

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

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Bonn, Germany, 25-27 September 2006)

Topic (iii): Editing microdata for release

**A NEW APPROACH FOR DISCLOSURE CONTROL IN THE IAB  
ESTABLISHMENT PANEL – MULTIPLE IMPUTATION FOR A BETTER DATA ACCESS**

**Invited Paper**

Prepared by Stefan Bender, Jörg Drechsler, Agnes Dundler and Susanne Rässler, Institute for  
Employment Research (IAB) and Thomas Zwick, Centre for European Economic Research

# A New Approach for Disclosure Control in the IAB Establishment Panel - Multiple Imputation for a Better Data Access<sup>1</sup>

Stefan Bender\*, Jörg Drechsler\*, Agnes Dundler\*,  
Susanne Rässler\*, and Thomas Zwick\*\*

\* Institute for Employment Research (IAB), Regensburger Straße 104,  
90478 Nürnberg, Germany

Stefan.Bender@iab.de, Joerg.Drechsler@iab.de,  
Agnes.Dundler@iab.de, Susanne.Raessler@iab.de

\*\* Centre for European Economic Research (ZEW), L 7,1,  
68161 Mannheim, Germany  
zwick@zew.de

**Abstract.** For every dataset, statistical agencies have to face the dilemma of guaranteeing the confidentiality of survey respondents on the one hand and offering sufficiently detailed data for scientific use on the other hand. For that reason a variety of methods to guarantee disclosure control is discussed in the literature.

In this paper we introduce two approaches based on multiple imputation and their possible application to the German IAB Establishment Panel. The first, proposed by Rubin (1993), generates fully synthetic datasets while the second imputes values only for selected variables that bear a high risk of disclosure.

We apply the two methods to a set of variables from the 1997 wave of the Establishment Panel and evaluate their quality by comparing results from an analysis done by Thomas Zwick (2005) with the original data with results we achieve for the same analysis run on the dataset after the imputation procedure.

**Keywords:** Confidentiality; Multiple Imputation; statistical disclosure control; IAB Establishment Panel; synthetic datasets; selective multiple imputation of key variables

---

<sup>1</sup> The research provided in this paper is part of the project “Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“ financed by the Federal Ministry for Education and Research (BMBF) and conducted by the following institutes: Federal Statistical Office Germany, Statistical Offices of the Länder, Institute for Applied Economic Research (IAW), Centre for European Economic Research (ZEW), Institute for Employment Research (IAB). For more information about this project see for instance Ronning and Rosemann (2006) or Ronning et al (2005).

## 1 Introduction

In recent years, the public demand for micro data increased dramatically. But statistical agencies face the dilemma that, although they might be willing to provide all the information required, a release of the datasets might not be possible for confidentiality reasons. The natural interest of enabling as much research as possible with the collected data has to stand back behind the confidentiality guaranteed to the survey respondent: Once the confidentiality is in doubt, potential respondents might be less willing to provide sensitive information, might give wrong answers on purpose or might even be unwilling to participate at all, with devastating consequences for the quality of the data collected.

For that reason, a variety of anonymization methods has been developed to provide as much information to the public as possible, while satisfying the disclosure restrictions needed to maintain the quality of the collected data. However, information loss is a disadvantage common to all these approaches. Furthermore, for some of them, the analyst needs to know the techniques used for anonymization or some special software is necessary to achieve valid inferences.

This paper discusses two new approaches based on multiple imputation and their possible application to a German panel of establishments (the IAB Establishment Panel).

The first, proposed by Rubin (1993), generates fully synthetic datasets while the second replaces only variables that bear a high risk of disclosure with imputed values (see for example Little and Liu 2002).

Rubin suggests to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public.

Because all the imputed values are random draws from the posterior predictive distribution of the missing values given the observed values, disclosure of sensitive information is impossible, especially if the released dataset doesn't contain any real data. Another advantage of this approach is the sampling design for the imputed datasets. As the released datasets are simple random samples from the population the analyst doesn't have to allow for a complex sampling design in his models.

However, the quality of this method strongly depends on the accuracy of the model used to impute the "missing" values. If the model doesn't include all the relationships between the variables that are of interest to the analyst or if the joint distribution of the variables is miss-specified, results from the synthetic dataset can be biased.

To reduce this potential bias, the second approach discussed in this paper replaces observed values with imputed values only for variables that are pub-

licly available in other databases and might enable intruders to identify a single respondent.

