Working Paper 10.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

Possible imputation procedures for the Census 2021

Prepared by Lydia Spies; Federal Statistical Office, Germany

I. Introduction

1. The German population Census 2011 combined a register based survey with a 10% sample survey. For the imputation of the missing or erroneous sample data a combination of cold-deck, deductive and nearest neighbour imputation was used. The employed imputation tool was developed and executed by IT-NRW, the Central statistical and IT services provider of one of the German Federal States, i.e. North Rhine-Westphalia.

2. In the course of preparing for the next Census 2021, we are currently evaluating different imputation methods and their implementation in different IT tools. Besides a comparison between the German imputation tool and CANCEIS, as two different implementations of nearest neighbour imputation, we also examine a possible application of multiple imputation. We are therefore looking into already available multiple imputation procedures in R and SAS.

3. The aim of the paper is to present the imputation methods under consideration, their application in different tools and to outline some practical issues. Special attention is drawn to the particular advantages and disadvantages of the different imputation procedures. Section 2 explains the motivation of the evaluation. The imputation methods and the tools are introduced in Section 3. Section 4 presents some advantages and disadvantages of the different options. The paper finishes with a final remark.

II. Motivation

A. Methodology of the German Census 2011

4. The main purpose of the household survey was the identification and statistical correction of outdated and missing entries in the population registers of large municipalities with 10.000 and more inhabitant because earlier tests showed that especially the registers of large municipalities are not always exact and reliable population figures cannot be derived from them as the only data source. Another objective was to provide information which is not, or not sufficiently, available in registers like data on migration background, education or employment data according to the definition of the International Labour Organization (ILO).

5. While the data collection was done by the German Federal States separately the whole validation and imputation process was executed by IT-NRW on a central server. For the imputation an extra ad hoc tool was developed. The imputation tool was specifically designed to be compatible with the respective standard validation tools of the German Statistical Offices and to be capable of being integrated to the IT-environment.

6. The imputation methods used were a combination of cold-deck imputation, deductive imputation and a single nearest neighbour imputation (Statistische Ämter des Bundes und der Länder, 2015). First,

implausible or missing values in variables like “sex” or “date of birth” were imputed by externally available information. Besides the population register another source for this cold-deck imputation were the lists with basic information about the household members compiled by the interviewers when doing the field work to establish the existence of the households.

7. In the next step a deductive imputation was applied. This was only possible for a small number of variables since it requires explicit relationships between variable values. For example for all respondents with plausible information about type of school and class level, a missing value in the filter question about if they had been visiting a school in the reference week were set to “yes”.

8. All values that were still missing after those two steps were imputed by a single nearest neighbour approach. For this purpose the data set was split into 4 blocks of variables. Similarity between donor and recipient was only required within a block. After adopting the information of the donor to all implausible or missing values of the recipient the validation step was repeated. Values that were plausible before and implausible after the imputation were also taken from the donor then. Some variables like “sex” and “date of birth” were excluded from the hot-deck procedure. The result was a single, plausible and complete data set.

9. Since the primary objective of the population census was the determination of the number of inhabitants on a detailed regional level an exact variance estimation of the population figures was part of the regular data production process. For the other survey estimates exact error estimation was not practically possible during the data production phase. However, according to the common publication practice of the Federal Statistical Office of Germany estimates with a relative standard deviation of more than 15% are not published. For implementation of this standard practice a simple rule of thumb criterion based on the observed frequencies was used to assess whether a particular estimate was fit for publication.

B. Ex post evaluation of the Methodology

10. Because “after the Census is before the Census” as we say we started the evaluation of the used methodology shortly after. An already well-known drawback of the used imputation methods is that the variance due to imputation cannot be determined from the final data itself. Since the errors have not been estimated exactly anyway this was irrelevant for the Census 2011. But since an exact error estimation is also part of the ex post evaluation this might be handled differently in the next Census.

11. So on behalf of the FSO a team of researchers from the University of Trier and the Leibnitz Institute for the Social Sciences (GESIS) was assigned to analyse different variance estimation methods in combination with several imputation approaches. In the study multiple imputation and resampling methods for variance estimation under single imputation, e.g. hot-deck imputation, were examined (Münnich et al., 2015).

