UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

Evaluation of Census 2011 survey estimates

Prepared by Lydia Spies; Federal Statistical Office, Germany

I. Introduction

1. In 2011 the first register-based population census was conducted in Germany. While in the past the approach was that of a complete enumeration this time only a 10% sample of the population was surveyed. The primary objective was the determination of the number of inhabitants on a detailed regional level. Basis for the calculation were the population registers maintained by the municipalities. Although registration is compulsory in Germany, 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.

2. Therefore a supplementary household survey was conducted. One major purpose of the household survey was the identification and statistical correction of outdated and missing entries in the population registers of large municipalities with 10.000 and more inhabitants. Besides the statistical correction of the population registers another objective of the household survey 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).

3. The target for accuracy of population figures was specified in the census law as a relative standard deviation of at most 0.5% (§ 7 para.1 ZensG 2011). Precise population figures are fundamental in Germany, because these estimates are used as calculation basis for community money transfer. Therefore, an exact variance estimation relating to those figures was part of the regular data production process. Due to a tight dissemination schedule for all other survey estimates only a simple rule of thumb criterion based on the observed frequencies was used to assess whether a particular estimate was fit for publication.

4. In preparation for the next census round we are now evaluating the reliability of this criterion. Therefore for some of the survey estimates we are attempting an exact estimation of the sampling variance as part of an ex post analysis. Since all additional survey variables contain imputed values the variance due to imputation should also be taken into account. One practical way to achieve this is by doing a multiple imputation (MI) and not only imputing once but several times. So another effort of the ex post study is to do multiple imputation of missing or erroneous data items of the Census 2011 data set prior to the estimation of the sampling variance. We will 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.

5. The focus of the present paper is on the imputation part of the ex post study. Work on this study has started only recently. Section 2 explains the motivation and the approaches actually used in the census 2011 in some more detail. The data of the ex post study and the imputation methods imagined for

the study are introduced in Section 3. Section 4 presents the design of the MI application tested so far and a few early, preliminary results. The paper finishes with an outlook section.

II. Motivation

6. As mentioned in the introduction the motivation for this study consists of two parts. Both are aimed to evaluate the methodology of the last population census and to decide whether we can apply them again in the next one or if we need to revise them. In this section the motivation for this study is explained further and the methodology of the previous census is introduced.

A. Evaluation of the survey estimates

7. The sample for the population census 2011 was designed with regard to the accuracy requirement formulated in the census law: The survey estimates for the total population in all municipalities with at least 10.000 inhabitants should have a relative standard deviation of at most 0.5%. This means that for a municipality with 20.000 estimated inhabitants, with a probability of 95% the difference to the real number of inhabitants is less than 200. For the estimation of the population figures a generalized regression model (GREG) was used (Münnich et al., 2011). Since for those estimates exact error estimation was conducted we know how precise they are.

8. On average the finally observed standard deviations are about 0.56%. In 63% of the municipalities the standard deviation was higher than 0.5%. In fact, in most cases it was just slightly higher, but anyway the targeted accuracy goal was only partially reached. The reason is that in preparative tests only a fraction of the information provided by the real census was available and the actual R^2 was much lower than expected (Berg and Bihler, 2014).

9. 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 rule a simple test criterion was employed, using the coefficient of variation of a binomial distribution with parameters *n* (sample size) and *f* (sampling fraction): $\sqrt{\frac{1-f}{n\,f}}$. Replacing the expected value *n f* by the observed sample mean *k* (e.g. the number of sample observations with a given property like for example "females aged 34 with academic degree in municipality Y") the criterion turns out $\sqrt{\frac{1-f}{k}} < 0.15$. Hence the criterion for an observed frequency *k* to be considered "fit for publication" was to exceed $\frac{1-f}{0.15^2}$, otherwise the frequency was suppressed (Statistische Ämter des Bundes und der Länder, 2015). The average sampling fraction over the stratums varies from one municipality to another. But in most cases the threshold was between 30 and 45 observed cases.

10. In preparation for the next census round we now want to check the performance of this rough estimate. So for some selected results we are now doing an error estimation based on more refined variance estimation techniques. Although the additional variance due to imputation was not considered when the publication limit of a relative standard deviation of 15% was defined, in theory it is part of the uncertainty of a survey estimate. For the evaluation study we plan to correctly consider this part of the variance. One way to achieve this is by using a multiple imputed dataset for the variance estimation, although for the original survey actually a nearest neighbour single imputation method was used.