The remainder of this paper is organized as follows: Section 2 provides a short overview of the multiple imputation framework and its modifications for disclosure control. Section 3 introduces the two datasets used. Section 4 describes the application of the two multiple imputation approaches for disclosure control to the IAB Establishment Panel. Section 5 evaluates these approaches by comparing results from an analysis done by Thomas Zwick (2005) with the original data with results achieved for the same analysis run on the dataset after the imputation procedure. The paper concludes with some final remarks.

## 2 Multiple Imputation

### 2.1 Multiple Imputation for Missing Data

Multiple imputation, introduced by Rubin (1978) and discussed in detail in Rubin (1987, 2004), is 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. Typically,  $m$  is small, such as  $m = 5$ . Each of the imputed (and thus completed) datasets is first analyzed by standard methods designed for complete data; the results of the  $m$  analyses are then combined in a completely generic way to produce estimates, confidence intervals and tests that reflect the missing-data uncertainty.

In this paper, we discuss analysis with scalar parameters only, for multidimensional quantities see Little and Rubin (2002, Section 10.2) To understand the procedure of analyzing multiply imputed datasets, think of an analyst interested in an unknown scalar parameter  $\theta$ , where  $\theta$  could be e.g. the mean of a variable, the correlation coefficient between two variables or a regression coefficient in a linear regression.

Inferences for this parameter for datasets with no missing values usually are based on a point estimate  $\hat{\theta}$ , an estimate for the variance of  $\theta$ ,  $\hat{V}$  and a normal or Student's  $t$  reference distribution. For analysis of the imputed datasets, let  $\hat{\theta}_i$  and  $\hat{V}_i$  for  $i = 1, \dots, m$  be the point and variance estimates for each of the  $m$  completed datasets. To achieve a final estimate over all imputations, these

estimates have to be combined using the combining rules first described by Rubin (1978).

For the point estimate, the final estimate simply is the average of the  $m$  point estimates  $\hat{\theta}_{MI} = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i$  with  $i=1, \dots, m$ . Its variance is estimated by  $T = W + (1 + m^{-1})B$ , where  $W = m^{-1} \sum_{l=1}^m \hat{V}_l$  is the “within-imputation” variance  $B = \frac{1}{m-1} \sum_{i=1}^m (\hat{\theta}_i - \hat{\theta}_{MI})^2$  is the “between-imputation” variance, and the factor  $(1 + m^{-1})$  reflects the fact that only a finite number of completed-data estimates  $\hat{\theta}_i$ ,  $i=1, \dots, m$  are averaged together to obtain the final point estimate. The quantity  $\hat{\gamma} = (1 + m^{-1})B/T$  estimates the fraction of information about  $\theta$  that is missing due to nonresponse.

Inferences from multiply imputed data are based on  $\hat{\theta}_{MI}$ ,  $T$ , and a Student’s  $t$  reference distribution. Thus, for example, interval estimates for  $\theta$  have the form  $\hat{\theta}_{MI} \pm t(1-\alpha/2)\sqrt{T}$ , where  $t(1-\alpha/2)$  is the  $(1-\alpha/2)$  quantile of the  $t$  distribution. Rubin and Schenker (1986) provided the approximate value  $\nu_{RS} = (m-1)\hat{\gamma}^{-2}$  for the degrees of freedom of the  $t$  distribution, under the assumption that with complete data, a normal reference distribution would have been appropriate (that is, the complete data would have had large degrees of freedom). Barnard and Rubin (1999) relaxed the assumption of Rubin and Schenker (1986) to allow for a  $t$  reference distribution with complete data, and suggested the value  $\nu_{BR} = (\nu_{RS}^{-1} + \hat{\nu}_{obs}^{-1})^{-1}$  for the degrees of freedom in the multiple-imputation analysis, where  $\hat{\nu}_{obs} = (1-\hat{\gamma})(\nu_{com})(\nu_{com} + 1)/(\nu_{com} + 3)$  and  $\nu_{com}$  denotes the complete-data degrees of freedom.

## 2.2 Fully Synthetic Datasets

In 1993, Rubin suggested to create fully synthetic datasets based on the multiple imputation framework. His idea was, to treat all units in the population that have not been selected in the sample as missing data, impute them according to the multiple imputation approach and then draw simple random samples from these imputed populations and release these synthetic data to the public.

For illustration, think of a dataset of size  $n$ , sampled from a population of size  $N$ . Suppose further, the imputer has information about some variables  $X$  for the whole population, for example from census records, and only the information from the survey respondents for the remaining variables  $Y$ . Let  $Y_{inc}$

be the observed part of the population and  $Y_{exc}$  the nonsampled units of  $Y$ . For simplicity, assume that there are no item-missing data in the observed dataset.