12. The result was a clear recommendation for multiple imputation mainly because resampling methods are computationally very expensive. While for correct variance estimation via resampling a lot of subsamples have to be drawn and imputed, e.g. 500, multiple imputation can already achieve good results with much less imputed datasets. In fact another result of the study was to use m=50.

13. Following up the results of the study one of the first steps was therefore to implement a multiple imputation procedure and test it on the data of the last Census. The aim was to try to assess the impact of the imputation variance and to see to what degree a single imputation approach, like it was used in the last census, leads to an underestimation of the errors.

14. But besides the compelling methodological advantages of multiple imputation it is still an imputation method we have not a lot of practical experience with. So far we have only had one very simple application of it in an actual statistics production process, i.e. the imputation of water consumption data of the German Agricultural Census 2010. The imputation of the population census is a much more complex task and the focus on multiple imputation would comprise a significant project risk.

15. So we are not only testing an implementation of multiple imputation but we are also looking into options to improve the nearest neighbour imputation procedure outlined in Section A. One possibility is to extend the features of the German ad hoc imputation tool. Until now only two distance measures are implemented. That resulted in a limited pool of close donors and some undesirable adjustments were necessary which I will talk more about in the next section. In order to improve the imputation process some modifications and expansions are worthwhile.

16. Another possibility is the application of the well-established nearest neighbour imputation software CANCEIS developed by Statistics Canada (CANCEIS Development Team, 2015). An application of CANCEIS has already been an option for the last population census and it was actually used for the German housing census 2011 (Grunwald et al., 2014). The decision to not use CANCEIS for the population census was mainly technology driven. The effort and costs of programming an own imputation tool were estimated to be less than the ones of embedding CANCEIS into the production workflow.

17. In the next census the frame conditions will possibly be different though and the execution of the imputation task might be conducted centrally by the Federal Statistical Office of Germany. So the cost-benefit analysis might now result in favour of CANCEIS since it already features a wide range of distance measures and provides flexibility through a wide range of user-defined parameters. Another part of our evaluation is therefore a comparison of the German imputation tool with CANCEIS.

III. Methods and tools

18. The decision about the imputation procedure is two-stage. First of all the imputation method has to be determined. While cold-deck imputation and deductive imputation as first process steps are already very likely to be reapplied it is still open which method to use for the final imputation step. The choice will be made between multiple imputation and nearest neighbour imputation. Only after that the second decision about the imputation tool can be made. In this section the methods and tools under consideration are described.

A. Multiple imputation

19. The idea of multiple imputation is to impute every missing value not just once but m>1 times (Rubin, 1987). This approach results in m imputed datasets which then have to be analysed separately. To get the overall estimate those m separate estimates are combined by simple combining rules. This proceeding makes it possible to include the uncertainty due to the missing values into the calculation of the errors.

20. To perform those multiple imputations there are plenty of methods available. Since here we are dealing with a multivariate missing data pattern the two possible options are Joint modelling and Fully conditional specification (van Buuren, 2012). The assumption for Joint modelling is that the data can be described by a multivariate distribution and the imputations are then drawn from this distribution. Most commonly a multivariate normal distribution is used. In the German population census all variables besides age are categorical. Thus the assumption of a multivariate normal distribution as required for Joint modelling does not hold. Another disadvantage of this approach is that it imputes continuous values even though categorical values are needed. (Schafer, 1997) for example suggests rounding of the continuous imputed values but this can lead to biased results.

21. We decided to only consider Fully conditional specification because this approach allows us to specify a separate imputation model for every single variable. Imputations are created iteratively per variable. The multivariate distribution $P(Y, X, R|\theta)$ is formed by the set of the conditional univariate distributions $P(Y_j|X, Y_{-j}, R, \theta_j)$ as described in (van Buuren, 2012). For most of the variables we used a polytomous regression model. However, following the advice of (van Buuren, 2012) for variables with a high number of categories we applied predictive mean matching. Although this method is based on the normal linear regression model only observed values are imputed and thus no rounding is necessary.

