Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and emerging methods

Simulating Multiple Imputation of Water Consumption in the German Agricultural Census 2010

Prepared by Lydia Spies, Sven Schmiedel, Katrin Schmidt; Federal Statistical Office, Germany

I. Introduction

1. Missing values due to nonresponse or implausible data are a problem every statistical office has to deal with in almost any survey. A simple and therefore very popular missing data method is to discard units with incomplete information. Although complete-case analysis is used by default in many statistical programs it is a poor way of handling missing values. Beside the potentially huge loss of information resulting in reduced statistical power it can produce seriously biased results if the respondents are systematically different from the non-respondents.

2. A much smarter missing data technique is imputation. Missing items are replaced with plausible values predicted by using the information contained in the observed items. To take account of the missing data uncertainty Rubin (1978, 1987) proposed to replace every missing datum by $m > 1$ plausible values. Each of these m resulting data sets is then analyzed separately and the results are combined by simple arithmetic. The consideration of the between imputation variance of the estimates allows valid variance estimation.

3. As the decision about publishing an estimated value in the Federal Statistical Office of Germany depends on its estimated coefficient of variation we favor multiple imputation over other methods. There are several different techniques how to do the repeated single imputations. In the German Agricultural Census 2010 for example Schreiner, Schmidt (2011) used a Bayesian linear regression model with a hot-deck component. Although the missing data scenario of this survey is very simple with only one incomplete variable and a relatively low rate of missing values (16%), it is nevertheless very complex and time consuming to find the “best” model. In practice another difficulty is that we do not know the true data. Thus it can be difficult to compare the performance of different approaches. While multiple imputation in theory leads to unbiased estimates, in practice also small deviations from model assumptions can cause bias.

4. That is why we implemented a software tool in R (2013) for a repeated simulation of missing values in complete-case data sets, multiple imputing them with different techniques and comparing the combined results. We hope this tool for simulating different types of nonresponse and for the comparison of the performance of different approaches will accelerate and ease the development and testing of sophisticated imputation methodologies at the Federal Statistical Office of Germany.

5. This paper presents our first test application to water consumption data of the German Agricultural Census 2010. In Section 2 the imputation methods and the test data are introduced. The simulation study is presented in Section 3. In Section 4 the main results are summarized. The paper finishes with a discussion section.

II. Methods and Test Data

6. As already stated above multiple imputation creates $m > 1$ complete data sets. No matter which approach is used to do the repeated single imputations the calculation of the overall estimates is always the same. After multiple imputing the missing values the *m* data sets are analyzed separately by complete-data methods. If we want to estimate a population quantity *Q* like the population mean we calculate the sample means $\hat{Q}_{*l}$ and the standard errors $\sqrt{U_{*l}}$, $l = 1,\ldots,m$. The *m* results are then combined by Rubin´s rules to obtain overall estimates and standard errors (see Rubin (1987)).

7. The overall estimate is simply the average of the *m* estimates $\bar{Q}_m = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_{*l}$. The overall standard error is the square root of the total variance T_m which is composed of the within-imputation variance $\bar{U}_m$ and the between-imputation variance B_m. The within-imputation variance is the average of the *m* variances $\bar{U}_m = \frac{1}{m}\sum_{l=1}^{m} U_{*l}$ and the between-imputation variance is the variance of the *m* estimates $B_m = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_{*l} - \bar{Q}_m)^2$**.** The total variance is calculated by $T_m = \bar{U}_m + (1 + \frac{1}{m})B_m$. A confidence interval for *Q* can be formed using a *t*-distribution with $v = (m-1)(1 + \frac{\bar{U}_m}{(1+\frac{1}{m})B_m})^2$ degrees of freedom.

8. In case of a census like the German Agricultural Census 2010, e.g. a full survey - not a sample survey, the within standard errors of the estimates are zero and the total variances equal the between-variances.

A. Methods

9. We applied two approaches for imputation of the missing data, both based on Bayes statistics. First, we implemented an imputation realized by draws out of the posterior predictive distribution (Rubin (1987), chapter 5.3). This distribution is represented by the following formula $y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and $\mathbf{X}\boldsymbol{\beta}$ is a linear regression model.