Now the synthetic datasets can be generated in two steps: First, construct  $m$  imputed synthetic populations by drawing  $Y_{exc}$   $m$  times independently from the posterior predictive distribution  $f(Y_{exc}|X, Y_{inc})$  for the  $N-n$  unobserved values of  $Y$ . If the released data should contain no real data for  $Y$ , all  $N$  values can be drawn from this distribution. Second, make simple random draws from these populations and release them to the public. The second step is necessary as it might not be feasible to release  $m$  whole populations for the simple matter of data-size. In practice, it is not mandatory to generate complete-data populations. The imputer can make random draws from  $X$  in a first step and only impute values of  $Y$  for the drawn  $X$ .

Analysis of the  $m$  simulated datasets follows the same lines as analysis after multiple imputation for missing values in regular datasets (see Section 2.1). However, the combination for the total variance slightly differs from the total variance in MI settings for treating missing data:

$$\hat{\text{var}}(\hat{\theta}_{MI}) = T_f = \frac{m+1}{m} B - W$$

This difference is due to the additional sampling from the synthetic units for fully synthetic datasets. Hence, the variance between the datasets  $B$  already reflects the variance within each population.

If  $m$  is large, inferences can be based on normal distributions. For moderate  $m$ , a  $t$  reference distribution is more adequate. The degrees of freedom are given by

$$v_f = (m-1)(1-r^{-1})^2 \quad \text{where } r = \frac{(1+m^{-1})B}{W}.$$

A disadvantage of this variance estimate is that it can become negative. For that reason Reiter (2002) suggests a slightly modified variance estimator that is always positive. Let  $n_{syn}$  be the number of observations in the released datasets sampled from the synthetic population, then the estimator can be calculated as follows:

$$T_f^* = \max(0, T_f) + \delta \left( \frac{n_{syn}}{n} W \right), \quad \text{where } \delta=1 \text{ if } T_f < 0, \text{ and } \delta=0 \text{ otherwise.}$$

### 2.3 Imputation of Selected Variables

In contrast to the creation of fully synthetic datasets, this approach replaces only observed values for variables that bear a high risk of disclosure (key variables) with synthetic values (see for example Little (1993)). This could be variables known to the public from other easily available databases or information from statements of accounts for incorporations. Masking these variables by replacing observed with imputed values prevents re-identification.

These values can be obtained by drawing from the posterior predictive distribution  $f(Y|X)$ , where  $Y$  indicates the variables that need to be modified to avoid disclosure and  $X$  are all variables that remain unchanged and are used as explanatory variables in the imputer's model (see Raghunathan et. al. (2003)).

Imputations are generated according to the multiple imputation framework as described in Section 2.1. But as in the full MI context, the variance estimation differs slightly from the MI calculations for missing data. Yet, it differs from the estimation in the full MI context as well. It is given by

$$T_p = W + m^{-1}B.$$

Similar to the variance estimator for multiple imputation of missing data  $m^{-1}B$  is the correction factor for the additional variance due to using a finite number of imputations. However, the additional  $B$  for missing data is not necessary. An intuitive justification is that the estimation of the imputed values is based on more information. The distribution from which the parameters  $\psi$  are drawn in the first step of the imputation procedure are conditioned not only on  $X$  but on  $Y$  as well. These true values are omitted only in the second step. So, using Rubin's variance estimator for the variance of  $\theta$  could lead to overestimation of the true variance. For a formal justification, see Reiter (2003). Inferences for  $\theta$  can be based on a Student's  $t$  reference distribution with

$$\nu_p = (m-1)\left(1 + \frac{W}{B/m}\right)^2 \text{ degrees of freedom.}$$

#### 2.4 Selective Multiple Imputation of Key Variables (SMIKe)

This approach, suggested by Little and Liu (2002), further reduces the number of substituted values by replacing only observations that face the highest risk of disclosure within each key variable with imputed values.

For illustration, think of a dataset composed of some categorical key variables  $X$  and several continuous variables  $Y$ . Cross tabulation of  $X$  yields the vector  $x$ , containing cell counts for all combinations of  $X$ . A previously defined sensitivity threshold  $s$  indicates whether a cell of  $x$  possibly allows re-identification. These sensitive cells are combined with selected non sensitive cells that are closely related to the sensitive cells with regard to  $Y$ . For this mixed set observations for  $X$  are replaced by imputed values and this modified dataset is released to the public.

Note that the sensitivity threshold  $s$  needs to be set carefully. Choosing  $s$  too low puts the original aim of confidentiality at risk. Choosing  $s$  too high means that data is unnecessarily changed and by that, information is unnecessarily lost.

