Comparing Fully and Partially Synthetic Data Sets for Statistical Disclosure Control in the German IAB Establishment Panel

Jörg Drechsler*, Stefan Bender*, and Susanne Rässler*

* Institute for Employment Research (IAB), Regensburger Straße 104,
90478 Nürnberg, Germany
joerg.drechsler@iab.de, stefan.bender@iab.de, susanne.raessler@iab.de

Abstract. For data sets considered for public release, 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 that address this problem can be found in the literature.
In this paper we discuss the advantages and disadvantages of two approaches that provide disclosure control by generating synthetic data sets: The first, proposed by Rubin (1993), generates fully synthetic data sets while the second suggested by Little (1993) 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 German IAB Establishment Panel and evaluate their quality by comparing regression results from the original data with results we achieve for the same analyses run on the data set after the imputation procedures.

Introduction

In recent years the demand for publicly available micro data increased dramatically. On the other hand, more sophisticated record linkage techniques and the variety of databases readily available to the public may enable an ill-intentioned data user (intruder) to identify single units in public use files provided by statistical agencies more easily. Since the data usually is collected under the pledge of confidentiality, the agencies have to decide carefully what information they are willing to release. Concerning release on the micro level, all agencies apply some statistical disclosure techniques that either suppress some information or perturb the data in some way to guarantee confidentiality. A certain amount of information loss is common to all these approaches. Thus, the common aim of all approaches is, to minimize this information loss while at the same time

minimizing the risk of disclosure. For that reason, a variety of methods for disclosure control has been developed to provide as much information to the public as possible, while satisfying necessary disclosure restrictions (Willenborg and de Waal, 2001, Abowd and Lane, 2004). Especially for German establishment data sets a broad literature on perturbation techniques with different approaches can be found (for example Brand 2000, Brand 2002, Brand et al. 1999, Gottschalk 2005, Ronning and Rosemann 2006, Ronning et al. 2005, Rosemann 2006).

A new approach to address this problem was suggested by Rubin in 1993: Generating fully synthetic data sets to guarantee confidentiality. His idea was 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 data sets are released to the public. Because all imputed values are random draws from the posterior predictive distribution of the missing values given the observed values, disclosure of sensitive information is nearly impossible, especially if the released data sets don't contain any real data. Another advantage of this approach is the sampling design for the imputed data sets. As the released data sets can be 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 data sets can be biased. Furthermore, specifying a model that considers all the skip patterns and constraints between the variables can be cumbersome if not impossible

To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that bear a high risk of disclosure or for variables that contain especially sensitive information leaving the rest of the data unchanged. This approach, discussed as generating partially synthetic data sets in the literature, has been adopted for some data sets in the US (see for example Abowd and Woodcock, 2001, 2004 or Kennickell, 1997).

In this paper we apply both methods to a German Establishment Survey (the Establishment Survey of the Institute for Employment Research (IAB) and discuss advantages and disadvantages for both methods in terms of data utility and disclosure risk.

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 data sets 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 regression results from the original data with results achieved for the same analyses run on the data set after the imputation procedures. The paper concludes with a general discussion of the findings from this study and their consequences for agencies willing to release synthetic data sets of their data.

2 Multiple Imputation

2.1 Multiple Imputation for Missing Data

Missing data is a common problem in surveys. To avoid information loss by using only completely observed records, several imputation techniques have been suggested. 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 data set 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 data set 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) data sets 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 θ, where θ 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 $\hat{\theta}$, $\hat{V}$ and a normal or Student's t reference distribution. For analysis of the imputed datasets, let $\hat{\theta}_i$ and $\hat{V}_i$ for $i = 1,...,m$ 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 $i = 1,...,m$. 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 $\hat{\theta}_i$, $i = 1,...,m$ is averaged together to obtain the final point estimate. The quantity $\hat{\gamma} = (1+m^{-1})B/T$ estimates the fraction of information about θ 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 θ 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 $v_{RS} = (m-1)\hat{\gamma}^{-2}$ 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 $v_{BR} = (v_{RS}^{-1} + \hat{v}_{obs}^{-1})^{-1}$ for the degrees of freedom in the multiple-imputation analysis, where $\hat{v}_{obs} = (1-\hat{\gamma})(v_{com})(v_{com}+1)/(v_{com}+3)$ and v_{com} denotes the complete-data degrees of freedom.