11. In case it turns out that (even ignoring the variance component of the imputation) the simplified variance estimate has a considerable tendency for underestimation, the next step would be to consider other estimation methods for the next census round. Depending on the results of the present study a follow-up study is foreseen with the purpose of testing small area estimation methods.

B. Testing multiple imputation

12. The imputation methods actually used in the last census 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.

13. 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 a 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”.

14. 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 variable blocks. 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.

15. Other variables like “occupation” needed a slightly different approach. This variable was asked as plaintext. In case of this variable the major part of the deductive imputation has been done during the codification phase. If it was not possible to assign an explicit occupation code, the encoders could add extra information in form of umbrella terms that were used in the hot-deck imputation.

16. The result was a single, plausible and complete data set. The question is to which degree ignoring the variance due to imputation leads to biased variance estimation. There are well known methods for variance estimation while correctly considering the variance induced by single imputation methods as for example described in (Davison et al., 2004). (Münnich et al., 2014) analyse different variance estimation methods in combination with several imputation approaches. The results of this work are the basis for our evaluation study.

17. Besides multiple imputation resampling methods for variance estimation under single imputation were examined in (Münnich et al., 2014). The basic concept of all resampling methods is to draw several different samples from the original survey sample. When a bootstrapping technique is applied, B samples of the same size as the original sample are drawn with replacement from the original survey sample (Shao and Sitter, 1996). For each subsample the single imputation step is processed and the respective survey estimate is calculated separately. The total variance of the survey estimate is then approximated by the empirical variance of the B bootstrap estimates.

18. One drawback of those resampling methods is that they are computationally very intensive. An even bigger disadvantage from a practical point of view is that they have to be applied again for every additional analysis. So the whole procedure of drawing the samples, imputing the missing values, editing the imputed values and calculating the survey estimates would have to be repeated over and over again. Especially for model based estimations like small area estimators this procedure is not applicable since it is not possible to adjust the model for every single sample.

19. Multiple imputation on the other hand seems to be a simple alternative, as it has to be applied only once. However, in this application one should consider a wide range of possible analysis of the imputed data set, in order to include the relationships of the variables relevant for an analysis in the imputation models. The multiple imputation step results in m>1 complete data sets useable for a variety of analysis. The m data sets could also be provided in the Research Data Centre in order to offer to interested scientist an easy way to achieve correct variance estimation.

20. After considering all pros and cons we decided to apply multiple imputation in our evaluation study in order to gain proper variance estimates. Also with regard to the envisioned test of small area

estimation, multiple imputation seems to be the only practical approach. And - last but not least - it will be interesting to find out how this theoretically appealing approach works in practice with a huge and complex data set like the German population census.

III. Evaluation study

21. Since the evaluation of the reliability of the rule of thumb threshold is one of the main purposes of this study and since most of the time the threshold lies between 30 and 50 we are concentrating on 4 to 6 dimensional tables with a high proportion of cells with frequencies between 30 and 50.

22. All selected tables contain regional information. For the present, first phase of the study we have limited the imputation to some specific regions. This should be enough to evaluate the threshold but it was also a decision due to limited computer resources. The results presented in this paper are based on some first test imputations for the Federal State Berlin.

23. In addition, we exclude responses from persons in special domains like retirement homes or prisons from the data base. Later on these domains might need an extra procedure as it was done in the "real" imputation procedure but this is a future research issue.

24. The remainder of this section will focus on one table to demonstrate the planned proceeding and to explain some practical difficulties that have to be faced when imputing missing values. The table is a hypercube formed by the variables "region", "sex", "occupation classification on main group level" (HGKLDB) and "section of the economic sector" (WZ_ABS). One difficulty here is the big number of categories in the variables HGKLDB with 37 categories and WZ_ABS with 23 categories. In return there are almost no edit rules regarding those variables.

A. The Data

25. The initial data we are working with is the population census data after the original editing and imputation process. So it is the complete and plausible data set that was used for publications and delivery of census hypercube data to Eurostat. In addition to the actually collected and the derived variables it also includes certain quality indicators. For every single collected variable there is one quality indicator of the same length stating exactly if and in what phase a respective value was changed or imputed during the editing and imputation process.

