Comparing Fully and Partially Synthetic Datasets 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
E-mail joerg.drechsler@iab.de; stefan.bender@iab.de
** Otto-Friedrich-University Bamberg,
Department of Statistics and Econometrics
Feldkirchenstraße 21,
96045 Bamberg, Germany
E-mail susanne.raessler@sowi.uni-bamberg.de

Abstract.[1] For datasets 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 datasets: The first, proposed by Rubin [1], generates fully synthetic datasets while the second suggested by Little [2] imputes values only for selected variables that bear a high risk of disclosure. Changing only some variables in general will lead to higher analytical validity. However, the disclosure risk will also increase for partially synthetic data, since true values remain in the datasets. Thus, agencies willing to release synthetic datasets will have to decide, which of the two methods balances best the trade-off between data utility and disclosure risk for their data. We offer some guidelines to help making this decision.

[1] 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 Federal Statistical Office Germany, the Statistical Offices of the Länder, the Institute for Applied Economic Research (IAW), and the Institute for Employment Research (IAB). We thank J.P. Reiter and two anonymous referees for their helpful comments on the paper and Thomas Zwick for sharing his STATA code.

To our knowledge, the two approaches never haven been empirically compared in the literature so far. 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 results from the original data with results we achieve for the same analyses run on the datasets after the imputation procedures. The results are as expected: In both cases the analytical validity of the synthetic data is high with partially synthetic datasets outperforming fully synthetic datasets in terms of data utility. But this advantage comes at the price of a higher disclosure risk for the partially synthetic data.

1 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 control 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 [3], [4]. Especially for German establishment datasets a broad literature on perturbation techniques with different approaches can be found [5-11].[2]

A new approach to address this problem was suggested by [1]: Generating fully synthetic datasets 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 datasets are

[2] For datasets considered for public release the regulations of the German law have to be fulfilled. In Germany, article 16 (6) of the Federal Statistic Law establishes a scientific privilege for the use of official statistics. Microdata that are de facto anonymous may be disseminated for scientific purposes (so called scientific use files). "Factual anonymity means that the data can be allocated to the respondent or party concerned only by employing an excessive amount of time, expenses and manpower." [12]. The concept of factual anonymity takes into account a rational thinking intruder, who calculates the costs and benefits of the re-identification of the data. Because factual anonymity depends on several conditions and is not further defined by law, it is necessary to estimate the costs and benefits of a re-identification for every dataset with a realistic scenario. [13] started to think of different methods for different scenarios. Additionally attempts were made in a German anonymization project [14-16] to re-identify individuals in a micro dataset under different realistic scenarios.

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 datasets don't contain any real data. Another advantage of this approach is the sampling design for the imputed datasets. As the released datasets 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 datasets can be biased. Furthermore, specifying a model that considers all the skip patterns and constraints between the variables in a large dataset can be cumbersome if not impossible

To overcome these problems, a related approach suggested by [2] 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 datasets in the literature, has been adopted for some datasets in the US [17-20].

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 describes the application of the two multiple imputation approaches for disclosure control to the IAB Establishment Panel. Section 4 evaluates these approaches by comparing the data utility and the disclosure risk for the two settings. The paper concludes with a general discussion of the findings from this study and their consequences for agencies willing to release synthetic versions of their datasets.

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 [21] and discussed in detail in [22,23], 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 [24], Section 10.2.

To understand the procedure of analyzing multiply imputed datasets, think of an analyst interested in an unknown scalar parameter Q, where Q 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 q, an estimate for the variance of q, u and a normal or Student's t reference distribution. For analysis of the imputed datasets, let q_i and u_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 in [21].

For the point estimate, the final estimate simply is the average of the m point estimates $\bar{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i$ with $i=1,...,m$. Its variance is estimated by $T_{MI} = \bar{u}_m + (1+m^{-1})b_m$, where $\bar{u}_m = m^{-1}\sum_{i=1}^{m} u_i$ is the "within-imputation" variance and $b_m = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - \bar{q}_m)^2$ is the "between-imputation" variance. The additional b_m / m reflects the fact that only a finite number of completed-data estimates q_i, $i=1,...,m$ is averaged together to obtain the final point estimate. The quantity $\hat{\gamma} = (1+m^{-1})b_m / T_{MI}$ estimates the fraction of information about Q that is missing due to nonresponse.