22. One drawback of multiple imputation is that there are not a lot of options to include edit constraints into the imputation process. This is especially true for edit rules regarding categorical variables. (Münnich et al., 2014) recommended the application of a so called shoot-out procedure to attain plausible datasets. The idea is to alternate the validation and the imputation step until plausibility is achieved. In case of a multiple imputation the shoot-out has to be applied to all m datasets separately.

23. One possible tool to perform Fully conditional specification is the R-package MICE (van Buuren and Groothuis-Oudshoorn, 2011). MICE is a very flexible tool with a lot of different options and user-defined parameters. But first tests showed that the performance regarding execution time and memory size is not fit for a survey as large as the German population census. Another possible tool is the MI procedure in SAS 9.4 (SAS Institute Inc., 2015). Up to SAS 9.2 the MI procedure was only capable of applying Fully conditional specification in case of monotone missing data patterns. But since version 9.3 it can also handle arbitrary missing data patterns. Since none of the tools is able to perform a shoot-out procedure this would have to be implemented separately.

B. Nearest neighbour imputation

24. Nearest neighbour imputation is a hot-deck imputation method where missing values are replaced with observed values from the same survey. The basic idea of nearest neighbour imputation is to impute the value of a donor unit that resembles the recipient unit as much as possible regarding the observed variables. The similarity between the units is measured by the (weighted) sum of the variable specific distance function scores. The lower the score is the better the donor resembles the recipient. Since only units that pass all edit rules serve as potential donors ideally the imputed recipient units are plausible too.

25. There are a lot of different distance functions that can be used to compare two values. In the German ad hoc imputation tool two distance measures are implemented. For most variables a simple comparison measure was used where the score is 0 if the variable values of the donor and the recipient are equal and 1 otherwise. Only for the variable “age” an extra linear scaled distance measure within age classes was implemented.

26. Because of the amount of variables with partially lots of categories the probability of finding a donor with a very small distance was quite low. So in order to improve the donor quality the variables were split into 4 blocks and the imputation was conducted for each block separately. The blocks were formed topic-wise in order to minimize the number of edit rules including variables from different blocks. The donor search was performed in different stages starting from sampling point level up to the respective Federal State level. After the imputation another validation step was performed in order to check whether originally plausible values had become implausible because of the imputation. If so those implausible values were also replaced with the donor values.

27. Besides the limitation of available distance functions the necessity to perform block-wise imputations is undesirable because this approach cannot assure the plausibility of the recipient units as soon as there are edit rules that include variables from different blocks. Furthermore the tool does not incorporate any measure about the amount of necessary change to the recipient unit into the donor search. But the ability of this imputation tool to automatically process the edit rules and outputs from our in house standard validation tools makes it worth to evaluate the feasibility of possible improvements of the tool.

28. Another option is to use CANCEIS which already has all the mentioned characteristics. The number of available distance functions makes it possible to conduct a much more specific donor search than just applying a distance of 1 to all values unequal to the recipients` value. Furthermore all edit rules are already taken into account and only feasible units can be potential donors. This approach makes a subsequent validation step unnecessary. Another advantage is that besides the distance between the recipient and a potential donor also the distance between the recipient and the imputed recipient is calculated when searching for the best donors. This approach favours donors that induce less necessary changes to the recipient unit.

29. Although CANCEIS is also able to perform deterministic imputation it cannot perform a cold-deck imputation. So we would still need a separate tool for this task. And since CANCEIS cannot interact with our standard validation tools all settings like edit rules would have to be entered separately.

IV. Advantages and disadvantages

30. The main and very compelling advantage of multiple imputation is that it enables a correct error estimation. Neglecting the uncertainty imputations come along with results in an underestimation of the standard errors of the estimates and the coverage rates of resulting confidence intervals will be below expectation. An application of multiple imputation is therefore very desirable from a methodological point of view. But the decision about the imputation method for the German population census 2021 is not only driven by methodological arguments but also practical ones.

31. Besides the advantage of resulting in correct error estimation, multiple imputation also comes along with a lot of practical challenges. One obvious is the fact that not only m datasets have to be stored but also every analysis has to be performed m times. This would increase the run-time of every single inquiry significantly. Also although the combining rules for the overall estimates are not complicated an extra user interface would have to be developed to enable everyone to work with the multiple datasets.