2.2 Fully Synthetic Data Sets

In 1993, Rubin suggested to create fully synthetic data sets 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 draw simple random samples from these imputed populations for release to the public.

For illustration, think of a data set 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 data set.

Now the synthetic data sets 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 populations. The imputer can make random draws from X in a first step and only impute values of Y for the drawn X.

The analysis of the m simulated data sets follows the same lines as the analysis after multiple imputation (MI) for missing values in regular data sets (see Section 2.1). However, the calculation of the total variance slightly differs from the calculation of 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 data sets. Hence, the variance B between the data sets already reflects the variance within each

imputation. For a formal justification see Raghunathan et al. (2003).

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

$\nu_f = (m-1)(1-r^{-1})^2$ 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:

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

Here, n_{syn} is the number of observations in the released data sets sampled from the synthetic population.

2.3 Partially Synthetic Data Sets

In contrast to the creation of fully synthetic data sets, this approach replaces only observed values for variables that bear a high risk of disclosure (key variables) or very sensitive variables with synthetic values (see for example Reiter 2003). The variables with a high risk of disclosure 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. The imputed 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 or variables that have been synthesized in an earlier step.

Imputations are generated according to the multiple imputation framework as described in Section 2.1, but comparable to the fully synthetic data context, the variance estimation differs slightly from the MI calculations for missing data. Yet, it differs from the estimation in the fully synthetic 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, necessary in the missing data context for averaging over the nonresponse mechanism (Rubin, 1987), is not necessary here, since the selection mechanism, set by the imputer, is not treated as stochastic. For a formal justification, see Reiter (2003). This variance estimate can never be negative, so no adjustments are necessary for partially synthetic data sets. Inferences for θ can be based on a Student's t reference distribution with $\nu_p = (m-1)(1+\frac{W}{B/m}))^2$ degrees of freedom.

3 The Data Sets[1]

For the imputation of the IAB Establishment Panel, we use additional information from the German Social Security Data. In the following Section both data sets 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) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973.[2] 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[3] 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 workplace 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 use 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[4] is based on the employment statistics aggregated via the establishment number as of June 30 of each 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. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 16 classes for the industry[5]. These cells are also used for 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 to include new or

---

1 This chapter follows the description given in Alda, Bender & Gartner (2005).

2 On the structure of the insurance number and on the data office of the pension insurance providers cf. Steeger (2000).

3 An overview of the data is given in Bender, Hass, and Klose (2000), a detailed description can be found in Bender, Hilzendegen, Rohwer, and Rudolph (1996).

4 The approach and structure of the establishment panel are described for example by Bellmann (2002) and Kölling (2000).

5 From 2000 onwards 20 industry classes are used.

reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. An overview of available information in 1997 is listed in the Appendix, Figure 2.

4 Application of the Two Synthetic Data Approaches to the IAB Establishment Panel

4.1 Generating Fully Synthetic Data Sets

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: Stratification by establishment size, region and industry (see Table 4 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 via establishment identification number. After matching, every data set is structured as follows: Let N be the total number of units in the newly generated data set, 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 industry 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} are the hypothetic missing data for the newly drawn values in X (see Figure 1).

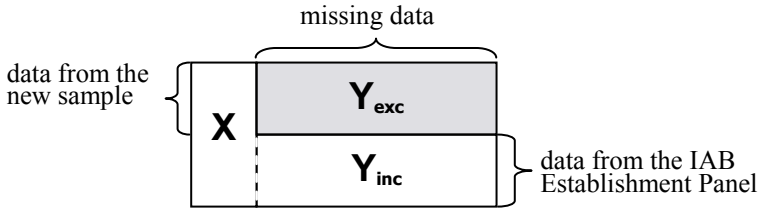

Fig. 1. The full MI approach for the IAB Establishment Panel

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 $f(Y_{exc}|X,Y_{inc})$ 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.