26. Therefore our first step is to re-create the missing values again. Since the variables in the tables we investigate in the study are almost all derived variables, one possibility is to create the missing values directly in the derived variables and to impute them. This approach has two drawbacks. First depending on the missing data pattern this can result in a high percentage of missing values in the derived variables. Second the variables contributing to the derived variable would have to get imputed anyway since they are obvious predictor variables in the imputation model for the derived variable.

27. Thus we created the missing values only in the collected variables or in variables that were only derived from one collected variable. Hence we are also going to impute only collected variables or variables that only depend on one collected variable. The variables used as spanning variables for the investigated tables are then derived again from the imputed data.

28. Because there is very little uncertainty about the values edited or imputed by the cold-deck and the deductive imputation procedure we decided to only delete the values imputed or changed by the hot-deck imputation procedure. In the variables that were not changed by the hot-deck procedure like "sex" and "date of birth" this approach results in complete variables.

29. On the other hand, for originally plausible values that were only changed because of the editing applied after the hot-deck imputation this results in pseudo missing values. The share of those pseudo missing values though is very small. For the variable WZ_ABS for example only 6.46% of the deleted values are pseudo missing values.

30. Table 1 shows the resulting proportions of missing values in the data set after and before cold-deck and deductive imputation for some selected variables.

Variable	Proportion of missing values	Proportion of missing values before cold-deck and deductive imputation
Occupation	8.00%	8.00%
Section of economic sector	3.77%	4.82%
Class level	0.65%	1.94%
Education graduation achieved	2.32%	2.93%
Highest education graduation	3.39%	5.32%
Highest apprenticeship graduation	3.29%	4.58%

Table1: Proportion of missing values

Obviously, for most variables the proportion of missing values is quite low. This proportion has been reduced further through the deductive imputation. The variable "occupation" contains the largest share of missing values. This is due to the difficult nature of plaintext data capture and the different imputation procedure applied.

B. The Method

31. When dealing with multivariate missing data one has to choose between 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, which is not imputed in this study, are categorial. 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 categorial values are needed. (Schafer, 1997) for example suggests rounding of the continuous imputed values but this can lead to biased results.

32. We decided to use 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 the implementation we are using the R-package MICE 2.22 (van Buuren and Groothuis-Oudshoorn, 2011).

33. One of the recommendations of (Münnich et al., 2014) was to favour a polytomous regression model over predictive mean matching when imputing categorial variables. Thus we started our tests with the MICE function *polyreg*. This is also the default method in MICE when dealing with factor variables. However, following the advice of (van Buuren, 2012) for variables with a high number of categories like HGKLDB we used the function *pmm* although this method is based on the normal linear regression model. But since predictive mean matching imputes the true value of the respondent with the closest predicted value only observed values are imputed and thus no rounding is necessary.

34. In MICE one can not only choose the form of the imputation model but also select the predictor variables for every model separately. In theory, if possible all other variables and their interactions should be included into the model in order to preserve the relationship. In practice, this is often not feasible because this can lead to huge models and thus computational problems, especially since categorical variables are split into dummy variables.

35. Therefore the used predictor variables have to be chosen carefully. Variables that are used in the intended analysis have to be included by all means because otherwise the results may be biased. Also, variables that are related to the nonresponse should be included. Those are variables for which the distributions of the response and nonresponse group differ or which are causative for the missing. Finally one should include variables that help to explain the variance.

36. MICE can handle both, *missing at random* (MAR) and *missing not at random* (MNAR) data. Data is *missing at random* when the missing probability only depends on observed information like other variables. On the other hand data is *missing not at random* when the missing probability also depends on the variable itself. For the sake of completeness: the third missing data model is *missing completely at random* (MCAR) for which the missing probability neither depends on the variable itself nor on other variables. Unfortunately, this assumption is often unrealistic in real survey data.

37. For the imputation of MNAR data one needs external information about the missing data mechanism since in this case the mechanism is not detectable through the data itself. So, to get started, we use the MAR assumption. If variables are included into the imputation model that are strongly related to the imputed variable one can yield good results also for MNAR because the strong relation between the variables makes the MAR assumption more plausible. Nevertheless, for variables where a MNAR assumption is realistic one should perform a sensitivity analysis of the results.