10. For $\boldsymbol{\theta}$ an improper prior distribution (the function used as a prior has an infinite integral and is thus not a probability density) is used. The slopes $\boldsymbol{\beta}$ are estimated by the completely observed data

$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_{obs}^t \mathbf{X}_{obs})^{-1} \mathbf{X}_{obs}^t \mathbf{Y}_{obs}$. The variance of the observed data $\hat{\sigma}_{obs}^2$ is calculated by $\hat{\sigma}_{obs}^2 = \frac{\hat{\boldsymbol{\varepsilon}}_{obs}^t \hat{\boldsymbol{\varepsilon}}_{obs}}{(n_1 - q)}$, where n_1 is the number of respondents and q is the number of covariates in the linear regression model.

11. For calculating the posterior variance σ_*^2, g is drawn from a chi²-distribution ($g \sim \chi_{n_1-q}^2$) and $\sigma_*^2 = \hat{\sigma}_{obs}^2 (n_1 - q)/g$. Afterwards q independent draws from $N(0, 1)$ are combined into a vector $\mathbf{z}$ and the slopes $\boldsymbol{\beta}_*$ of the posterior are computed by $\boldsymbol{\beta}_* = \widehat{\boldsymbol{\beta}} + \sigma_* [\mathbf{X}_{obs}^t \mathbf{X}_{obs}]^{-1/2} \mathbf{z}$, where the triangular square root $[\mathbf{X}_{obs}^t \mathbf{X}_{obs}]^{1/2}$ is calculated via Cholesky factorization. Then the missing values of y are calculable as $y_{i_{mis}} = \mathbf{X}_{i_{mis}} \boldsymbol{\beta}_* + z_i \sigma_*$, where $\mathbf{X}_{i_{mis}}$ represents the covariates of survey respondents where y is missing.

12. In the procedure we used, the random component $\mathbf{z}$ of the residuals is replaced by the "hot-deck" component of the standardized residuals $\hat{\boldsymbol{\varepsilon}}_{*,sd}$. Hot-deck means that residuals from the set of observed residuals $\hat{\boldsymbol{\varepsilon}}_{obs}$ are chosen randomly. Hence, $y_{i_{mis}} = \mathbf{X}_{i_{mis}} \boldsymbol{\beta}_* + \hat{\boldsymbol{\varepsilon}}_{*,sd} \sigma_*$.

13. Beside the first method, which in the following will be referred to as "hot-deck", the second implemented method was predictive mean matching (PMM) (Rubin (1987)). This approach does not use the hot-deck component. The value is estimated by $y_{i_{mis}} = \mathbf{X}_{i_{mis}} \boldsymbol{\beta}_*$. Thereafter the nearest observed value is used. Thus, only values that are observed are imputed.

14. All described methods were used to impute multiple, in our case 200 times.

B. The Test Data

15. The database consists of 14 213 farms from the German agricultural census 2010 ("original data"). The water consumption in m³ is the only incomplete item. Missing data were observed on 2 088 records. Furthermore, we deleted 213 observations because of implausible values (value of water consumption in m³/irrigated land in m² greater than 10 000 or less than 10). Overall, 2 301 (16 %) values were missing.

16. Further covariates were:

- Irrigated acreage by 18 crop types: on metric scale in m²
- Climatic water balance: on metric scale in millimeter, comparison of evaporation and precipitation,
- Usable field capacity: on ordinal scale in categories, a value which indicates how much water a soil can buffer
- Irrigation method: sprinkler and drip irrigation
- Water source: nominal scale, states the origin of the used water

III. The Simulation Study

17. Our simulation is based on the following steps. First, the cases with missing data were deleted from the original data. This was the database for the simulation study and in the following will be referred to as

“complete data”. In this database we randomly generated missing values and used the resulting data sets as test sets for the imputation methodologies. This way, it was possible to do a proper comparison of imputation result and true data.

A. Missing pattern

18. As mentioned in the introduction in our scenario we only have to deal with the simplest missing pattern. Only for one variable values are missing. The performance of the multiple imputation methods was investigated by using the ”complete data” and generating missing values in the data by the following method.

19. In order to simulate a missing at random (MAR) scenario, we determined the probability of a value to be missing by a logistic regression. In the original dataset the farms with a missing value for the water consumption are coded with “1”, all others with “0”. The logistic regression model with all other variables as covariates determines the probability for the item nonresponse. This probability is used, in order to delete values randomly. In every simulation, the average proportion of missing data in the reduced dataset is (2 301/14 213 =) 16.19 %.