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 newly founded establishments because in that year the questionnaire had a focus on new foundations. Therefore, 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. 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 data sets using the establishment identification number reveals that 278 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, industry and region match in both data sets. Merging indicates that there are some differences between the two records. If the data sets differ, values from the employment statistics are adopted.

Cross tabulation of the stratum parameters for the 7,332 observations in our sample 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 with 20 to 49 employees (see Table 4 in the Appendix). Now, a new data set can be generated easily by drawing establishments from the German Social Security Data according to this matrix.

4.3 Generating Partially Synthetic Data Sets for the IAB Establishment Panel

For this study, we replace only two variables (the number of employees and the industry, coded in 16 categories) with synthetic values, since these are the only two variables that might lead to disclosure in the analyses we use to evaluate the data utility of the synthetic data sets. If we intended to release the complete data to the public, some other variables would have to be synthesized, too. Identifying all the variables that provide a potential disclosure risk is an important and labour intensive task. The research project that deals with this problem is still running. Nevertheless, the two variables mentioned above definitely impose a high risk of disclosure, since they are easily available in public databases and especially large firms can be identified without difficulty using only these two variables.

We define a multinomial logit model for the imputation of the industry code and a linear model stratified by four establishment sizes defined by quartiles for the number of employees.

5 Comparison Between the Original and the Imputed Data Sets

5.1 Data Utility

To create the fully synthetic data sets we draw ten new samples from the German Social Security Data as described in Section 4.2 and impute every sample ten times using chained equations as implemented in the software IVEware by Raghunathan, Solenberger and Hoewyk. For the imputation procedure we use 26 variables from the GSSD and reduce the number of panel variables to be imputed to 48 to avoid multicollinearity problems. For the partially synthetic data sets, we use the same number of variables in the imputation model (26 from the GSSD 48 from the establishment panel), but no samples are drawn from the GSSD, since the original sample is used. We generate the same number of synthetic data sets, but the modelling is performed using own coding in R.

For an evaluation of the data utility of the synthetic data, we compare analytic results achieved with the original data with results from the synthetic data. The first regression is based on an analysis by Thomas Zwick: 'Continuing Vocational Training Forms and Establishment Productivity in Germany' published in the German Economic Review, Vol. 6(2), pp. 155-184 in 2005.

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 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 demonstrates 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 5 in the Appendix for details) shows that establishments increase training if they expect to loose workers. One reason could be that the market for skilled labour 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. This is also the case if establishments give a higher priority to additional apprenticeship training and continuing vocational training efforts instead of hiring externally qualified employees when 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 information and communication technology (IT) it is also essential to offer more training. Collective wage agreements are often associated with fringe benefits such as training, while works councils usually attach high importance to continuing vocational training. Therefore both have a positive effect on the amount of training offered.

In the regression, Zwick uses two variables (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 eliminating newly founded establishments, we have to rerun the regression, using all observations except for newly founded establishments and deleting the two variables which are not part of the 1997 wave. Results from this regression are given in Table 6 in the Appendix and it is evident that the new regression differs only slightly from the original regression. All the variables significant in Zwick's analysis are still significant. Only for the variable "high number of maternity leaves expected", the significance level decreases from 1% to 5%.

For his analysis, Zwick runs the regression only on units with no missing values for the regression variables, loosing all the information on establishments that did not respond to all questions used. This might lead to biased estimates if the assumption of a missing pattern that is completely at random (see for example Rubin, 1987) does not hold. For that reason, we compare the regression results from the synthetic data sets that by definition have no missing values, with the results, Zwick would have achieved if he would have run his regression on a data set with all the missing values multiply imputed. Comparing results from Zwick`s regression run on the original data and on the synthetic data sets are presented in Table 1.