To select a set of non sensitive cells for a sensitive cell, it is essential to find non sensitive cells that are as similar as possible to the sensitive cells

with respect to  $Y$ . This guarantees that the imputed values are distributed equally over the sensitive and the non sensitive cells and by this increases protection. One way to determine eligible cells is to measure the distance between the  $Y$ s of the different cells. If all variables in  $Y$  are, as assumed, continuous variables the Mahalanobis Distance can be a means of calculating this distance. Defining the number of non sensitive cells to be selected again needs careful weighting between information loss and protection gain.

Once the mixed dataset  $M$  of sensitive and non sensitive cells is determined, the values of  $X$  are discarded and new values are drawn from the posterior predictive distribution  $f(X|Y,M)$ .

For the parametric imputation model the values are drawn in two steps analogous to the multiple imputation settings for missing data: First, values for  $\psi$  are drawn from its posterior distribution given  $X$  and  $Y$ ,  $f(\psi | X, Y, M)$ .

Then, new values  $\tilde{X}$  are drawn for cases in  $M$ , given  $Y$  and the drawn  $\psi$ ,  $f(\tilde{X}_i | Y, \psi, M)$ .

Repeating this procedure  $m$  times and combining the imputed datasets for  $M$  each time with the unchanged data yields  $m$  complete datasets with sensitive data replaced by imputed values.

Statistical inference from these  $m$  datasets can be obtained according to Rubin's (1978) combining rules for multiply imputed datasets for missing data. Again, the variance estimator differs from the one given by Rubin (1978) for the reasons discussed in Section 2.3. As in the context of imputing selected variables completely, it is given by  $T_p = W + m^{-1}B$  with inferences for  $\theta$  based on a Student's  $t$  reference distribution with  $\nu_p = (m-1)(1 + \frac{W}{B/m})^2$  degrees of freedom.

### 3 The Datasets<sup>2</sup>

For the imputation of the IAB Establishment Panel, we use additional information from the German Social Security Data. In the following Section both datasets, the German Social Security Data and the IAB Establishment Panel will be described in detail.

#### 3.1 The German Social Security Data

The German employment register contains information on all employees covered by social security. The basis of the German Social Security Data (GSSD)

---

<sup>2</sup> This chapter follows the description given in Alda, Bender & Gartner (2005).



is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973.<sup>3</sup> This procedure requires employers to notify the social security agencies about all employees covered by social security.

As by definition the German Social Security Data only includes employees covered by social security - civil servants and unpaid family workers for example are not included - approx. 80% of the German workforce<sup>4</sup> are represented. However, the degree of coverage varies considerably across the occupations and the industries.

The notifications of the GSSD include for every employee, among other things, the working place and the establishment identification number. We use this number to match the selected establishment characteristics aggregated from the employment register with the IAB Establishment Panel. As we start with the 1997 wave of the panel, data are taken from the register for June, 30th 1997 (see Figure 2 in the Appendix for all characteristics used).

### **3.2 The IAB Establishment Panel**

The IAB Establishment Panel<sup>5</sup> is based on the employment statistics aggregated via the establishment number as of 30 June of a year. Consequently the panel only includes establishments with at least one employee covered by social security. The sample is drawn following the principle of optimum stratification according to the stratification cells of the establishment size class (10 categories) and the industry (16 categories<sup>6</sup>). These stratification cells are also used in the weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. 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. The response rate of units that have been interviewed repeatedly is over 80%. Each year the panel is accompanied by supplementary samples and follow-up samples in order to interview new or reviving establishments and to compensate for panel mortality. The list of questions includes detailed information about the firms' personnel structure, development and personnel pol-

---

<sup>3</sup> On the structure of the insurance number and on the data office of the pension insurance providers cf. Steeger (2000).

<sup>4</sup> An overview of the data is given in Bender, Hass, and Klose (2000), a detailed description can be found in Bender, Hilzendege, Rohwer, and Rudolph (1996).

<sup>5</sup> The approach and structure of the establishment panel are described for example by Bellmann (2002) and Kölling (2000).

<sup>6</sup> From 2000 onwards the stratification is done according to 20 industries.

icy. An overview of available information in 1997 is listed in the Appendix, Figure 2.