B. The model

20. In order to impute the missing values we set up a regression model. We observed that the acreage of irrigated farmland on log scale explains about 70 % of the variance in the data. Therefore the overall acreage of irrigated farmland was used adding the crop types as dummy variables. We included all other available covariates (climatic water balance, usable field capacity, irrigation method, water source) in the model. Because the distribution of water consumption in m^3 (and the respective residuals in the regression model) did not show a normal distribution we compared the natural logarithm and the cubic root transformations in the simulation.

C. The implementation

21. An extreme result is possible when the simulation is applied only once. When many repetitions of the simulation are used, one receives also the distribution of the quality measures and ensures that the results are not extreme by chance. Therefore, we implemented a bootstrap procedure which chooses 1 000 times 1 000 farms with replacement from the “complete data”. In the so called “complete sample data” the item nonresponse was then randomly generated, as explained above. Afterwards we imputed these artificial missings 200 times and calculated quality measures. The sample data without the cases with randomly generated item nonresponse will be referred to as “complete-case data”.

D. The quality measurements

22. Main focus of statistical data production are tables of point estimates (means and totals). In our example those tables present the data by

- size ranges of water consumption,
- the 16 German federal states,
- size ranges of acreage.

23. However, the calculation of these tables during the search of the best imputation method is not feasible as the variance in the tables is too broad because only 1 000 farms were used in each imputation step and the sample was drawn completely at random and not stratified. We only calculated these tables therefore in the final analysis step where the imputation method considered as "best method" was used for a "final" imputation in the "complete data". In order to decide which imputation method is best we computed the root mean square error (RMSE).

24. During the search of the best method the bias (average estimate compared to the true value) and the coverage rate for the overall sample mean were considered. The coverage rate is the percentage of how often the true value of mean water consumption is found in the 95 % confidence intervals (CI) of the imputations. We also considered the average CI range and at the between and within variance.

25. We calculated those measures of goodness for the different imputation approaches, the "complete sample data" and the "complete case data".

IV. Results

26. Table 1 presents the results after $m = \{5, 20, 200\}$ imputations. It shows the median results after 1 000 simulation loops. The minimum has negative values for some of the hot-deck methods. This is possible if the randomly chosen hot-deck residual is very large. We found the smallest median bias (34.0) for $m = 5$ imputations without transformation using the PMM method.

27. In figure 1 we see that the complete case analysis clearly overestimates the mean. This is reasonable as the probability for a value being missed depends on the acreage of a farm which is an indicator for the size of the enterprise. Large enterprises might use a counting device which leads to an exact measure for water consumption. The likelihood for a small farm to show a missing value is therefore higher as for a larger farm which leads to this overestimation. For all other methods a good estimation of the overall mean can be found, except for the hot-deck method with log transformation.