All estimates are very close to the estimates from the real data and except for the variable "high number of maternity leaves expected", for which the significance level decreases to 5% for the fully synthetic data, remain significant on the same level when using the synthetic data. Obviously Zwick would have come to the same conclusions in his analysis, no matter if he would have used the fully synthetic data or the partially synthetic data instead of the real data.

However, if we compare the results from the partially synthetic and the fully synthetic data sets more closely, we see that the estimates from the partially synthetic data sets are closer to the original estimates for most coefficients, although the industry dummies are used as covariates in the regression. Note that the univariate distribution of the industry will always be identical to the true distribution for the fully synthetic data sets, because the industry code is part of the sampling design which is identical for the original and for the fully synthetic data.

Table 1. Comparison between the regression coefficients from the real data and the coefficients from the synthetic data

Exogenous variables	**Coeff. from org. data**	**Fully synthetic data**	**Partially synt. data**	$\beta_{fully} - \beta_{org}$	$\beta_{partially} - \beta_{org}$
Redundancies expected	0.250***	0.251***	0.260***	0.001	0.010
Many employees are expected to be on maternity leave	0.266**	0.244*	0.318**	-0.021	0.052
High qualification need exp.	0.648***	0.625***	0.642***	-0.023	-0.006
Apprenticeship training reaction on skill shortages	0.113*	0.147*	0.118*	0.034	0.005
Training reaction on skill shortages	0.527***	0.523***	0.547***	-0.004	0.019
Establishment size 20-199	0.686***	0.645***	0.702***	-0.041	0.017
Establishment size 200-499	1.355***	1.203***	1.329***	-0.152	-0.027
Establishment size 500-999	1.347***	1.340***	1.359***	-0.007	0.012
Establishment size 1000 +	1.964***	1.778***	1.815***	-0.187	-0.149
Share of qualified employees	0.778***	0.820***	0.785***	0.043	0.008
State-of-the-art technical equipment	0.169***	0.168***	0.170***	-0.001	0.001
Collective wage agreement	0.254***	0.313***	0.268***	0.059	0.014
Apprenticeship training	0.484***	0.406***	0.503***	-0.078	0.020
Number of observations	7,332	7,332	7,332		
15 industry dummies and East Germany dummy					

Notes: *** Significant at the 0.1% level,** Significant at the 1% level, * Significant at the 5% level; the standard errors are heteroscedasticity-corrected.

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

Another way to determine the data utility is to look at the overlap between the confidence intervals for the estimates from the original data and the confidence intervals for the estimates from the synthetic data as suggested by Karr et al. (2006). For every estimate the average overlap is calculated by:

$$J_k = \frac{1}{2}\left(\frac{U_{over,k} - L_{over,k}}{U_{org,k} - L_{org,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}} \right),$$

where $U_{over,k}$ and $L_{over,k}$ denote the upper and the lower bound of the overlap of the confidence intervals for the estimate k, $U_{org,k}$ and $L_{org,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the original data, and $U_{syn,k}$ and $L_{syn,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the synthetic data. This utility measure is more accurate in the sense that it also considers the significance level of the estimate, because estimates with low significance might still have a high confidence interval overlap and by this a high data utility even if their point estimates differ considerably from each other, because the confidence intervals will increase with lower significance. For more details on this method see Karr et al. (2006). Results for our regression example are presented in Table 2.

Table 2: Comparison of the average confidence interval overlap between the original data set and the synthetic data sets

Exogenous variables	**CI overlap for the fully synthetic data**	**CI overlap for the partially synth. data**
Redundancies expected	0.950	0.954
Many employees are expected to be on maternity leave	0.945	0.861
High qualification need exp.	0.923	0.980
Apprenticeship training reaction on skill shortages	0.846	0.973
Training reaction on skill shortages	0.897	0.908
Establishment size 20-199	0.760	0.901
Establishment size 200-499	0.421	0.923
Establishment size 500-999	0.955	0.973
Establishment size 1000 +	0.735	0.792
Share of qualified employees	0.846	0.972
State-of-the-art technical equipment	0.953	0.996
Collective wage agreement	0.675	0.916
Apprenticeship training	0.594	0.883
Average overlap	0.808	0.926