Inferences from multiply imputed data are based on $\bar{q}_m$, T_{MI}, and a Student's t reference distribution. Thus, for example, interval estimates for Q have the form $\bar{q}_m \pm t(1-\alpha/2)\sqrt{T_{MI}}$, where $t(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the t distribution. [25] 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). [26] relaxed that assumption 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 ν_{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 draw simple random samples from these imputed populations for release to the public. Most surveys are conducted using complex sampling designs. Releasing simple random samples simplifies research for the potential user of the data, since the design doesn't have to be incorporated in the model. It is not necessary however to release simple random samples. If a complex design is used – we use a stratified sample in our study to take advantage of the efficiency gained by the original stratification – the analyst accounts for the design in the within variance u_i.

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. The approach also applies if there are missing data. Details about generating synthetic data for datasets subject to item nonresponse are described in [27].

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 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 datasets follows the same lines as the analysis after multiple imputation (MI) for missing values in regular datasets (see Section 2.1). The final point estimate is still given by $\bar{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i$. However, the calculation of the total variance slightly differs from the calculation of the total variance in MI settings for treating missing data:

$T_f = (1+m^{-1})b_m - \bar{u}_m$

This difference is due to the additional sampling from the synthetic units for fully synthetic datasets. Hence, the variance b_m between the datasets already reflects the variance within each imputation. For a formal justification see [28].

A disadvantage of this variance estimate is that it can become negative. For that reason, [29] suggests a slightly modified variance estimator that is always positive:

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

Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic 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

$\nu_f = (m-1)(1-r^{-1})^2$ where $r = \frac{(1+m^{-1})b_m}{\bar{u}_m}$.

2.3 Partially Synthetic Datasets

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) or very sensitive variables with synthetic values [30]. 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 while the point estimate stays the same, 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 $T_p = \bar{u}_m + m^{-1}b_m$.

Similar to the variance estimator for multiple imputation of missing data, b_m / m is the correction factor for the additional variance due to using a finite number of imputations. However, the additional b_m, necessary in the missing data context, is not necessary here, since $\bar{u}_m$ already captures the variance of *Q* given the observed data. This is different in the missing data case, where $\bar{u}_m$ is the variance of *Q* given the completed data and $\bar{u}_m + b_m$ is the variance of *Q* given the observed data. For a formal justification, see [30]. This variance estimate can never be negative, so no adjustments are necessary for partially synthetic datasets. Inferences for *Q* can be based on a Student's *t* reference distribution with $\nu_p = (m-1)(1+\frac{\bar{u}_m}{b_m / m}))^2$ degrees of freedom.

3 Application to the IAB Establishment Panel

To generate fully synthetic datasets for the IAB Establishment Panel, we need information from the sampling frame of the Establishment Panel. We obtain this information by aggregating the German Social Security Data (GSSD) to the establishment level. From this aggregated dataset, we can sample new records that provide the basis for the generation of the synthetic datasets.

3.1 Description of the Two Datasets: The German Social Security Data and the IAB Establishment Panel

The German Social Security Data (GSSD) contains information on all employees covered by social security.[3] The notifications of the GSSD include for every employee, among other things, the workplace and the establishment identification number. By aggregating records with the same establishment identification number it is possible to generate establishment information from the GSSD. As we use the 1997 wave of the IAB Establishment Panel for our analysis, data are taken and aggregated from the GSSD for June, 30th 1997 (see Figure 2 in the Appendix for all characteristics used). We use the establishment identification number again to match the aggregated establishment characteristics from the GSSD with the IAB Establishment Panel.