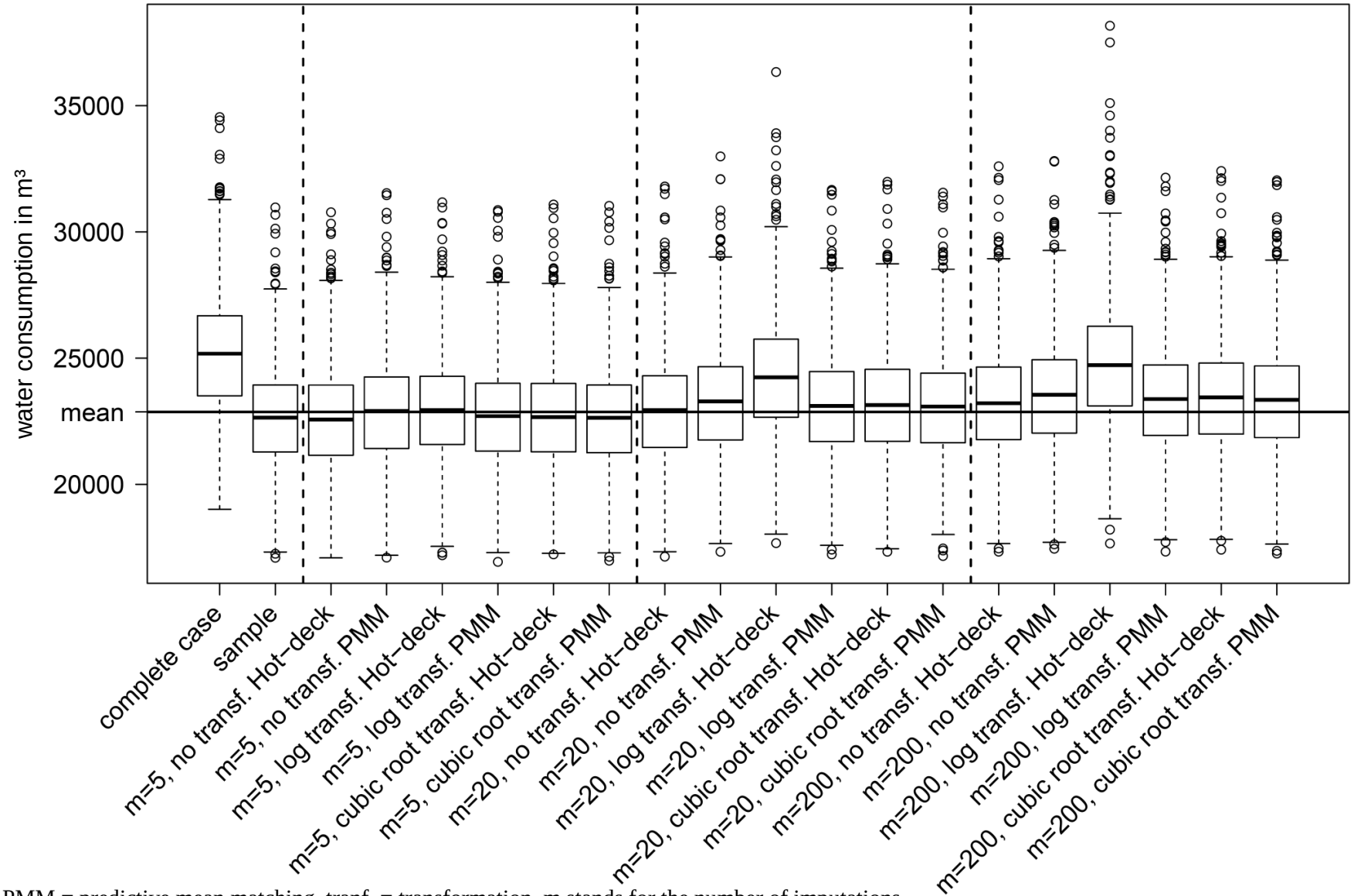
28. Furthermore, one can observe that the estimate for the mean has a positive bias which is even more pronounced the more imputations were used. This finding was not altered by further increasing m (data not shown). Because we were not sure, if the variance is stable after 200 imputations, we looked at the development of the variance after 1 000 imputations (Figure 2). We see that the total variance is still very instable even after 200 imputations.

Table 1: Results of the multiple imputation after 1 000 repetitions, measures show the median value of the 1 000 imputations