The confidence interval overlap is high for both approaches, often more than 90%, but again the partially synthetic approach yields better results than the fully synthetic approach. The overlap is higher for all estimates except for the variable that indicates whether the establishment expects many employees to be on maternity leave. Especially, if we look at the average CI overlap over all estimates, the improvements for the partially synthetic datasets become clearly evident with an increase of the average overlap from 80.8% to 92.6%.

The advantages of the partially synthetic approach become even more obvious, if we look at a regression of the number of employees transformed on a logarithmic scale on the 16 industry dummies. This model might not be the most interesting model from an economic perspective (the R^2 is low, 0.134 for the original data) but it provides useful information for our study, since it contains exactly the two variables that are synthesized for the partially synthetic approach. Table 3 shows the estimates for both approaches compared to the real estimates and the average confidence interval overlap.

Again, the partially synthetic approach provides better results, although the estimates for the fully synthetic data sets are based on exact marginal distribution for the industry. The deviation from the original estimates is lower for eleven of the 16 estimates. The significance level changes slightly for six estimates when using the fully synthetic data sets, but only for two estimates when using the partially synthetic data sets. The confidence interval overlap is higher for 13 estimates if only some variables are synthesized and the average overlap over all estimates further underlines the higher data utility for partially synthetic data sets.

Table 3: Comparison of the estimates and confidence interval overlaps for a regression of the number of employees on industry dummies (the 16th dummy is the reference category)

Exogenous variables	Coefficients from org. data	Fully synthetic data	Partially synthetic data	CI overlap fully synt. data	CI overlap part. synt. data
Industry dummy 1	-1.606***	-1.794***	-1.531***	0.653	0.834
Industry dummy 2	0.774***	0.757***	0.723***	0.849	0.919
Industry dummy 3	0.098	-0.006	0.148	0.731	0.878
Industry dummy 4	-0.029	-0.204*	0.016	0.470	0.864
Industry dummy 5	-0.96***	-1.162***	-0.923***	0.477	0.908
Industry dummy 6	-1.276***	-1.495***	-1.234***	0.470	0.880
Industry dummy 7	-1.696***	-1.884***	-1.600***	0.507	0.718
Industry dummy 8	-0.505***	-0.286*	-0.605***	0.515	0.786
Industry dummy 9	0.334*	0.362**	0.320*	0.871	0.975
Industry dummy 10	-0.547*	-0.62**	-0.713***	0.914	0.799
Industry dummy 11	-1.431***	-1.531***	-1.342***	0.781	0.781
Industry dummy 12	-0.318**	-0.346***	-0.258*	0.929	0.851
Industry dummy 13	-0.442***	-0.623***	-0.395***	0.537	0.883
Industry dummy 14	-1.641***	-1.844***	-1.529***	0.589	0.731
Industry dummy 15	-0.703***	-0.719**	-0.820***	0.966	0.841
Intercept	4.831***	4.85***	4.779***	0.926	0.774
Average overlap				0.699	0.839

Notes: *** Significant at the 0.1% level,** Significant at the 1% level, * Significant at the 5% level

Of course, partially synthetic data sets will always provide results that are at least as good as the ones from the fully synthetic data set for analyses that are based solely on variables left unchanged in the partially synthetic data. So, in terms of data utility, partially synthetic data sets will outperform fully synthetic data sets in most cases. Furthermore, there might be instances where defining models for all variables is simply impossible, because there are so many logical constraints, bounds, and skip patterns in the data that a useful model cannot be obtained. And if it is possible to come up with a model, the imputed values might be biased and this bias is then introduced in all the other variables that are imputed on a later stage, based on the imputations for this variable.

However, the data utility benefits of the partially synthetic data sets come at the price of an increased disclosure risk that should be discussed in the following Section.