## 4 Applications to the IAB Establishment Panel

### 4.1 The Full MI Approach

In a first step, we only impute values for a set of variables from the 1997 wave of the IAB Establishment Panel. As it is not feasible to impute values for the millions of establishments contained in the German Social Security Data for 1997, we sample from this frame, using the same sampling design as for the IAB Establishment Panel stratified by establishment size, region and economic branch (see Table 1 in the Appendix for an example). Every stratum contains the same number of units as the observed data from the 1997 wave of the Establishment Panel. We gain further information by adding variables from the German Social Security Data and matching these variables to the observations in the Establishment Panel by establishment identification number. After matching, every dataset is structured as follows: Let  $N$  be the total number of units in the newly generated dataset, that is the number of units in the sample  $n_s$  plus the number of units in the panel  $n_p$ ,  $N=n_s+n_p$ . Let  $X$  be the matrix of variables with information for all observations in  $N$ . Then  $X$  consists of the variables establishment size, region and economic branch and the variables added from the German Social Security Data (see Figure 2 in the Appendix). Let  $Y$  be the selected variables from the Establishment Panel, with  $Y=(Y_{inc}, Y_{exc})$ , where  $Y_{inc}$  are the observed values from the Establishment Panel and  $Y_{exc}$  is the hypothetic missing data for the newly drawn values in  $X$ . Variables in  $Z$  are not imputed, but are used as explanatory variables in the imputation model for  $Y_{exc}$  (see Figure 1).

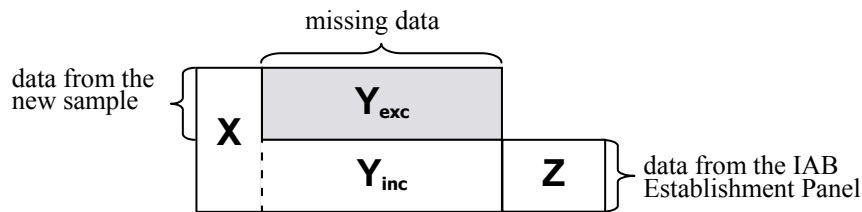


Fig. 1. The full MI approach for the IAB Establishment Panel.

Now, values for the missing data can be imputed as outlined in Section 2 by drawing  $Y_{exc}$   $m$  times independently from the posterior predictive distribution  $f(Y_{exc}|X, Y_{inc}, Z)$  for the  $N-n_p$  unobserved values of  $Y$ .

After the imputation procedure, all observations from the Establishment Panel are omitted and only the imputed values are kept for analysis. Results from this analysis can be compared with the results achieved with the real data.

As the Establishment Panel consists of many skip patterns leading to high collinearity in the dataset, we use only some of the variables from the panel not selected for imputation as explanatory variables  $Z$ . To select these, we use stepwise regression to find the five variables for every variable to impute, which explain the highest amount of the variance of this variable.

## **4.2 Drawing a New Sample From the German Social Security Data**

Due to panel mortality a supplementary sample has to be drawn for the IAB Establishment Panel every year. In the 1997 wave, this supplementary sample primarily consisted of new founded establishments, because in that year the questionnaire had a focus on new foundations. So, start-ups are overrepresented in the sample. Arguably, answers from these establishments differ systematically from the answers provided by establishments existing for several years. Drawing a new sample without taking this oversampling into account could lead to a sample after imputation that differs substantially from that in the Establishment Panel.

For simplicity reasons, we define establishments not included in the German Social Security Data before July 1995 as new foundations and delete them from the sampling frame and the Establishment Panel. For the 1997 wave of the Establishment Panel, this means a reduction from 8,850 to 7,610 observations. In a later stadium of the project, we will analyse the influence of new foundations on answers given in the survey.

Additionally, we have to make sure that every establishment in the survey is also represented in the German Social Security Data for that year. Merging the two datasets using the establishment identification number reveals that 278 remaining units from the panel are not included in the Employment Statistics. These units are also omitted leading to a final sample of 7,332 observations.

Furthermore, we have to verify that the stratum parameters size, economic branch and region match in both datasets. Merging indicates that there are some differences between the two records. If the datasets differ, values from the Employment Statistics are adopted.

Cross tabulation of the stratum parameters for the 7,332 observations left provides a matrix containing the number of observations for each stratum. For example, one cell of the matrix contains companies specialized in investment goods that are located in Berlin-West and employ 20 to 49 employees (see

Table 1 in the Appendix). Now, a new dataset can be generated easily, by drawing from the German Social Security Data according to this matrix.

### 4.3 Imputation of Selected Variables Applied to the IAB Establishment Panel

We consider the variables establishment size, region and economic branch as the variables with the highest risk of disclosure as they are easily available from other data sources. For them, we replace all observed values with imputed values (the SMiKE approach is more laborious and will be addressed in a later stadium of this project).

