

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Vienna, Austria, 21-23 April 2008)

Topic (iv): New and emerging methods

**Evaluating Different Approaches for Multiple Imputation Under Linear Constraints**

**Invited Paper**

Prepared by Joerg Drechsler, Institute for Employment Research, Germany and  
Trivellore E. Raghunathan, University of Michigan

**I. INTRODUCTION**

1. Missing data is a common problem in surveys. To avoid information loss by using only completely observed records, several imputation techniques have been suggested. In this paper we focus on multiple imputation (Rubin, 1978), an approach that retains the advantages of imputation while allowing the uncertainty due to imputation to be directly assessed. With multiple imputation, the missing values in a dataset are replaced by  $m > 1$  simulated versions, generated according to a probability distribution for the true values given the observed data. More precisely, let  $Y_{obs}$  be the observed and  $Y_{mis}$  the missing part of a dataset  $Y$ , with  $Y = (Y_{mis}, Y_{obs})$ , then missing values are drawn from the Bayesian posterior predictive distribution of  $(Y_{mis} | Y_{obs})$ , or an approximation thereof. Raghunathan et al. (2001) developed an iterative algorithm called sequential regression multiple imputation (SRMI) based on the ideas of Gibbs sampling to avoid otherwise necessary assumptions about the joint distribution of the missing data given the observed data. Imputations are generated variable by variable where the missing values for any variable  $Y_k$  are imputed for the conditional distributions of  $(Y_k | Y_{-k})$ , where  $Y_{-k}$  represents all variables in the dataset except  $Y_k$ . This allows for different imputation models for each variable. Continuous variables can be imputed with a linear model, binary variables can be imputed using a logit model, etc. Under some regularity assumptions iterative draws from these conditional distributions will converge to draws from the joint multivariate distribution of the data.

2. However, none of the multiple imputation approaches in the literature deals with the problem of imputations under linear constraints. Many surveys contain variables that have to add up to a given total. For example in establishment data the total number of employees will be the sum of the number of employees with different qualification levels, the turnover reported for different regions will have to add up to the total turnover, or the investment can be divided into different subcategories. Sometimes variables are even subject to multiple constraints, e.g. if the number of employees with different qualification levels is further divided into male and female employees. In this paper we will focus only on imputation methods for variables that are subject to one linear constraint. In a simulation study based on data from the German IAB Establishment Survey, we will evaluate five different multiple imputation approaches that address the linear constraints in different ways.

3. The remainder of the paper is organized as follows: Section II describes the data used for the simulation study. Section III introduces the five different imputation approaches. In Section IV the simulation design is illustrated. Section V presents the simulation results. The paper concludes with some final remarks and ideas for future research.

## II. LINEAR RESTRICTIONS IN THE IAB ESTABLISHMENT PANEL

4. The IAB Establishment Panel contains detailed information about German firms' personnel structure, development, and policy. For the first wave, 4,265 establishments were interviewed in Western Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually – since 1996 with over 4,700 establishments in Eastern Germany in addition. In 2007 more than 16,000 establishments participated in the survey.

5. Linear restrictions have to be fulfilled for variables concerning the employment structure and the turnover in different regions. In our simulation study we focus on a set of questions on the personnel structure. The linear restriction is given by the following equation:

$$Y_t = Y_{work} + Y_{train} + Y_{exec} + Y_{own} + Y_{marg} + Y_{other} \quad (1)$$

The total number of employees ( $Y_t$ ) has to be equal to the sum of the following six subcategories: The number of blue and white collar workers ( $Y_{work}$ ), the number of trainees ( $Y_{train}$ ), the number of executives ( $Y_{exec}$ ), the number of owners and working family members ( $Y_{own}$ ), the number of marginal workers not covered by social security ( $Y_{marg}$ ) and the number of other workers ( $Y_{other}$ ). Additionally, all variables have to be nonnegative.

6. For our simulation we delete all observations with missing values, treating the remaining 11536 units as the population. Table 1 contains some (unweighted) summary statistics for the seven variables included in the balance restrictions. Some features of the data are worth noticing: First, as expected, the distribution of the variables is heavily skewed. Second, most variables are semi-continuous, that is many establishments don't have any employees in certain subcategories. Especially for the number of executives and the number of other employees only 729 and 555 establishments respectively have any employees in these subcategories. Third, not surprisingly, with a maximum of 21 there is only little variation in the number of owners and working family members. And finally, most of the employees are workers, leaving only a small proportion of the total number of employees for the other five subcategories.