5.2 Disclosure risk

In general, the disclosure risk for the fully synthetic data is very low, since all values are synthetic values. It is not zero however, because, if the imputation model is too good and basically produces the same estimated values in all the synthetic data sets, it doesn't matter that the data is all synthetic. It might look like the data from a potential survey respondent an intruder was looking

for. And once the intruder thinks, he identified a single respondent and the estimates are reasonable close to the true values for that unit, it is no longer important that the data is all made up. The potential respondent will feel that his privacy is at risk. Still, this is very unlikely to occur since the imputation models would have to be perfect and the intruder faces the problem that he never knows if the imputed values are anywhere near the true values.

The disclosure risk is higher for partially synthetic data sets especially if the intruder knows that some unit participated in the survey, since true values remain in the data set and imputed values are generated only for the survey participants and not for the whole population. So for partially synthetic data sets assessing the risk of disclosure is an equally important evaluation step as assessing the data utility. It is essential that the agency identifies and synthesizes all variables that bear a risk of disclosure. A conservative approach would be, to also impute all variables that contain the most sensitive information. Once the synthetic data is generated, careful checks are necessary to evaluate the disclosure risk for these data sets. Only if the data sets proof to be useful both in terms of data utility and in terms of disclosure risk, a release might be considered.

For this study, the disclosure risk evaluation is still in progress. First results show however that the disclosure risk is still very low for the partially synthetic data sets considered here.

6. Discussion and Conclusion

Releasing microdata to the public that guarantees confidentiality for survey respondents on the one hand, but also provides a high level of data utility for a variety of analyses on the other hand is a difficult task. In this paper we discussed two closely related approaches based on multiple imputation: The generation of fully and partially synthetic data sets. While fully synthetic data sets will never contain any originally observed values, original values are replaced only for key identifiers and/or sensitive values in partially synthetic data sets. Since imputed values can be generated for the whole population with fully synthetic data sets, but only for the survey respondents with partially synthetic data sets, knowing that a certain unit participated in a survey will be a benefit for the intruder only for the partially synthetic data sets.

Nevertheless, partially synthetic data sets have the important advantage that in general the data utility will be higher, since only for some variables the true values have to be replaced with imputed values, so by definition the correlation structure between all the unchanged variables will be exactly the same as in the original data set. The quality of the synthetic data sets will highly depend on the quality of the underlying model and for some variables it will be very hard to define good models. But if these variables don't contain any sensitive information or information that might help identify single respondents, why bother to find these models? Why bother to perturb these variables first place? Furthermore, the risk of biased imputations will increase with the number of variables that are imputed. For, if one of the variables is imputed based on a 'bad' model, the biased imputed values for that variable could be the basis for the imputation of another variable and this variable again could be used for the imputation of another one and so on. So a small bias could increase to a really problematic bias over the imputation process.

The findings in this paper underline these thoughts. The partially synthetic data sets provide higher data quality in terms of lower deviation from the true estimates and higher confidence interval overlap between estimates from the original data and estimates from the synthetic data almost for all estimates. Still, this increase of data utility comes at the price of an increase in the risk of disclosure. Although the disclosure risk for fully synthetic data sets might not be zero, the disclosure risk will definitely be higher if true values remain in the data set and the released data is based only on survey participants. Thus, it is important to make sure that all variables that might lead to disclosure are perturbed in a way that confidentiality is guaranteed. This means that a variety of disclosure risk checks are necessary before the data can be released, but this is a problem common to all perturbation methods that are based only on the information from the survey respondents.

Agencies willing to release synthetic public use files will have to consider carefully, which approach suites best for their data sets. If the data consists only of all small number of variables and imputation models are easy to set up, the agencies might consider releasing fully synthetic data sets, since these data sets will provide the highest confidentiality protection, but if there are many variables in the data considered for release and the data contains a lot of skip patterns, logical constraints and questions that are asked only to a small subgroup of survey respondents, the agencies might be better off to release partially synthetic data sets and include a detailed disclosure risk study in their evaluation of the quality of the data sets considered for release.

