we do not have any practical implementations in our NSI yet.

40. For the different tools we plan to work application requirements out. That includes the tool's assumptions and advantages and disadvantages. In addition we want to generate a list of use cases for the tools and methods so that subject matter experts can have a quick overview and pick a method which also seems suitable for their survey. This work should be done in collaboration within the working group of the Federal Statistical Offices and the Statistical Offices of the States. Furthermore, we are very interested in the experiences other NSI made and want to learn from their knowledge.

References

Bankier, M. (2012), Imputing Numerical and Qualitative Variables Simultaneously, Social Survey Methods Division, Statistics Canada.

CANCEIS Development Team (2015), "CANCEIS User's Guide Version 5.2", Social Survey Methods Division, Statistics Canada.

Rekatsinas, T., Chu, X., Ilyas, I. F., Ré, C. (2017b). HoloClean: Holistic Data Repairs with Probabilistic Inference. Proceedings of the VLDB Endowment 10(11), 1190-1201.

Statistische Ämter des Bundes und der Länder (2018): Digitale Agenda des Statistischen Verbunds.

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