**Table 1: Summary statistics for the seven variables under linear constraints**

	<i>Min.</i>	<i>1st Quart.</i>	<i>Median</i>	<i>Mean</i>	<i>3rd Quart.</i>	<i>Max.</i>	<i>nb of obs != 0</i>
total nb of emp	1	6	19	128.6	79	22920	11536
workers	0	3	14	109.9	66	19410	11211
trainees	0	0	0	6.101	3	1552	5232
executives	0	0	0	6.124	0	6323	729
owners	0	0	0	0.6667	1	21	5735
marginal workers	0	0	1	5.413	3	2492	5772
others	0	0	0	0.4577	0	566	555

## III. FIVE DIFFERENT MULTIPLE IMPUTATION APPROACHES

7. We consider five different approaches for the imputation under linear constraints. For the first three methods missing values are imputed ignoring the linear constraint and an additional correction step after the imputation procedure guarantees consistency between the variables. The last two approaches include the additional constraint in the imputation model. Before imputation all continuous variables are transformed by taking the cubic root to get rid of the skewness in the data. All methods are described in this section:

## A. Simple Imputation

8. For this approach all variables are imputed independently variable by variable. All models are based on linear regressions. To avoid uncongeniality problems (Meng, 1994) all available information is included, i.e. all variables in the survey are considered as potentially explanatory variables. Due to multicollinearity problems, some variables have to be removed from some of the models, however. Still the regressions are based on roughly 100 explanatory variables for most models. Only for the number of executives and the number of other workers that contain only a small number of nonzero observations, the models are reduced to roughly 20 explanatory variables. Depending on the number of units with nonzero values, the variables are stratified by Eastern and Western Germany and two categories for the establishment size, i.e. up to 4 different regressions are modelled per variable. If negative values are imputed for some units, new values are drawn from the posterior predictive distribution for these units, until a positive value is imputed. If no positive value has been imputed after 200 draws, the value is set to zero.

9. To get consistent estimates for each record, some corrections are applied after the imputation process. If the total number of employees was originally missing and the imputed value for the total is less than the sum of the originally observed values of the subcategories, the total is set equal to the sum of the subcategories. (We investigated an approach, where we forced each imputed value for the total to be higher than the sum of the observed subtotals, but this led to an overestimation of the totals). If the total number of employees originally was observed, or if the imputed value of the total is higher than the sum of the observed subcategories, we weight every imputed subcategory  $Y_{imp}^{(i)}$  with the factor  $\frac{Y_t - \sum Y_{obs}^{(i)}}{\sum Y_{imp}^{(i)}}$ ,

where  $Y_{obs}^{(i)}$  are the observed subcategories.

## B. Independent imputation

10. This approach is almost similar to the first approach, with the difference that we address the problem of semi-continuity directly. The imputation is performed in two steps for each variable. In the first step it is decided whether the missing value is zero or not. For that, missing values are imputed using a logit model with outcome 1, if unit  $j$  has a positive value for that variable. In the second step a linear regression model based on the observed positive values is used to predict the actual value for the units with a predicted positive outcome in step one. We use the same imputation models and correction methods as in the simple imputation approach.

## C. Nested Imputation of proportions

11. The idea behind this method is related to the imputation approach described by Schenker et al. (2006). All subcategories are expressed as the proportion of the total, i.e. all values for the subcategories are divided by the total and thus are bounded between zero and one. A logit transformation of the variables guarantees that the variables will have values in the full range  $]-\infty, \infty[$  again. Missing values for these transformed variables can be imputed with a linear regression. To avoid problems on the bounds of the proportions, the two step imputation approach described in B is used to determine zero values and proportions greater than 0.999999 are set to 0.999999 before the logit transformation. After the imputation all values are transformed back to get proportions again and finally all values are multiplied with the totals to get back the absolute values.

12. Since the distribution of the proportions across the subcategories is heavily skewed – the number of workers is by far the biggest share of the total for almost all units -, the estimation for the other categories with very small proportions didn't provide good results. For that reason we use a nested imputation approach. That is, the number of workers is imputed as described above, but for all other variables the proportions are defined as the proportion of the total number of employees subtracted by the number of workers  $p^{(i)} = \frac{Y^{(i)}}{Y_t - Y_{work}}$ . These proportions are more equally distributed for the remaining

variables and we obtain far better results with this approach. After the imputation process all variables are corrected again to get consistent estimates as described in A.

#### D. Non-Bayesian Dirichlet Approach

13. This approach follows an idea by Tempelman (2007). Again, we calculate the proportions of the total for each subcategory, but now we assume that these proportions follow a Dirichlet distribution. To generate initial values for the imputation procedure we use the EM-Algorithm. The imputation algorithm for this procedure is based on the data augmentation algorithm by Tanner and Wong (1987). First we generate new values for the parameter  $\alpha$  by drawing from  $\alpha | Y_{obs}, Y_{mis} \sim N(\hat{\alpha}, \hat{V}(\hat{\alpha}))$  (P-Step), where  $\hat{\alpha}$  and  $V(\hat{\alpha})$  are obtained by ML-estimation. After that we generate imputed values  $Y_{imp}^*$  by drawing from  $Y_{j,mis}^* | Y_{obs}, \alpha \sim Dir_{m_j}(\alpha_{j,mis})$  (I-Step), where  $m_j$  is the number of missing values for unit  $j$  and  $\alpha_{j,mis}$  are the drawn alphas from the P-Step for the subcategories with missing values for unit  $j$ . This procedure will generate imputed values that sum up to one, but we need to make sure that the sum over the imputed and the observed values for one unit equals one. For this, we need to reweight our imputed values to get our final estimates:  $Y_{j,mis} = (1 - Y_{j,obs}' i_{6-m_j}) Y_{j,mis}^*$ , with  $i_{6-m_j}$  being a vector of ones of length  $6-m_j$ .