IV. First results

38. Our evaluation is still in a rather early, preparatory stage. Hence, this paper presents only a few, very preliminary results, and is rather meant to illustrate the approaches considered for the proceeding we have in mind in order to impute missing data with the regard to the above mentioned example table/hypercube. Note, that all results presented in this section are based exclusively on the data from one of the German States, e.g. the Land Berlin. Although in principle the desired goal is an imputation cycle that includes all variables at one time, for the preliminary studies we portion the variables into blocks.

39. In the beginning, for every variable certain decisions have to be made. First the imputation model has to be chosen. As already mentioned we are only dealing with categorial variables and thus we are starting with the functions *polyreg* and *pmm* for variables with a high number of categories. But there are several other functions available in MICE we can resort to in case of unexpected problems.

40. All variables used in the intended analysis always have to be included into the imputation models. Thus “sex”, HGKLDB and WZ_ABS are included as predictor variables in the models for the respective other variables. Since the considered analysis is a table of those variables all interactions would have to be included as well. The inclusion of the interaction between HGKLDB and “sex” would make the imputation model very large and lead to very long computing times. Therefore, during the current, preparatory stage of the study we did not include interactions even though in theory we should and we still need to examine the consequences of this relaxation.

41. Secondly, the missing data mechanism has to be examined and variables related to it have to be identified. For this purpose we are going to have a look at the distributions of the potentially related variables for the group with missing values and the group with observed values and to compare them. Figure 1 shows the kernel densities of “age” in the two groups of the variable WZ_ABS. The red solid line is the kernel density for respondents with observed WZ_ABS and the green dashed line for the ones where it is unobserved. The difference between the densities is not surprising since we did not delete any values changed by the deductive imputation. Thus for all respondents younger than 15 WZ_ABS corresponds a structural zero since this item was only required for people aged 15 and older. Therefore it was already set to “0”.

42. Another option to detect relationships is to calculate the correlation between a dummy variable marking the missing values of the variable of interest (1: missing, 0: observed) and the other variables. For this purpose, the idea is to use Cramer’s V. For example, the strongest relationship to the missing dummy of WZ_ABS observed so far results from the variable “current position”. This variable should therefore be included as a predictor variable into the imputation model for WZ_ABS.

43. The next step is to detect variables that help to explain variance. For this purpose one can again calculate correlations. We use again Cramer’s V, to measure this time the relationship between the observed parts of the variables. A first observation is that both variables of interest (HGKLDB and

WZ_ABS) have a strong relationship with “current position”. “Current position” on the other hand is strongly related to all variables concerning graduation. Those variables should therefore be included into the imputation models as well.

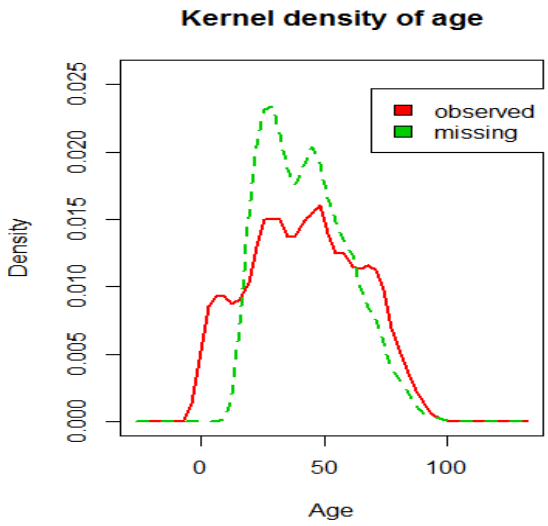

Figure 1: Kernel density of age for “section of the economic sector” (WZ_ABS) observed and missing

44. Although if the observed parts of two variables are strongly related this does not necessarily mean that they are useful predictors for each other. The usefulness also depends on the missing pattern of the two variables. In the worst case they are both missing for the same respondents and are not useful for the imputation of each other at all. A practical indicator to detect variables that contain only little information about the missing part of another variable is the proportion of usable cases. So the proportion of usable cases I_{jk} for imputing variable Y_j from variable Y_k is the number of pairs (Y_j, Y_k) with Y_j missing and Y_k observed, divided by the total number of missing cases in Y_j (Van Buuren, 2012).