imputation approach	minimum	25 % quan-tile	median	75 % quantile	maximum	mean	standard devi-ation	bias	coef. var.	confidence interval	width of CI	cover-age in %	RMSE	within var.	between var.
complete data	.	600.0	3 674.0	19 700.0	.	22 870.0	-	-	-	-	-	-		-	-
sample data (before deletion)	3.0	600.0	3 700.0	19 630.0	853 000.0	22 645.0	1 855.1	-225.0	8.2	[18 838.3 , 26 416.5]	7 272.0	92.0	1868.7	-	-
complete case (before imput.)	4.0	700.0	4 209.5	22 610.0	853 000.0	25 175.0	1 979.6	2 305.0	7.9	[21 053.1 , 29 111.5]	7 759.7	84.8	3038.4	-	-
m=5, id, hot-deck	-210 350.0	500.0	4 374.0	21 860.0	853 000.0	22 566.0	1 963.1	-304.0	8.7	[18 506.0 , 26 539.3]	7 706.9	92.7	1986.5	3 614 092.8	179 700.0
m=5, id, PMM	4.0	614.0	3 950.0	20 000.0	853 000.0	22 904.0	1 870.8	34.0	8.2	[19 021.7 , 26 625.5]	7 334.9	91.6	1871.1	3 439 240.2	37 700.0
m=5, log, hot-deck	1.3	600.0	3 612.0	19 415.0	1 069 000.0	22 945.0	1 982.9	75.0	8.6	[18 853.9 , 27 013.6]	7 807.6	93.6	1984.3	3 681 975.8	123 255.0
m=5, log, PMM	4.0	600.0	3 650.0	19 520.0	853 000.0	22 702.0	1 871.6	-168.0	8.2	[18 790.4 , 26 415.7]	7 336.7	92.2	1879.1	3 439 051.4	37 250.0
m=5, cubic root, hot-deck	-6 720.5	600.0	3 900.0	20 000.0	853 000.0	22 664.0	1 854.0	-206.0	8.1	[18 869.8 , 26 458.5]	7 267.6	91.5	1865.4	3 406 586.9	21 070.0
m=5, cubic root, PMM	4.0	600.0	3 622.5	19 640.0	853 000.0	22 636.0	1 862.4	-234.0	8.2	[18 788.3 , 26 370.2]	7 300.4	91.5	1877.0	3 423 684.5	22 075.0
m=20, id, hot-deck	-226 500.0	500.0	4 326.8	23 212.5	853 000.0	22 943.8	2 104.8	73.8	9.2	[18 567.3 , 27 287.7]	8 263.8	94.3	2106.1	3 840 366.2	599 526.2
m=20, id, PMM	4.0	600.0	3 875.0	20 000.0	853 000.0	23 285.5	1 989.5	415.5	8.5	[19 179.3 , 27 299.5]	7 801.3	93.3	2032.4	3 597 405.1	272 955.1
m=20, log, hot-deck	0.7	600.9	3 591.5	19 447.5	4 856 500.0	24 241.0	3 124.7	1 371.0	13.0	[17 503.4 , 30 442.8]	12 341.1	98.6	3412.2	6 853 478.4	2 709 967.8
m=20, log, PMM	4.0	600.0	3 737.8	20 000.0	853 000.0	23 105.0	1 990.0	235.0	8.6	[19 000.0 , 27 155.0]	7 812.9	93.4	2003.9	3 618 156.0	265 504.6
m=20, cubic root, hot-deck	-10 590.0	600.0	3 997.3	20 000.0	982 450.0	23 139.8	1 977.5	269.8	8.5	[19 100.6 , 27 103.0]	7 754.1	93.6	1995.8	3 557 221.4	264 641.1
m=20, cubic root, PMM	4.0	600.0	3 724.0	20 000.0	853 000.0	23 082.0	1 986.8	212.0	8.6	[19 003.4 , 27 050.3]	7 788.9	93.5	1998.1	3 582 092.2	256 757.2
m=200, id, hot-deck	-231 600.0	500.0	4 398.3	23 915.0	900 000.0	23 214.3	2 141.2	344.3	9.2	[18 811.5 , 27 627.0]	8 394.5	95.1	2168.7	3 912 276.9	677 083.2
m=200, id, PMM	4.0	600.0	3 930.5	20 000.0	853 000.0	23 548.7	1 991.2	678.7	8.5	[19 411.8 , 27 531.4]	7 805.6	93.7	2103.7	3 639 247.8	318 334.7
m=200, log, hot-deck	0.2	615.0	3 676.8	19 775.0	18 215 000.0	24 722.5	4 114.7	1 852.5	16.6	[16 272.3 , 33 126.9]	16 146.2	99.8	4512.5	10 790 111.8	6 226 763.0
m=200, log, PMM	4.0	600.0	3 794.5	20 000.0	853 000.0	23 380.0	2 001.7	510.0	8.6	[19 239.8 , 27 464.4]	7 847.1	93.8	2065.6	3 672 670.5	310 483.6
m=200, cubic root, hot-deck	-22 110.0	608.2	4 000.0	20 047.5	1 307 500.0	23 445.5	1 987.9	575.5	8.5	[19 337.6 , 27 411.9]	7 792.6	93.8	2069.5	3 620 329.9	295 540.6
m=200, cubic root, PMM	4.0	600.0	3 785.0	20 000.0	853 000.0	23 349.3	1 995.9	479.3	8.5	[19 199.1 , 27 346.7]	7 824.3	93.8	2052.6	3 652 152.6	290 791.5