14. This procedure will provide consistent estimates and in theory no additional adjustments are necessary. However, this approach can not deal with proportions that are actually zero, so before imputation we set all values less than 0.001 to 0.001 and change values less or equal to 0.001 back to zero after imputation. This leads to minor inconsistencies after imputation. Additionally, missing values for the total number of employees have to be imputed before the Dirichlet approach can be applied. Here it still can happen that the imputed value is less than the sum of the observed values. So we run the same correction methods as for the other three approaches discussed above, to make sure that all estimates are consistent. Still, the necessary changes are relatively small compared to the changes for the three independent approaches described in A-C. Compared to the other methods this imputation method is not fully Bayesian, however, since we don't draw our parameters from their posterior predictive distribution. Instead we approximate this distribution by a normal distribution with parameters obtained from the maximum likelihood estimation.

#### E. Bayesian Dirichlet/Multinomial Imputation

15. To obtain fully Bayesian imputations that directly take the linear restrictions into account, we consider a three stage imputation that consists of the following three steps: Since the Dirichlet approach cannot directly incorporate covariates in the imputation model, we impute missing values with the simple imputation approach described in A as a first step. The idea is to get consistent estimates that incorporate as much information from the covariates as possible. In the second step we obtain a vector of proportions  $p_j$ , with  $j=1, \dots, n$  for each unit by taking random draws from the Dirichlet distribution  $p_j \sim Dir(\alpha_j)$ , with  $\alpha_j = (c * Y_{j,work}, c * Y_{j,train}, c * Y_{j,exec}, c * Y_{j,own}, c * Y_{j,marg}, c * Y_{j,other})$  and  $c=500$ . Since the parameter alpha has to be greater than zero, all values less than 0.005 are set to 0.005. The constant factor  $c$  is necessary, because for small establishments where the imputed total value can be less than one, the initial imputed value for a nonzero subcategory can be very small since values are imputed for cubic root transformed variables. For the Dirichlet approach the values have to be transformed back so imputed values less than one will be further reduced. But we still want to make sure that the imputation model differentiates between actual zeros and subcategories with a positive value, i.e. the probability vector obtained from the draw from the Dirichlet distribution should still have a higher probability mass for nonzero values. The constant factor  $c$  will guarantee this. Since the expected value for the Dirichlet distribution is given by

$$E(p_i) = \frac{\alpha_i}{\sum \alpha_i}, \text{ with } 0 \leq p_i \leq 1, \text{ multiplying each alpha with a constant } c \text{ does not change the expected}$$

value of  $p_i$  (it will affect the variance though, increasing  $c$  will decrease the imputation variance). The value of  $c$  is selected somewhat arbitrarily. We investigated the influence of  $c$  on the simulation results for  $c=250$ ,  $c=750$ , and  $c=10,000$  and didn't find significant differences in any of the results.

16. In the third step the missing values for each unit are imputed by taking random draws from the multinomial distribution with  $size = Y_t - \sum Y_{obs,i}$  and the vector of probabilities  $p_{mis}^*$  only for the missing values, where the probabilities are adjusted to sum up to one. Again, the correction step described in A is performed after the imputation to correct the imputed totals and inconsistencies due to rounding errors. Since the simple imputation in the first step is mainly performed to incorporate the covariables into the imputation model, but on the other hand also is the most time consuming step, it is only performed once every imputation round and not every iteration.

#### IV. THE SIMLUATION DESIGN

17. To assess the quality of our imputation models our main focus is on repeated sampling properties. We use the following sampling design: Starting point is the wave 2007 from the IAB Establishment Panel. Deleting all records with missing values, we obtain a dataset of 11536 records and treat these data as the population. From this population we take a random sample with replacement of the same sample size ( $n=11536$ ). Missing values are generated for each of the seven variables described in II using a MAR design described below. After that the missing values are imputed again using the five different imputation approaches described in III. For every imputation method the number of imputations is set to 10 and the number of iterations between the imputations is set to 20. The whole process of sampling, generating missings and imputation is repeated 100 times.