Imputing only some variables has the advantage that we don't have to sample from the original frame. Instead, we use the dataset from the Establishment Panel and replace parts of the data with missing values. After imputing these values, a different part of the original dataset is changed to missing and these values are imputed the same way. This procedure is repeated, until all observed values are once replaced by imputed values. Then all parts with imputed values from the different imputation rounds are combined to generate a new dataset with all values for the three variables replaced by imputed values. For the imputation the same explanatory variables, found by stepwise regression, are used as in the full MI approach.

For illustration, assume the three variables establishment size, region and economic branch are imputed in three steps. Let  $N$  be the number of observations in the dataset, divided in three parts:  $a=1, \dots, n_1$ ,  $b=n_1+1, \dots, n_2$  and  $c=n_2+1, \dots, N$ . In the first imputation round, without loss of generality, we replace all values in  $a$  with missing values for the variable establishment size  $x_e$ , all values in  $b$  for the region variable  $x_r$  and all values in  $c$  for the economic branch variable  $x_b$ . Then we impute all missing values  $m$  times and store the complete datasets as  $D_{1i}$ , with  $i=1, \dots, m$ . Afterwards, we go back to the original dataset, now deleting all values in  $b$  for  $x_e$  all values in  $c$  for  $x_r$  and all values in  $a$  for  $x_b$ . Imputing these values provides datasets  $D_{2i}$ . In the last round all values in  $c$  for  $x_e$ , all values in  $a$  for  $x_r$  and all values in  $b$  for  $x_b$  are set to missing and then imputed yielding  $D_{3i}$ . Datasets with the three variables completely imputed are achieved, by combining the parts with imputed values from every imputation round. Note, that it is irrelevant that the new dataset is formed from different imputation rounds. One of the basic assumptions for multiple imputation is that the draws from the posterior predictive distribution are independent, so the procedure described here might even be a way to make sure that this assumption holds.

Obviously, it is not necessary that the dataset is divided in three parts. A higher number of divides means more information is used in every imputation round. Theoretically, it would be possible to impute only one observation in

every imputation round, although this might not be feasible for two reasons. First, imputing every single value for every variable that needs disclosure control separately, becomes labour-intensive quickly. Second, with only one value missing, the model for the imputation, especially for categorical variables, might be “too” good, leading to imputed values that never differ from the real values and by this countering the confidentiality aim.

So the number of imputation rounds should be carefully selected to address the trade-off between using as much information as possible for the imputation model and guaranteeing confidentiality and the practicality of the approach. Further research is needed on this topic.

## **5 Comparison Between the Original and the Imputed Dataset**

To evaluate the two approaches discussed here, we will compare analytic results achieved with the original data with results from the imputed data. Basis is an analysis done by Thomas Zwick (2005) in ‘Continuing Vocational Training Forms and Establishment Productivity in Germany’ published in the *German Economic Review*, Vol. 6(2), 155-184.

Zwick analyses the productivity effects of different continuing vocational training forms in Germany. He argues that vocational training is one of the most important measures to gain and keep productivity in a firm. For his analysis he uses the waves from 1997 to 2001 from the IAB Establishment Panel.

In 1997 and 1999 the Establishment Panel included the following additional question that was asked if the establishment did support continuous vocational training in the first part of 1997 or 1999 respectively: ‘For which of the following internal or external measures were employees exempted from work or were costs completely or partly taken over by the establishment?’ Possible answers were: formal internal training, formal external training, seminars and talks, training on the job, participation at seminars and talks, job rotation, self-induced learning, quality circles and additional continuous vocational training. Zwick examines the productivity effects of these training forms and is able to demonstrate that formal external training, formal internal training and quality circles do have a positive impact on productivity. Especially for formal external courses the productivity effect can be measured even two years after the training.

To detect why some firms offer vocational training and others not, Zwick runs a probit regression using the 1997 wave of the Establishment Panel. The regression (see Table 2 in the Appendix for details) shows that establishments increase training if they expect to lose workers. One reason could be that the external skilled labour market in Germany is small and establishments have difficulties in finding new skilled workers. Furthermore, establishments tend

to offer more training, if high qualification needs are expected. Just like establishments that give a higher priority to additional apprenticeship training and continuing vocational training efforts instead of hiring externally qualified employees if they have vacancies for skilled jobs. Larger establishments tend to qualify employees more often because they usually have own training departments and can therefore train workers more efficiently. For firms with a high share of qualified employees, state-of-the-art technical equipment or investments in IT (information and communication technology) it is also essential to offer more training. Collective wage agreements are often associated with fringe benefits such as training, while workers councils usually attach high importance to continuing vocational training, so both have a positive effect on the amount of training offered.