The IAB Establishment Panel is an annually conducted survey, which started in 1993 (West Germany) and 1996 (East Germany). In 2008 the sample contained around 16,000 establishments. The sampling frame of the IAB Establishment Panel[4] is based on the GSSD, which is aggregated via the establishment identification number as of 30 June of each year. Consequently the IAB Establishment 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 establish-

[3] 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. 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 are represented. However, the degree of coverage varies considerably across the occupations and the industries.

[4] The approach and structure of the establishment panel are described for example by [31] and [32].

ment, 16 classes for the region, and 16 classes for the industry[5]. These cells are also used for weighting and extrapolation of the sample. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or 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.

3.2 Generating Fully Synthetic Datasets

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. Every stratum contains the same number of units as the observed data from the 1997 wave of the Establishment Panel.

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 establishment births. 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 establishments 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

Merging the GSSD and the IAB Establishment Panel using the establishment identification number reveals that 278 units from the panel are not included in the GSSD[6]. 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 datasets. Merging indicates that there are some differences between the two records. If the datasets differ, values from the GSSD are adopted.

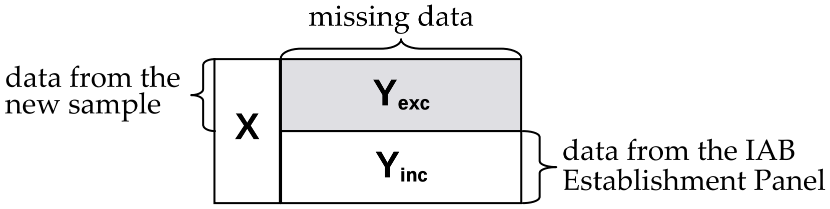
[5] From 2000 onwards 20 industry classes are used.

[6] There are several possible reasons for this, e.g. re-organization of the firm leading to new establishment identification numbers, coding errors, or delays in the notifications for an establishment in the GSSD.

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. Now, a new dataset can be generated easily by drawing establishments from the German Social Security Data according to this matrix.

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 (from the GSSD), region and industry and the other variables added from the German Social Security Data (see Figure 2 in the Appendix). Note that the variable establishment size is included in both, the GSSD and the establishment panel. These two variables need not necessarily be identical, since they are reported at different points in time. However, we use the establishment size from the GSSD as a very strong predictor when synthesizing the establishment size in the establishment panel. 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).

Figure 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 GSSD and all originally observed values from the Establishment Panel are omitted and only the imputed values for the Panel are released. Results from an analysis on these released data can be compared with the results achieved with the real data.

To create the fully synthetic datasets we draw ten new samples from the German Social Security Data and impute every sample ten times using a sequential imputation approach as implemented in the software IVEware by Raghunathan, Solenberger and Hoewyk [33]. This software uses an iterative algorithm called sequential regression multiple imputation (SRMI, [34]) that is based on the ideas of Gibbs sampling and avoids otherwise necessary assumptions about the joint distribution of the missing data given the observed data. Imputations are generated variable by variable where the missing values for any variable Y_k are imputed by drawing from the conditional distributions of $(Y_k | Y_{-k})$, where Y_{-k} represents all variables in the dataset except Y_k.. This allows for different imputation models for each variable. Continuous variables can be imputed with a linear model, binary variables can be imputed using a logit model, etc. Under some regularity assumptions iterative draws from these conditional distributions will converge to draws from the joint multivariate distribution of the data. Since most of the continuous variables like establishment size are heavily skewed, these variables are transformed by taking the cubic root before imputation to get rid of the skewness. In general, all variables are used as predictors in the imputation models in hopes of reducing problems from uncongeniality [35]. In the multinomial logit model for the categorical variables some explanatory variables are dropped for multicollinearity reasons.

The imputations described in this Section are performed in two stages. A new sample is drawn from the GSSD on stage one and given the drawn values from the GSSD, new values are multiply imputed for the Establishment Panel on stage two. The imputations within the same sample from stage one are correlated so that inferential methods must account for that correlation. Thus, the estimators described in Section 2.2 have to be modified as suggested by [36].