##### A. The missing data generator

18. Generally the implicit assumption for complete case analysis that units are missing completely at random (MCAR) is seldom fulfilled. For this reason we establish a simulation design that is based on the more realistic assumption of a missing at random (MAR) mechanism. Since we have only a few missings in the questions on the personnel structure, we develop a model that doesn't necessarily represent the true missing mechanism, but at least makes some plausible assumptions. In our simulation the probability to be missing is based on three completely observed covariates and can be described with the following

model  $p_{mis} = \frac{\exp(Y)}{1 + \exp(Y)}$  with  $Y = 1.4 - 0.5X_1 - 0.01X_2 - X_3$ , where  $X_1$  is the expected development for the

number of employees in the next five years (coded in six categories from more than 10% increase to more than 10% decrease),  $X_2$  is the number of unskilled workers and  $X_3$  is a dummy that indicates whether the establishment has accepted an industry-wide wage agreement or not. Note that all  $X$  variables have a negative influence on the probability to be missing. The idea behind the model is that establishments that expect to have an increase in the number of employees tend to be the establishments with a good order situation and they might be less willing to spend time on long questionnaires, so the probability of missing values is higher for these establishments. On the other hand the probability that an establishment has a department of its own for the human resources will increase with the number of unskilled workers and this again will lead to a decrease in the probability of missing values since it will be easy for the human resources department to provide information about the personnel structure and the department can spend more time on the questionnaire than owners of a small establishment that fill out the questionnaire on their own. Lastly the IAB Establishment Panel is supported by the big German labour organizations and establishments that participate in an industry-wide wage agreement might be more willing to follow their request to fill in all questions in the survey.

19. The same missing data mechanism is applied to all seven variables generating roughly 30% missings for each variable. Units with only one missing value are imputed deductively before the five regression based imputation methods are applied.

##### B. Estimates of interest

20. We are interested in two analyses. The first one contains some simple descriptive statistics: The (unweighted) means for the seven variables computed for the 16 German Länder. The second one is a typical economic analysis that tries to explain the existence of a collective wage agreement in an establishment by the establishment size. To include more information about the quality of our imputation models, we use the number of employees that are covered by social security instead of the total number

of employees as the explanatory variable. This number actually is the sum of the number of workers and the number of trainees. This allows us to incorporate two of the subcategories in our model. The actual model is given by the following equation:

$$Y \sim emp_{<10} + emp_{<50} + emp_{<100} + emp_{<250} + emp_{<750} + emp_{>750} + industry.dummies \quad (2)$$

, with  $Y$  being the dummy for collective wage agreement,  $emp$  the number of employees covered by social security in 6 categories and 17 industry dummies as covariates. In our analysis we are interested in the estimates for the intercept and the five remaining establishment size categories.

### C. Quality measures

21. We are interested in four quality measures: The average point estimates across the 100 samples, the 95% coverage rate, i.e. how often the true value lies within the 95% confidence interval around the estimated value, the confidence interval overlap between the original and the estimated confidence interval and the variance ratio between the estimated and the original variance. All measures are computed three times: For the bootstrap samples, for the datasets with missings and for the imputed datasets.

## V. Results

**Table 2: Imputation results for the number of workers using the simple imputation approach**

	NUMBER OF WORKERS													number of obs
	org mean	sample mean	mis mean	imp mean	sample cov	mis cov	imp cov	sample overl	mis overl	imp overl	sample var_ ratio	mis var_ ratio	imp var_ ratio	
Reg1	79.85	81.67	94.16	81.78	0.97	0.81	0.97	0.81	0.65	0.81	1.01	1.24	1.02	392
Reg2	303.47	294.44	359.77	294.59	0.85	0.93	0.86	0.80	0.75	0.80	0.90	1.10	0.90	180
Reg3	110.78	111.65	126.29	111.65	0.93	0.70	0.93	0.79	0.60	0.78	0.99	1.19	0.99	834
Reg4	56.08	55.37	62.64	55.41	0.93	0.86	0.93	0.78	0.71	0.78	0.96	1.19	0.96	781
Reg5	181.80	183.64	213.98	183.49	0.83	0.74	0.83	0.77	0.61	0.77	0.98	1.20	0.98	1126
Reg6	122.21	123.29	141.48	123.31	0.95	0.76	0.95	0.79	0.63	0.79	1.00	1.19	1.00	746
Reg7	83.05	85.24	96.92	85.30	0.97	0.80	0.97	0.80	0.62	0.80	1.02	1.21	1.02	560
Reg8	158.33	159.85	187.97	159.87	0.93	0.89	0.93	0.79	0.69	0.79	0.98	1.20	0.98	902
Reg9	217.41	218.12	256.89	218.06	0.90	0.89	0.90	0.80	0.69	0.80	0.97	1.16	0.97	866
Reg10	71.62	72.53	91.07	72.58	0.89	0.87	0.89	0.80	0.70	0.80	0.96	1.21	0.96	396
Reg11	109.49	108.11	124.26	108.03	0.70	0.88	0.70	0.79	0.74	0.79	0.84	1.01	0.84	594
Reg12	50.06	49.64	53.91	49.65	0.92	0.91	0.92	0.78	0.73	0.78	0.97	1.16	0.97	777
Reg13	60.71	61.61	66.80	61.59	0.95	0.90	0.94	0.80	0.72	0.80	1.01	1.20	1.01	682
Reg14	86.23	87.16	97.63	87.32	0.89	0.89	0.90	0.77	0.71	0.77	0.97	1.23	0.97	949
Reg15	73.57	73.75	79.30	73.87	0.95	0.87	0.95	0.80	0.71	0.80	0.99	1.17	0.99	793
Reg16	60.69	61.23	68.29	61.30	0.96	0.93	0.97	0.81	0.71	0.81	0.99	1.25	0.99	958
Aver.					0.91	0.85	0.91	0.79	0.69	0.79	0.97	1.18	0.97	