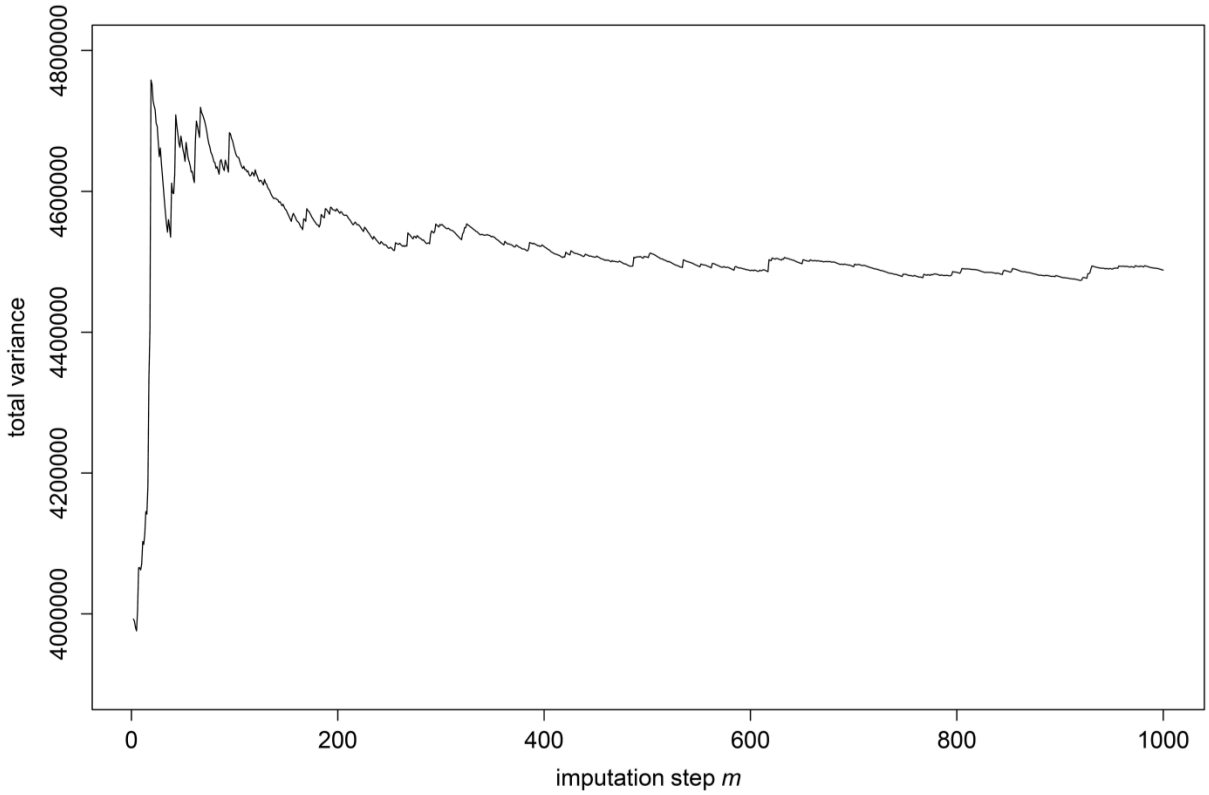
id = identity, PMM = predictive mean matching, coef. var. = coefficient of variation, CI = confidence interval, RMSE = root mean square error var. = variance,
. = deleted due to confidentiality, - = value not available

Figure 1: Boxplots of all imputation approaches, including the complete case analysis (data after deletion of cases with artificial missing) and the analysis of each sample (data before deletion of missings)



PMM = predictive mean matching, tranf. = transformation, m stands for the number of imputations

Figure 2: Total variance development by number of imputations



29. As the PMM method showed the lowest RMSE for 200 imputations we decided to use this model with the cubic root as transformation for a "final" imputation in the "complete data".

30. In tables 2-4 we see the results for the final imputation of the complete data. Here too the missing values for the water consumption variable were artificially generated and those missing data were imputed 200 times.

31. For the tabulation by water consumption ranges (table 2) we see that the simulated mean always exceeds the true mean. At the same time the coefficient of variation is always far below 15 % which is the threshold at the Federal statistical office of Germany for publication. Hence, all these results would be considered as proper for publication.

Table 2: Water consumption in size ranges, result for the imputation in the original data

water consumption in m³	farms	mean	simulation mean	bias	s.d.	coefficient of variation
0- <2 000	4 831	581.9	642.9	61.0	16.3	2.5 %
2 000- <5 000	1 640	3 170.6	3 466.2	295.6	18.6	0.5 %
5 000- <10 000	1 287	7 096.8	7 588.8	492.0	34.9	0.5 %
10 000- <20 000	1 187	13 890.9	14 869.2	978.3	89.7	0.6 %
20 000- <50 000	1 511	32 363.1	33 451.4	1 088.3	136.9	0.4 %
50 000- <100 000	822	70 708.0	72 743.5	2 035.5	284.1	0.4 %
≥ 100 000	634	207 829.9	213 856.7	6 026.8	2 577.7	1.2 %

s.d. = standard deviation

32. The result for the tabulation by size of acreage (table 3) is different. We find a positive bias in smaller size ranges whereas it turns out to be negative in higher ones. Here we see that only the first class has a large coefficient of variation, however, still below the 15 % limit.