References

1. Abowd, J.M., Lane, J.: New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. Privacy in Statistical Databases. Springer Verlag, New York (2004) 282-289

2. Abowd, J.M., Woodcock, S.D.: Disclosure limitation in longitudinal linked data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam (2001) 215-277

3. Abowd, J.M., Woodcock, S.D.: Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. Privacy in Statistical Databases. Springer Verlag, New York (2004) 290-297

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

5. Barnard, J., Rubin, D.B.: Small-sample Degrees of Freedom With Multiple Imputation. Biometrika, Vol. 86 (1999) 948-955

6. Bellmann, L.: Das IAB-Betriebspanel - Konzeption und Anwendungsbereiche. Journal of the German Statistical Society, Vol. 86 (2002) 177-188

7. Bender, S., Haas, A., Klose, C.: The IAB Employment Subsample 1975-1995. Journal of Applied Social Science Studies, Vol. 120 (2000) 649-662

8. Bender, S., Hilzendegen, J., Rohwer, G., Rudolph, H.: Die IAB Beschäftigtenstichprobe 1975-1990. Beiträge zur Arbeitsmarkt- und Berufsforschung, No. 197 (1996)

9. Brand, R.: Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. Beiträge zur Arbeitsmarkt- und Berufsforschung, Bd. 237 (2000)

10. Brand, R.: Masking through Noise Addition. Inference Control in Statistical Databases. Springer Verlag, Berlin Heidelberg (2002) 97-116

11. Brand, R., Bender, S., Kohaut, S.: Possibilities for the creation of a scientific-use file for the IAB-Establishment-Panel. Statistical Data Confidentiality Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality Held in Thessaloniki in March 1999. Eurostat, Brüssel 57-74.

12. Gottschalk, S.: Unternehmensdaten zwischen Datenschutz und Analysepotenzial. ZEW Wirtschaftsanalysen, Bd. 76, Nomos Verlag, Baden Baden (2005)

13. Karr, A.F., Kohen, C.N., Oganian, A., Reiter, J.P. and Sanil, A. P.: A framework for evaluating the utility of data altered to protect confidentiality, The American Statistician, Vol. 60, (2006) 224 - 232

14. Kennickell, A.B.: Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. Record Linkage Techniques. National Academy Press, Washington D.C. (1997) 248-267

15. Kölling, A.: The IAB-Establishment Panel. Journal of Applied Social Science Studies, Vol. 120 (2000) 291-300

16. Little, R.J.A.: Statistical Analysis of Masked Data, Journal of Official Statistics, Vol. 9 (1993) 407-426

17. Little, R.J.A., Rubin, D.B.: Statistical Analysis With Missing Data. John Wiley & Sons, Hoboken (2002)

18. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 19 (2003) 1-16

19. Reiter, J.P.: Satisfying Disclosure Restrictions with Synthetic Data Sets. Journal of Official Statistics, Vol. 18 (2002) 531-544

20. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Survey Methodology, Vol. 29 (2003) 181-188.

21. 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

22. 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

23. Rosemann, M.: Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW (2006)

24. Rubin, D.B.: Multiple Imputation in Sample Surveys - a Phenomenological Baysian Approach to Nonresponse. American Statistical Association Proceedings of the Section on Survey Research Methods (1978) 20-40

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

26. Rubin, D.B.: Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 9 (1993) 462-468

27. 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

28. 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

29. Steeger, W.: 25 Jahre Datenstelle der Rentenversicherungsträger (DSRV). Deutsche Rentenversicherung, 10-11/2000, (2000) 648-684

30. Willenborg, L. and de Waal, T.: Elements of Statistical Disclosure Control. Springer-Verlag, New York (2001)

31. Zwick, T.: Continuing Vocational Training Forms and Establishment Productivity in Germany. German Economic Review, Vol. 6(2), (2005) 155-184