22. Table 2 is an example for the results for the number of workers using the simple imputation approach. There are some noteworthy results we can see in this table. First of all, the average sample from the bootstrap samples always differs to some extent from the original sample mean. Especially for region 2 there is a significant difference. The results for the complete case analysis for the datasets with missings overestimate the means in every region. This is not surprising, since the missing mechanism generates more missings for small establishments. The imputations correct this bias. The average estimate over the 100 simulations is close to the original estimate and it is almost identical to the estimates from the bootstrap sample, which actually are the benchmark for the imputations. For the 95% coverage rates, we can see that even the bootstrap samples seldom reach the expected coverage of 95%. For region 11 it is as low as 70%. This is probably a result of the fact that we don't use weights for these estimates and thus taking random samples from the population can have significant effects on the estimation of means that are highly affected by outliers. The coverage rate for the complete case analysis is lower for most

estimates, although there are some regions (region 1 and region 11) where the coverage actually is higher. This stems from the fact that with a reduced number of observations, the precision of the estimates decreases (the ratio between the complete case variance and the true variance is more than 1.2 for most estimates) and this will widen the 95% confidence intervals. The coverage for the imputed dataset is almost identical to the coverage of the bootstrap sample, again indicating the good quality of the imputations. For the average confidence interval overlap we find similar results: The overlap is more or less identical between the bootstrap sample and the imputed dataset, whereas the overlap for the complete case analysis is lower for all estimates. If we look at the variance ratios we can see that the low coverage rates for some of the bootstrap samples are a result of the reduced variability. The two regions with the lowest coverage rate (region 2 and 11) also have a significant reduction in the variability (0.9 and 0.84 respectively for the variance ratios). All other variance ratios are close to 1. For the complete case analysis we can see a significant increase of the variance for almost all of the estimates, which is a result of the reduced number of observations. The variance ratio for the imputed dataset again is very close to the variance ratio for the bootstrap sample, indicating that the additional variance introduced by the imputation is only very small.

23. To compare the results for the different imputation approaches we will only compare the average estimates across the 16 regions provided in the last row of Table 2. Tables 3-5 provide some results for all the different imputation methods. The results for the number of trainees, the number of executives and the number of marginal workers are almost identical for all five methods and therefore are moved to the Appendix (Table 10-12) for brevity.

**Table 3: Average imputation results from the different imputation methods for the number of workers**

	<i>workers</i>								
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.908	0.852	0.909	0.793	0.686	0.793	0.972	1.182	0.971
independent	0.895	0.861	0.893	0.788	0.691	0.788	0.963	1.173	0.966
proportions	0.906	0.868	0.899	0.798	0.696	0.797	0.969	1.175	0.969
Dirichlet	0.906	0.876	0.924	0.791	0.692	0.785	0.972	1.184	1.261
Bayesian Dir.	0.884	0.866	0.914	0.785	0.698	0.766	0.964	1.166	1.747

24. For the number of workers it seems that coverage is a little higher for the Dirichlet and the Bayesian Dirichlet approach, but this is mainly a result of an increased variability in the estimates. The variance ratios are 1.261 and 1.747 respectively for these two methods. The confidence interval overlap indicates that the first three methods obtain overlaps that are almost identical to the bootstrap sample overlaps, whereas the overlap is slightly reduced for the Dirichlet method and significantly reduced for the Bayesian Dirichlet method. Since the imputation variance introduced by the first three methods is small, these methods seem preferable to the Dirichlet methods.

**Table 4: Average imputation results from the different imputation methods for the number of owners**

	<i>owners</i>								
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.937	0.798	0.945	0.787	0.671	0.708	0.996	1.128	1.694
independent	0.946	0.802	0.943	0.791	0.674	0.685	0.995	1.126	2.576
proportions	0.938	0.806	0.951	0.797	0.674	0.590	0.996	1.128	4.394
Dirichlet	0.943	0.778	0.795	0.791	0.661	0.519	0.996	1.126	3.470
Bayesian Dir.	0.920	0.780	0.984	0.778	0.653	0.717	0.990	1.124	2.366

25. The number of owners and working family members is the most difficult to impute variable. The confidence interval overlap is reduced for all methods, with a very significant reduction for the proportions and the Dirichlet approach. This is not surprising, since both methods model this variable as the proportion of the total and there is hardly any correlation between the total number of employees and

this variable. The Dirichlet approach also performs poorly in terms of the coverage rate. The Bayesian Dirichlet approach on the other hand performs surprisingly well in terms of coverage and overlap.