Table 3: Farmland in hectare in size ranges, result for the imputation in the original data

Farmland in hectare	farms	mean	simulation mean	Bias	s.d.	coefficient of variation
< 5	2 202	1 453.2	2 486.1	1 032.9	350.9	14.1 %
5 - <10	1 041	3 771.2	4 570.7	799.5	487.8	10.7 %
10 - <20	1 371	4 931.4	5 952.5	1 021.1	491.7	8.3 %
20 - <50	2 160	10 119.2	11 136.5	1 017.3	482.1	4.3 %
50 - <100	2 317	21 447.4	22 714.7	1 267.3	479.3	2.1 %
100 - <200	1 628	46 105.0	45 889.2	-215.8	604.3	1.3 %
200 - <500	599	98 370.4	97 291.5	-1 078.9	911.8	0.9 %
500 -< 1 000	143	126 895.4	124 800.3	-2 095.1	1 755.5	1.4 %
> 1 000	156	172 545.1	170 611.8	-1 933.3	1 413.3	0.8 %

s.d. = standard deviation

33. For the federal states (table 4) we find that the coefficient of variation is above 15 % for four federal states.

Table 4: Results for the Federal States, result for the imputation in the original data

Federal State	farms	mean	simulation mean	bias	s.d.	coefficient of variation
1	580	5 955.5	7 147.5	1 192.0	840.8	11.8 %
2	223	2 083.6	3 204.4	1 120.8	737.7	23.0 %
3	3 668	43 780.2	44 230.9	450.7	380.8	0.9 %
4	4	1 788.8	1 788.8	0.0	0.0	0.0 %
5	1 369	7 923.4	8 521.6	598.2	501.0	5.9 %
6	654	20 253.6	22 595.7	2 342.1	1 063.2	4.7 %
7	823	25 393.3	24 548.7	-844.6	677.6	2.8 %
8	2 123	4 776.5	5 727.6	951.1	474.1	8.3 %
9	1 498	4 849.0	5 915.6	1 066.6	458.6	7.8 %
10	38	3 759.5	4 574.6	815.1	1 478.0	32.3 %
11	13	2 565.8	4 195.6	1 629.8	3 994.8	95.2 %
12	346	43 423.3	44 597.7	1 174.4	979.1	2.2 %
13	140	101 017.4	98 790.3	-2 227.1	1 627.9	1.6 %
14	198	6 431.2	7 616.3	1 185.1	1 432.5	18.8 %
15	175	78 927.8	78 570.6	-357.2	1 459.6	1.9 %
16	60	18 056.4	18 388.2	331.8	1 026.1	5.6 %

s.d. = standard deviation

V. Discussion

34. We analyzed the results of multiple imputation by a complex investigation, which is usually not possible in the daily routine of a statistical office. When evaluating data from an agricultural census (which means all farms of the whole country are included) with 16 % missing values one might be

tempted to think that using only the complete cases to compile a statistic is sufficient. Our study shows that this analysis would lead to large overestimation of water consumption.

35. This overestimation is still present with multiple imputation, however, on a much lower level. Hence, by using multiple imputation one gets much better results. The PMM method is an attractive method because it results in “realistic” imputed values and shows stable results without any extreme imputations, which is likely to happen with the hot-deck approach. Even though those extreme imputations on some individual records may not harm the estimates, it is difficult to explain a user of a statistic a negative value as a minimum where only positive values are possible.

36. A noticeable result of our simulation is that the total variance estimate has still not completely converged after 200 imputations. This needs further investigation.

References

Rubin, D.B. (1978). Multiple imputation in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34

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

Schreiner, C., Schmidt, K. (2011). Nacherhebung Bewässerung zur Landwirtschaftszählung 2010. *Wirtschaft und Statistik*, Dezember 2011, Statistisches Bundesamt, 1202-1211, URL https://www.destatis.de/DE/Publikationen/WirtschaftStatistik/LandForstwirtschaft/BewaesserungLandwirtschaftszaehlung2010.pdf

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/.