Appendix

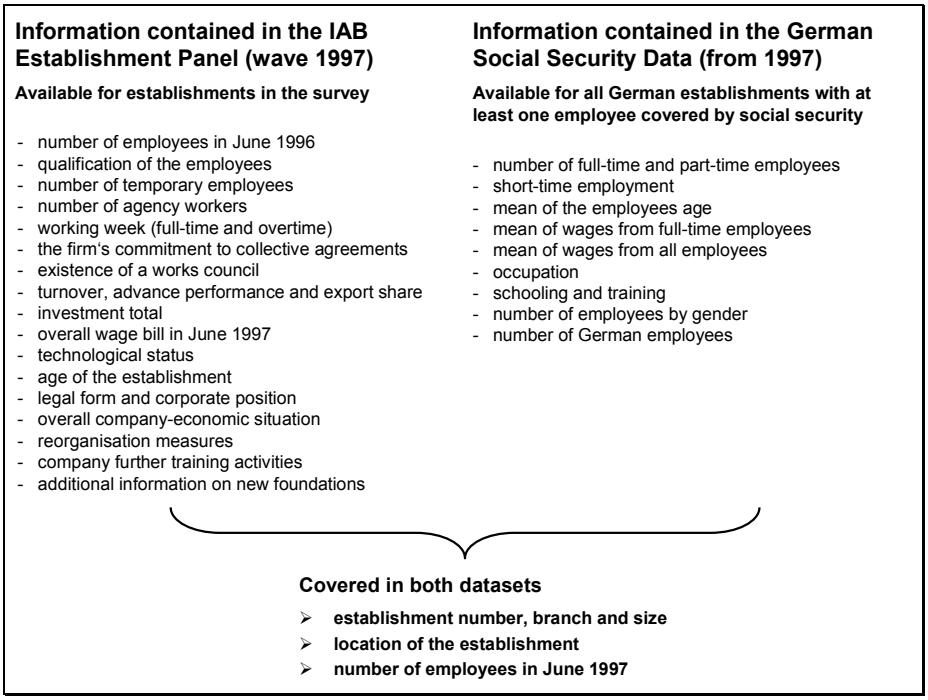

Fig. 2. Data comparison

Table 4. Stratification matrix

Federal state	Branch of trade (16 categories)						
	Establishment size[6]	1 Agriculture, forestry	2 Mining and quarrying	3 Raw material processing	...	16 Non-profit organization	Total
Berlin-West	1 0-4	0	0	1	...	6	42
	2 5-9	2	0	0	...	0	25
	3 10-19	1	0	2	...	3	35
	4 20-49	0	1	1	...	5	29
	5 50-99	0	0	1	...	1	13
	6 100-199	1	0	2	...	2	31
	...	...	...	...	...	...	...
	10 5,000+	0	1	0	...	1	5
	Total	4	3	9	...	40	275
Berlin-East	1 0-4	0	0	0	...	1	52
	2 5-9	0	0	1	...	3	45
	...	...	...	...	...	...	...
	10 5,000+	0	0	0	...	1	1
	Total	3	2	4	...	41	303
Branden-burg	1 0-4	5	0	2	...	8	96
	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...

[6] Number of employees covered by social security

Table 5. Probit estimation to explain if an establishment trains or not from Zwick (2005)

Exogenous variables	**Coefficients**	**z-Value**
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 1,000 +	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		
Pseudo-R^2	0.32	
Number of observations	5,629	

Notes: *** Significant at the 0.1% level, **Significant at the 1% level; the standard errors are heteroscedasticity-corrected.
Source: Zwick (2005), p. 169.

Table 6. Probit estimation to explain if an establishment trains or not after modifications described in Section 5.1

Exogenous variables	**Coefficients**	**z-Value**
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		
Pseudo-R^2	0.32	
Number of observations	6,258	

Notes: *** Significant at the 0.1% level, **Significant at the 1% level; * Significant at the 5% level, the standard errors are heteroscedasticity- corrected.
Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the employment statistics of the German Federal Employment Agency; regression according to Zwick (2005).