26. The variance ratio of the imputed datasets is very high for all methods. Note however that in the original dataset this variable contains only integer values between 1 and 21, with roughly 87% of the values being 1 or 0 and only 15 values higher than 5. It is not surprising that any imputation model that is based on the assumption of a continuous variable will introduce more variability here. Still the introduced variability is significantly larger for the proportions and the Dirichlet approach.

**Table 5: Average imputation results from the different imputation methods for the number of others**

	<b>others</b>								
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.803	0.808	0.852	0.790	0.760	0.783	0.912	1.118	1.045
independent	0.804	0.811	0.866	0.793	0.762	0.777	0.916	1.118	1.220
proportions	0.800	0.822	0.937	0.790	0.765	0.746	0.903	1.101	2.150
Dirichlet	0.819	0.825	0.905	0.793	0.763	0.735	0.928	1.139	1.560
Bayesian Dir.	0.805	0.817	0.869	0.787	0.759	0.768	0.939	1.159	1.272

27. For the number of other employees we find again that the simple imputation approach works best in terms of confidence interval overlap. The proportions and Dirichlet approach have a higher coverage rate, but this is a result of the increased variance and they both perform poorly on the confidence interval overlap with an overlap that is actually lower than the overlap for the complete case analysis.

28. To further investigate the quality of the different imputation approaches we also computed the average absolute difference between the estimate from the bootstrap sample and the estimate from the imputed dataset across the 100 simulation runs and then averaged again across the 16 German Länder. Results are provided in Table 6

**Table 6: Average absolute deviation between the bootstrap sample mean and the imputed mean**

	<b>Average absolute deviation</b>				
	<i>simple</i>	<i>independent</i>	<i>proportions</i>	<i>Dirichlet</i>	<i>Bayesian Dirichlet</i>
employees total	0.344	0.381	0.340	4.877	11.770
workers	0.219	0.349	0.836	4.172	11.369
trainees	0.073	0.130	0.328	0.298	0.147
executives	0.050	0.078	0.267	0.186	0.077
owners	0.043	0.069	0.160	0.143	0.054
marginal workers	0.079	0.133	0.234	0.283	0.233
others	0.048	0.070	0.151	0.131	0.061

29. Obviously the simple imputation approach provides the best results for all variables. The Dirichlet and the Bayesian Dirichlet approach perform rather poorly for the number of workers and the total number of employees. Since the imputation method for the total number of employees is identical for all five approaches, the bad results for this variable seem to be a result of the correction methods that are applied, if the imputed total is less than the sum of the observed subcategories (see Section III.A). For the number of executives, the number of owners and the number of other employees the Bayesian Dirichlet approach seems to be a useful alternative to the simple imputation method. The proportions and the Dirichlet approach constantly provide less useful results than the other methods.

30. Results for the regression are provided in Tables 7-9. Here the coverage rates of the bootstrap samples are close to their expected value of 0.95 and the expected value across the 100 samples is almost identical to the true value. The results for the complete case analysis are severely biased and coverage rates are as low as 0.52 for some estimates.

**Table 7: Results from a logit regression to explain if establishments accept collective wage agreements (simple imputation)**

	<i>simple</i>												
	<i>org</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>
	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>cov</i>	<i>cov</i>	<i>cov</i>	<i>overl</i>	<i>overl</i>	<i>overl</i>	<i>var_</i>	<i>var_</i>	<i>var_</i>
											<i>ratio</i>	<i>ratio</i>	<i>ratio</i>
Intercept	-1.09	-1.09	-0.79	-1.10	0.94	0.52	0.93	0.78	0.47	0.79	1.01	1.24	1.01
10<x<=50	0.84	0.83	0.91	0.84	0.92	0.83	0.92	0.80	0.65	0.81	1.00	1.25	1.02
50<x<100	1.29	1.29	1.41	1.27	0.94	0.77	0.96	0.80	0.61	0.80	1.00	1.33	1.02
100<x<=250	1.81	1.81	1.89	1.81	0.97	0.86	0.96	0.79	0.70	0.79	1.00	1.30	1.02
250<x<=750	2.35	2.36	2.32	2.35	0.92	0.90	0.93	0.77	0.76	0.78	1.00	1.26	1.01
>750 emp.	3.86	3.93	3.81	3.93	0.98	0.93	0.97	0.80	0.76	0.80	1.03	1.21	1.04

31. The simple approach provides very good results for the regression. The average estimate is very close to the bootstrap sample estimate, the coverage rate is about 0.95 and the confidence interval overlap is more or less identical to the overlap of the bootstrap samples. The variance introduced by the imputation is very low. The independent and the proportions approach provide almost similar results as the simple approach. The tables can be found in the Appendix (Tables 13-14).