32. Another issue concerns the comparability of the quality of the results with the quality of the results from other surveys. According to the common publication practice of the Federal Statistical Office of Germany estimates with a relative standard deviation of more than 15% are not published. When the uncertainty due to the imputation is taken into account the relative standard deviation will rise for sure but not due to a lower quality of the results but due to an improved methodology. Communicating this fact would most probably be a challenging task. Furthermore a deviation from the common publication practice would be necessary in order to avoid a reduction of publishable results.

33. There are also some concerns about the feasibility of multiple imputation in an actual statistics production process. Due to a tight schedule there will most probably not be a lot of time for the imputation step. But since multiple imputation with Fully conditional specification is a model-based approach the final adjustment of the models can only be done when the actual data is available. Furthermore first tests showed that the imputation itself is very time-consuming because the models are computationally very expensive especially due to the high amount of categorical variables.

34. But the most serious disadvantage of multiple imputation is the inability to incorporate edit rules into the imputation process. The shoot-out procedure has a significant influence on the execution time especially in case of multiple imputation where e.g. m=50 datasets have to be validated and imputed and validated again and imputed again and so on. (de Waal et al., 2011) propose a method for the adjustment of the imputed data. But to the best of our knowledge there is no implementation of this method available so far. So from our point of view this is in practice still an unsolved issue and a lot of work still has to be done before it can be applied in an actual statistics production process.

35. If the benefit is worth the effort strongly depends on the decision about whether an exact error estimation will be applied. Multiple imputation makes only sense if standard errors are calculated. If the quality of the survey estimates will again be assessed by a simple rule of thumb criterion then there is no need for multiple imputation.

36. In this case we will have to make a decision between the German ad hoc imputation tool and CANCEIS. Compared to CANCEIS the German ad hoc imputation tool is a very rudimentary tool. The major advantage is its ability to interact with the standard validation tools and its potential to be extended to become a standard imputation tool for the German statistical offices in the long run. While there is no possibility to include own distance functions into CANCEIS we would be able to modify a German imputation tool according to our needs. But on the other hand we are well aware that CANCEIS is the result of over 20 years of work and experience and that it would be very challenging to develop an equally efficient tool.

V. Final remark

37. The aim of the paper was to outline the imputation methods and tools under consideration for the next German population census and to describe the relevant criteria which have to be taken into account. So far no final decision has been made. While in the next future we will focus on a rigorous assessment and comparison of the potential of different Nearest Neighbour imputation tools, in principle even multiple imputation is still an option.
As the theoretical advantages of multiple imputation are definitely convincing, even if the decision will be not to use it in the population Census 2021, we will nevertheless continue working on it, and assess its potential for application in the production process of other statistics.

References

CANCEIS Development Team (2015). CANCEIS User's Guide. Version 5.2., Ottawa: Statistics Canada

de Waal, T.; Pannekoek, J.; Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New Jersey: John Wiley & Sons

Grunwald, S.; Krause, A. (2014) Umgang mit fehlenden Angaben in der Gebäude- und Wohnungszählung 2011. Wirtschaft und Statistik, Ausgabe 8/2014, S. 437-449.

Münnich, R. ; Gabler, S.; Bruch, C.; Burgard, J.; Enderle, T.; Kolb, J.-P.; Zimmermann, T. (2015). Tabellenauswertungen im Zensus unter Berücksichtigung fehlender Werte. AStA Wirtschafts- und Sozialstatistisches Archiv 9, Nr. 3/4, S. 269-304.

Münnich, R.; Gabler, S.; Bruch, C.; Burgard, J.; Enderle, T.; Kolb, J.-P.; Zimmermann, T. (2014). Imputationsprojekt zum deutschen Zensus, Report, Statistisches Bundesamt

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/.

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

SAS Institute Inc. (2015). The MI procedure. SAS/STAT® 14.1 User's Guide, Chapter 75

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data, Boca Raton: CRC Press

Statistische Ämter des Bundes und der Länder (2015). Zensus 2011. *Methoden und Verfahren*, June 2015, Wiesbaden: Statistisches Bundesamt, URL: https://www.destatis.de/GPStatistik/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000960/Zensus2011_Methoden_und_Verfahren.pdf

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