In the regression, Zwick uses two variables, containing information about investment in IT and the co-determination of the employees, that are only included in the 1998 wave of the Establishment Panel. Moreover, he excludes some observations based on information from other years. As we impute only the 1997 wave of the Establishment Panel eliminating new founded establishments, we have to rerun the regression, using all observations except for new founded establishments and deleting the two variables from the 1998 wave from the model. Results from this regression are given in Table 3 in the Appendix and it is evident that the new regression differs only slightly from the regression with the original model. All the variables significant in Zwick's analysis are still significant.

The comparison of the results from this regression with the results from the imputed dataset will reveal whether valid inferences can be achieved with the manipulated dataset.

## **6 Concluding Remarks**

In this paper we introduce multiple imputation as a means of guaranteeing confidentiality in datasets and provide possible applications to the IAB Establishment Panel. Due to the short amount of time since the start of the project, no results are presented here, but we are positive to have first results within the next months.

At a later date, the more laborious but very promising SMiKE approach will be addressed.

## **References**

1. Alda, H., Bender, S., Gartner, H.: The Linked Employer-Employee Dataset of the IAB (LIAB). IAB Discussion Paper, No. 6. (2005)

2. Barnard, J., Rubin, D.B.: Small-sample Degrees of Freedom With Multiple Imputation. *Biometrika*, Vol. 86. (1999) 948-955
3. Bellmann, L.: Das IAB-Betriebspanel - Konzeption und Anwendungsbereiche. *Journal of the German Statistical Society*, Vol. 86. (2002) 177-188
4. Bender, S., Haas, A., Klose, C.: The IAB Employment Subsample 1975-1995. *Journal of Applied Social Science Studies*, Vol. 120. (2000) 649-662
5. Bender, S., Hilzendegen, J., Rohwer, G., Rudolph, H.: Die IAB Beschäftigtenstichprobe 1975-1990. *Beiträge zur Arbeitsmarkt- und Berufsforschung*, No. 197. (1996)
6. Kölling, A.: The IAB-Establishment Panel. *Journal of Applied Social Science Studies*, Vol. 120. (2000) 291-300
7. Little, R.J.A.: Statistical Analysis of Masked Data, *Journal of Official Statistics*, Vol. 9 (1993) 407-426
8. Little, R.J.A., Liu, F.: Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. *American Statistical Association Proceedings of the Joint Statistical Meetings*. (2002) 2133-2138
9. Little, R.J.A., Rubin, D.B.: *Statistical Analysis With Missing Data*, John Wiley & Sons, Hoboken (2002)
10. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple Imputation for Statistical Disclosure Limitation, *Journal of Official Statistics*, Vol. 19 (2003) 1-16
11. Reiter, J.P.: Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics*, Vol. 18. (2002) 531-544
12. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. *Survey Methodology*, Vol. 29 (2003) 181-189
13. Ronning, G., Rosemann, M.: Estimation of the Probit Model From Anonymized Micro Data. *Work Session on Statistical Data Confidentiality*, Geneva, 9-11 November 2005. *Monograph of Official Statistics*. Eurostat, Luxemburg (2006) 207-216
14. Ronning, G., Rosemann, M., Strotmann H.: Post-Randomization under Test: Estimation of a Probit Model. *Journal of Economics and Statistics*, Vol. 225. (2005) 544-566
15. Rubin, D.B.: Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. *American Statistical Association Proceedings of the Section on Survey Research Methods*. (1978) 20-40
16. Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York (1987)
17. Rubin, D.B.: Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, Vol. 9. (1993) 462-468

18. Rubin, D.B.: The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys. *The American Statistician*, Vol. 58. (2004) 298-302
19. Rubin, D.B., Schenker, N.: Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, Vol. 81. (1986) 366-374
20. Steeger, W.: 25 Jahre Datenstelle der Rentenversicherungsträger (DSRV). *Deutsche Rentenversicherung*, 10-11/2000. (2000) 648-684
21. Zwick, T.: Continuing Vocational Training Forms and Establishment Productivity in Germany. *German Economic Review*, Vol. 6(2). (2005) 155-184