**Table 8: Results from a logit regression to explain if establishments accept collective wage agreements (Dirichlet imputation)**

	<i>Dirichlet</i>												
	<i>org</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>
	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>cov</i>	<i>cov</i>	<i>cov</i>	<i>overl</i>	<i>overl</i>	<i>overl</i>	<i>var_</i>	<i>var_</i>	<i>var_</i>
											<i>ratio</i>	<i>ratio</i>	<i>ratio</i>
Intercept	-1.09	-1.08	-0.79	-1.08	0.95	0.58	0.96	0.80	0.48	0.80	1.00	1.24	1.00
10<x<=50	0.84	0.83	0.92	0.84	0.98	0.81	0.98	0.82	0.62	0.83	1.00	1.25	1.03
50<x<100	1.29	1.29	1.42	1.27	0.94	0.79	0.95	0.80	0.59	0.79	1.00	1.32	1.02
100<x<=250	1.81	1.81	1.89	1.77	0.95	0.88	0.95	0.80	0.69	0.79	1.00	1.30	1.01
250<x<=750	2.35	2.35	2.30	2.31	0.94	0.96	0.96	0.80	0.76	0.79	1.00	1.25	1.01
>750 emp.	3.86	3.88	3.77	3.67	0.95	0.96	0.84	0.80	0.76	0.73	1.01	1.19	0.95

32. The Dirichlet approach provides good results, too, but we can see a little bias for the largest establishment size category and this bias leads to a reduced coverage rate and a reduced confidence interval overlap. Even the variance estimate for that category is lower than the original estimate

**Table 9: Results from a logit regression to explain if establishments accept collective wage agreements (Bayesian Dirichlet imputation)**

	<i>Bayesian Dirichlet</i>												
	<i>org</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sampl</i>	<i>mis</i>	<i>imp</i>	<i>sampl</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>
	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>e cov</i>	<i>cov</i>	<i>cov</i>	<i>e overl</i>	<i>overl</i>	<i>overl</i>	<i>var_</i>	<i>var_</i>	<i>var_</i>
											<i>ratio</i>	<i>ratio</i>	<i>ratio</i>
Intercept	-1.09	-1.10	-0.81	-1.10	0.99	0.56	0.99	0.82	0.50	0.82	1.00	1.24	1.00
10<x<=50	0.84	0.84	0.91	0.85	0.94	0.79	0.94	0.82	0.61	0.82	1.00	1.25	1.02
50<x<100	1.29	1.29	1.40	1.26	0.96	0.84	0.94	0.78	0.63	0.78	1.00	1.32	1.03
100<x<=250	1.81	1.81	1.88	1.78	0.92	0.84	0.91	0.80	0.72	0.81	1.00	1.30	1.02
250<x<=750	2.35	2.36	2.34	2.31	0.94	0.93	0.89	0.80	0.79	0.77	1.00	1.26	1.01
>750 emp.	3.86	3.75	3.61	3.46	0.87	0.69	0.57	0.75	0.64	0.53	0.97	1.12	0.92

33. The same effect can be observed for the Bayesian Dirichlet approach. Here the coverage and the overlap are reduced even further. These results probably stem from the fact that for large establishments the proportion of the number of workers to the total number of employees will be close to one for most establishments and imputation models based on these proportions might have difficulties differentiating there, which could lead to the observed overestimation for the number of workers in large establishments.

## VI. Conclusions

34. In this paper we evaluated different multiple imputation methods that provide consistent estimates for variables that are subject to one linear constraint. All methods provide good results with coverage rates and confidence interval overlaps close to the values obtained from bootstrap samples from the population. In our simulation based on variables on the personnel structure of establishments that participated in the IAB Establishment Panel in 2007, we found that the simplest method – imputing all variables separately ignoring the semi-continuity and correcting after the imputation process – provided the best results. Especially compared to the methods that are based on proportions (the proportions, Dirichlet and the Bayesian Dirichlet method) the additional variability introduced by the imputation is low and the average absolute deviation from the true value is very small for all variables. We also found a small bias for the number of workers for the Dirichlet and the Bayesian Dirichlet approach. Further studies are necessary to determine the reason for this bias.

35. There might be two possible reasons why the simplest approach outperforms all the other methods: First, the distribution of the proportions of the total number of employees is heavily skewed across the six subcategories. Especially for the large establishments by far the largest share goes to the first category (the number of workers) leaving very often less than 1% of the total to be distributed across the other categories. Even with the nested imputation approach we used in our simulations, it will be difficult to model these proportions. Second, for all the seven variables in our simulation, we have information from last year. This information always is a very strong predictor for the actual values. We also use these information in the methods based on proportions, but some of the predictive power might be lost if the information from last year is computed as the proportion of the actual total number of employees.

36. We expect that the methods based on proportions will work better in settings where the proportions are more equally distributed across the subcategories. Therefore we will evaluate the performance of the five imputation methods on other variables that are subject to linear constraints in future studies.