45. The proportion of usable cases for imputing HGKLDB from WZ_ABS is rather poor, only 0.24. Thus for only 24% of the missing values in HGKLDB the WZ_ABS is observed. Generally, variables with a low proportion of usable cases should be excluded from the imputation model in order to keep it small. However, since in this case we need the relationship of the variables for our intended analysis we will have to keep it anyway.

46. After the selection of the imputation models one has to decide about the number of imputations *m*. We decided to begin with m=10 imputations, even though another recommendation of (Münnich et al., 2014) was to impute 50 times. So we will probably raise this number later on.

47. Another option in MICE is that one can change the visiting scheme of the algorithm. That means we can determine the order in which the variables are imputed. Although in our models we do not use any interactions as predictors, in which case it would be essential to impute the main effects and to derive the respective interaction before it is used as a predictor, in our experiments so far we changed the visiting scheme of the algorithm so that the variables of interest for our intended analysis are imputed after the other variables. Since the only editing rule for the table is that HGKLDB has to be a structural zero if WZ_ABS is a structural zero we impute WZ_ABS first. Afterwards we set HGKLDB to “0” for all cases were WZ_ABS was imputed as “0”.

48. After the imputation is done and this can take quite a while, if the models are large, the convergence of the MICE algorithm has to be examined. By default, MICE stores the mean and standard deviation of the imputations per iteration. Since our variables are categorial the mean and the variance are no adequate measures. For a meaningful analysis we need to replace them by other parameters of interest. But for a first inspection the default plots in mice can still be helpful to give some impression of the behaviour of the algorithm. Figure 2 shows the convergence plot for WZ_ABS. Each line represents one

of the 10 imputations. Since the streams are freely mixed with one another and there are no visible trends there is no evidence that the algorithm did not converge.

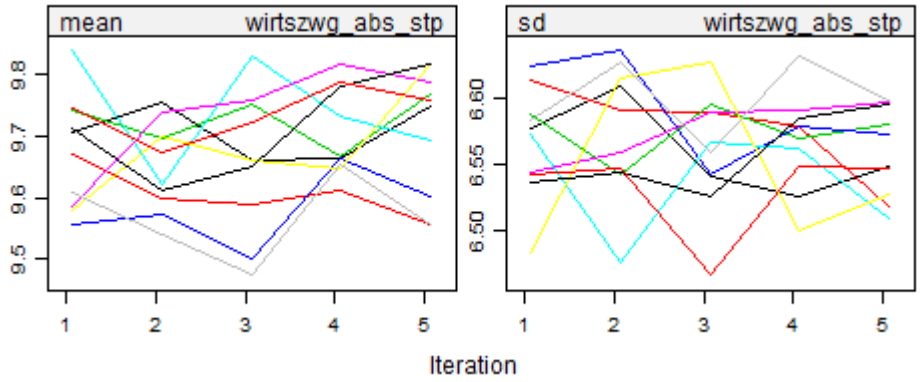

Figure 2: Convergence plot of "section of the economic sector"

49. Last but not least the imputations themselves have to be inspected. Imputations are good if they are plausible in the sense that they could have been real values if they had been observed. In order to assess the plausibility of imputations one could examine the distributions of the observed and imputed data. In case of MAR or MNAR the distributions cannot be expected to be identical. But serious discrepancies can be a hint that something went wrong and should be examined further.