## Appendix

Information contained in the IAB Establishment Panel (wave 1997)	Information contained in the German Social Security Data (from 1997)
<p><b>Available for establishments in the survey</b></p> <ul style="list-style-type: none"> <li>- number of employees in June 1996</li> <li>- qualification of the employees</li> <li>- number of temporary employees</li> <li>- number of agency workers</li> <li>- working week (full-time and overtime)</li> <li>- the firm's commitment to collective agreements</li> <li>- existence of a works council</li> <li>- turnover, advance performance and export share</li> <li>- investment total</li> <li>- overall wage bill in June 1997</li> <li>- technological status</li> <li>- age of the establishment</li> <li>- legal form and corporate position</li> <li>- overall company-economic situation</li> <li>- reorganisation measures</li> <li>- company further training activities</li> <li>- additional information on new foundations</li> </ul>	<p><b>Available for all German establishments with at least one employee covered by social security</b></p> <ul style="list-style-type: none"> <li>- number of full-time and part-time employees</li> <li>- short-time employment</li> <li>- mean and standard deviation of the employees age</li> <li>- mean and standard deviation of wages from full-time employees</li> <li>- mean and standard deviation of wages from all employees</li> <li>- occupation</li> <li>- schooling and training</li> <li>- number of women and men</li> <li>- number of German employees</li> </ul>
<p><b>Covered in both datasets</b></p> <ul style="list-style-type: none"> <li>➤ establishment number, branch and size</li> <li>➤ location of the establishment</li> <li>➤ number of employees in June 1997</li> </ul>	

Fig. 2. Data comparison.

Table 1. Stratification matrix.

Federal state	Branch of trade (16 categories)							
	Establishment size <sup>7</sup>	1 Agriculture, forestry	2 Mining and quarrying	3 Raw material processing	4 Investment goods	...	16 Non-profit organization	Total
Berlin-West	1 0-4	0	0	1	1	...	6	42
	2 5-9	2	0	0	2	...	0	25
	3 10-19	1	0	2	4	...	3	35
	4 20-49	0	1	1	4	...	5	29
	5 50-99	0	0	1	3	...	1	13
	6 100-199	1	0	2	2	...	2	31
	10 5000+	0	1	0	0	...	1	5
Total	4	3	9	28	...	40	275	
Berlin-East	1 0-4 svb	0	0	0	0	...	1	52
	2 5-9	0	0	1	6	...	3	45
	10 5000+	0	0	0	0	...	1	1
	Total	3	2	4	30	...	41	303
Bran-	1 0-4 svb	5	0	2	7	...	8	96
den-	...	...	...	...	...	...	...	...
burg	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

<sup>7</sup> Number of employees covered by social security

Table 2. Probit estimation to explain if an establishment trains or not from Zwick (2005).

<b>Exogenous variables</b>	<b>Coefficients</b>	<b>z-Value</b>
Redundancies expected	0.303***	4.72
Many employees are expected to be on maternity leave	0.332***	3.21
High qualification need exp.	0.565***	6.94
Apprenticeship training reaction on skill shortages	0.222***	4.32
Training reaction on skill shortages	0.652***	13.08
Establishment size 20-199	0.616***	12.67
Establishment size 200-499	1.119***	10.47
Establishment size 500-999	1.239***	7.32
Establishment size 1000 +	1.661***	5.38
Co-determination	0.258***	3.81
Share of qualified employees	0.633***	9.03
State-of-the-art technical equipment	0.199***	4.65
Investor in IT	0.244***	5.29
Collective wage agreement	0.213***	4.82
Apprenticeship training	0.457***	10.01
15 sector dummies and East Germany dummy	Yes	
Pseudo-R <sup>2</sup>	0.32	
Number of observations	5,629	

*Notes:* \*\*\*Significant at the 1% level; the standard errors are heteroscedasticity-corrected.

*Source:* Zwick, 2005: 169

Table 3. Probit estimation to explain if an establishment trains or not after modifications described in Section 5.

<b>Exogenous variables</b>	<b>Coefficients</b>	<b>z-Value</b>
Redundancies expected	0,261***	4,58
Many employees are expected to be on maternity leave	0,252***	2,49
High qualification need expected	0,641***	8,10
Apprenticeship training reaction on skill shortages	0,176***	3,40
Training reaction on skill shortages	0,597***	11,91
Establishment size 20-199	0,683***	15,19
Establishment size 200-499	1,351***	15,71
Establishment size 500-999	1,398***	11,75
Establishment size 1,000 +	1,972***	9,15
Share of qualified employees	0,766***	10,28
State-of-the-art technical equipment	0,175***	4,16
Collective wage agreement	0,245***	5,46
Apprenticeship training	0,420***	9,31
15 sector dummies and East Germany dummy	Yes	
Pseudo-R <sup>2</sup>	0.32	
Number of observations	6,258	

*Notes:* \*\*\*Significant at the 1% level; the standard errors are heteroscedasticity-corrected.

*Source:* IAB Establishment Panel 1997 without new founded establishments and establishments not represented in the Employment Statistics of the German Federal Employment Agency; regression according to Zwick (2005)