## References

- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input, *Statistical Science* 9, 538–558.
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001): A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* 27, 85–96.
- Rubin, D.B. (1978): Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 20-40
- Schenker, N., Raghunathan, T.E., Chiu, P., Makuc D., Zhang G., and Cohen, A.(2006): Multiple Imputation of Missing Income Data in the National Health Interview Survey, *Journal of the American Statistical Association*, 101, 924-933
- Tanner, M.A., Wong, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation (with discussion), in: *Journal of the American Statistical Association*, 82, S. 528-550.
- Tempelman, C. (2007): Imputation of ‘Restricted Data – Applications to Business Surveys, Ph.D. thesis, Statistics Netherlands

## Appendix

Table 10: Average imputation results from the different imputation methods for the number of trainees

<i>trainees</i>									
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.893	0.888	0.902	0.794	0.769	0.796	0.955	1.139	0.962
independent	0.898	0.892	0.909	0.793	0.775	0.794	0.958	1.138	1.004
proportions	0.894	0.900	0.937	0.798	0.776	0.797	0.963	1.143	1.153
Dirichlet	0.890	0.885	0.907	0.795	0.773	0.791	0.968	1.148	1.167
Bayesian Dir.	0.883	0.881	0.896	0.787	0.762	0.786	0.963	1.138	1.008

Table 11: Average imputation results from the different imputation methods for the number of executives

<i>executives</i>									
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.863	0.897	0.868	0.801	0.783	0.802	0.938	1.120	0.943
independent	0.827	0.861	0.833	0.786	0.769	0.786	0.930	1.101	0.942
proportions	0.855	0.889	0.886	0.795	0.776	0.798	0.949	1.127	1.025
Dirichlet	0.850	0.874	0.861	0.790	0.770	0.789	0.954	1.139	1.095
Bayesian Dir.	0.831	0.867	0.846	0.784	0.773	0.784	0.926	1.096	0.933

Table 12: Average imputation results from the different imputation methods for the number of marginal workers

<i>marginal workers</i>									
	<i>sample cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sample overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
simple	0.865	0.921	0.873	0.792	0.757	0.793	0.948	1.238	0.957
independent	0.882	0.928	0.899	0.797	0.762	0.797	0.959	1.250	1.025
proportions	0.876	0.929	0.916	0.802	0.759	0.793	0.947	1.237	1.126
Dirichlet	0.888	0.919	0.916	0.799	0.759	0.791	0.956	1.250	1.113
Bayesian Dir.	0.878	0.911	0.894	0.786	0.747	0.785	0.958	1.254	1.067

Table 13: Results from a logit regression to explain if establishments accept collective wage agreements (independent imputation)

<i>independent</i>													
	<i>org mean</i>	<i>sample mean</i>	<i>mis mean</i>	<i>imp mean</i>	<i>sampl e cov</i>	<i>mis cov</i>	<i>imp cov</i>	<i>sampl e overl</i>	<i>mis overl</i>	<i>imp overl</i>	<i>sample var_ratio</i>	<i>mis var_ratio</i>	<i>imp var_ratio</i>
Intercept	-1.09	-1.08	-0.78	-1.08	0.96	0.52	0.96	0.79	0.45	0.79	1.01	1.24	1.01
10<x<=50	0.84	0.84	0.92	0.85	0.88	0.74	0.93	0.76	0.61	0.76	1.00	1.26	1.02
50<x<100	1.29	1.30	1.41	1.30	0.96	0.81	0.96	0.82	0.62	0.82	1.00	1.32	1.03
100<x<=250	1.81	1.82	1.89	1.82	0.95	0.86	0.95	0.80	0.69	0.81	1.00	1.30	1.02
250<x<=750	2.35	2.35	2.31	2.35	0.98	0.95	0.98	0.80	0.76	0.81	1.00	1.25	1.01
>750 emp.	3.86	3.91	3.76	3.91	0.95	0.89	0.97	0.77	0.71	0.78	1.03	1.18	1.03

**Table 14: Results from a logit regression to explain if establishments accept collective wage agreements (imputation based on proportions)**

	<i>proportions</i>												
	<i>org</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>	<i>sample</i>	<i>mis</i>	<i>imp</i>
	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>mean</i>	<i>cov</i>	<i>cov</i>	<i>cov</i>	<i>overl</i>	<i>overl</i>	<i>overl</i>	<i>var_</i>	<i>var_</i>	<i>var_</i>
											<i>ratio</i>	<i>ratio</i>	<i>ratio</i>
Intercept	-1.09	-1.08	-0.78	-1.08	0.91	0.51	0.91	0.77	0.44	0.77	1.00	1.24	1.00
10<x<=50	0.84	0.83	0.91	0.85	0.97	0.81	0.96	0.82	0.63	0.81	1.00	1.25	1.03
50<x<100	1.29	1.29	1.40	1.29	0.97	0.74	0.97	0.80	0.62	0.81	1.00	1.32	1.03
100<x<=250	1.81	1.81	1.88	1.82	0.91	0.90	0.96	0.79	0.71	0.79	1.00	1.30	1.03
250<x<=750	2.35	2.34	2.30	2.36	0.94	0.93	0.94	0.81	0.75	0.79	1.00	1.25	1.02
>750 emp.	3.86	3.94	3.85	3.95	0.96	0.93	0.96	0.80	0.76	0.80	1.04	1.24	1.07