50. Figure 3 shows the distributions of observed and imputed "section of the economic sector" (WZ_ABS).

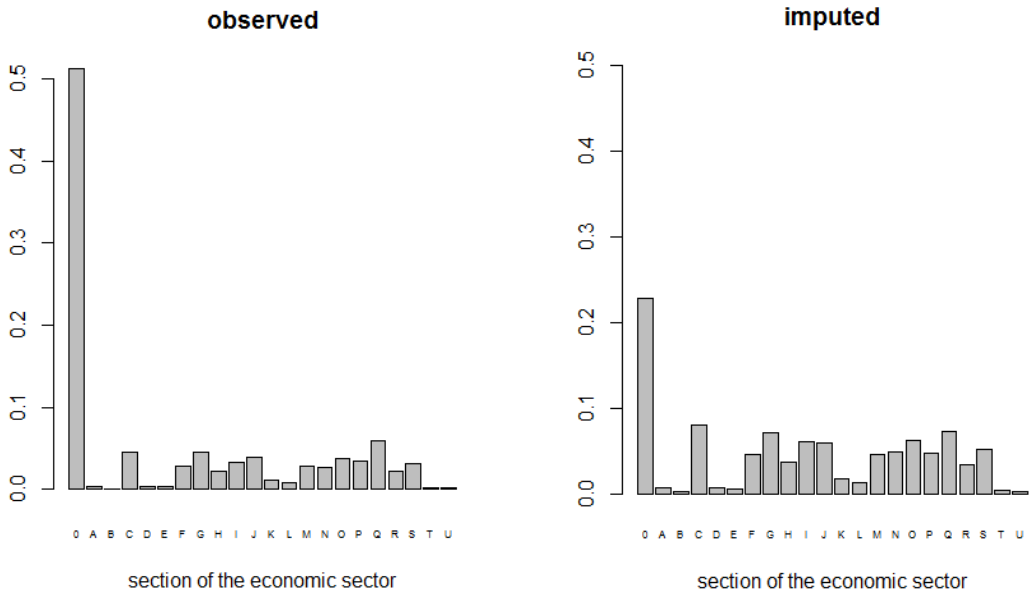

Figure 3: Distributions of observed and imputed "section of the economic sector" (WZ_ABS)

Besides the first category that corresponds to the structural zeros, the distributions look pretty similar. The difference in the fractions of the structural zeros however is in this case not an indicator that the imputation went wrong. Since the distribution of the observed group includes all persons younger than 15 for whom WZ_ABS is always a structural zero and the imputed group includes only persons aged 15 and older this difference is completely plausible.

51. Unfortunately, at this early stage of our study we cannot present any results regarding the general objectives of the study, e.g. concerning variance estimation and the impact of the imputation variance. The calculation of the survey estimates and the respective error estimation still has to be implemented. Only after that work is done we will be able to evaluate the influence of the imputation variance on the accuracy of the estimates.

V. Summary and Outlook on Future Work

52. This paper has presented the intended procedure we plan to execute on the whole population census. We started with a simple example but developing the final procedure will be much more complicated. Especially the implementation of the compliance of all edit rules needs further studies and effort. One open question is to what degree the so far applied shoot-out procedure that just replaces implausible imputations by a plausible value distorts the distribution of the variable. And what happens if no deductive imputation for implausible imputations is possible and one has to impute implausible values over and over again?

53. Although MICE is a very flexible tool for multiple imputation for our purposes we probably will have to expand its functionality. First of all the convergence plots have to be adjusted so that proper parameters are traced. Another issue is that the possibilities to correct or re-impute implausible imputations within MICE are very limited.

54. Also concerning the models further work is needed. Especially the bias of the results due to the negligence of interaction effects has to be examined. Of course the best solution is to include all interaction effects but this could result in models that can probably not be computed in reasonable time any more. But computation time is an unsolved issue in our study anyway so yet it is not clear if we can use the MICE algorithm for the imputation of the whole data set at all.

55. But since the main purpose of the study is to assess the impact of the imputation variance on the accuracy of the survey estimates the most urgent task is to implement the calculation of the survey estimates and the variance estimation. Only then we can decide, if it is actually necessary to account for the variance due to imputation and hence to apply multiple imputation, or if ignoring the imputation variance might be justified for the future studies on suitable estimation methods for the next Census round.

References

Berg, A.; Bihler, W. (2014). Das Hochrechnungsverfahren zur Ermittlung der Einwohnerzahl im Zensus 2011. *Wirtschaft und Statistik*, April 2014, Wiesbaden: Statistisches Bundesamt, URL: https://www.destatis.de/DE/Publikationen/WirtschaftStatistik/Zensus/HochrechnungsverfahrenEinwohnerzahl_42014.pdf

Davison, A.; Sardy, S. (2004). Resampling Methods for Variance Estimation / Dacseis Project - Report

Münnich, R.; Gabler, S.; Ganninger, M.; Burgard, J. P.; Kolb, J.-P. (2011). Das Stichprobendesign des registergestützten Zensus 2011. *Methoden - Daten –Analysen*, 2011, Jg. 5 Heft 1

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

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

R Core Team (2013). 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.

Shao, J.; Sitter, R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association, Vol.* 91, No. 435, pp. 1278-1288

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