Let m be the number of new samples to be drawn and r be the number of imputations within each sample ($m=r=10$ in our case). Let $q^{(i,j)}$ and $u^{(i,j)}$ be the point estimate and its variance estimate from the imputed dataset j in sample i, with $i=1,\ldots,m$ and $j=1,\ldots,r$ Let $\bar{q}_r^{(i)} = \sum_{j=1}^{r} q^{(i,j)} / r$, and $\bar{q}_m = \sum_{i=1}^{m} \bar{q}_r^{(i)} / m$. Let $b_m = \sum_{i=1}^{m} (\bar{q}_r^{(i)} - \bar{q}_m)^2 / (m-1)$ and $w_r^{(i)} = \sum_{j=1}^{r} (q^{(i,j)} - \bar{q}_r^{(i)})^2 / (r-1)$. Finally, let $\bar{u}_m = \sum_{i=1}^{m} \sum_{j=1}^{r} u^{(i,j)} / (mr)$.

The final point estimate in this context equals $\bar{q}_m$, while the estimate for its variance is given by $T_{2step} = (1+m^{-1})b_m + (1-r^{-1})\bar{w}_m - \bar{u}_m$, where $\bar{w}_m = \sum_{i=1}^{m} w_r^{(i)} / m$.

For modest m and r inferences can be based on a t-distribution with υ_{2step} degrees of freedom, where
$$\upsilon_{2step} = \left(\frac{((1+m^{-1})b_m)^2}{(m-1)T_{2step}^2} + \frac{((1-r^{-1})\bar{w}_m)^2}{(m(r-1))T_{2step}^2} \right)^{-1}.$$

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 (Figure 2 in the appendix provides a broad description of the information contained in these variables)[7].

[7] A detailed description of all variables can be obtained from the authors on request.

3.3 Generating Partially Synthetic Datasets 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 datasets. 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. For the partially synthetic datasets, we use the same number of variables in the imputation model (26 from the GSSD 48 from the establishment panel), but all imputations are generated on one stage since the original sample is used and no additional samples are drawn from the GSSD. We generate the same number of synthetic datasets, but the modelling is performed using own coding in R.

4 Data Utility versus Disclosure Risk

4.1 Data Utility

For an evaluation of the 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 Zwick [37].

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, 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 shows that establishments increase training if they expect to loose workers. In the regression, Zwick uses two variables (investment in IT and the codetermination 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. We find that results from the adjusted regression differ 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%.

The original data from the 1997 wave of the Establishment Panel contains units with missing values for the regression variables. Methods for providing valid inferences for synthetic datasets in the presence of missing data can be found in in [27]. Similar methods still need to be developed for the two stage imputation described in this paper. Basically, a different two step imputation procedure is necessary in this context. All missing values are multiply imputed on stage one and on the second stage all confidential values are replaced with synthetic values in every fully imputed dataset. Since we are only interested in the comparison of the partially and the fully synthetic approach and the amount of missingness is low - most of the times less than 1%, never exceeding 5% - we don't take this extra step here. Instead we impute all missing values only once and treat the imputed data as the true data, that is, we assume for the synthesis that the original data was fully observed and don't account for the additional uncertainty due to imputation of the missing values. We use the same models for the imputation of the missing values and the generation of the synthetic data. This can lead to overoptimistic results, since the model for the synthesis will automatically be the correct model for the originally missing values. However, the main aim of this paper is not to evaluate the performance of synthetic data in general. We aim to compare two different synthetic data generation approaches and both approaches are affected in the same way, so our conclusions should not be influenced by this effect. Comparing results from Zwick`s regression run on the fully imputed data and on the synthetic datasets are presented in Table 1.

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 synhetic data	Partially synthetic data	β_{fully} - β_{org}	$\beta_{partially}$ - β_{org}
Redundancies expected	0.253***	0.251***	0.264***	-0.002	0.011
Many employees expected to be on maternity leave	0.262**	0.244*	0.314**	-0.018	0.052
High qualification need exp.	0.646***	0.625***	0.639***	-0.021	-0.007
Apprenticeship training reaction on skill shortages	0.113*	0.147*	0.112*	0.034	-0.001
Training reaction on skill shortages	0.540***	0.523***	0.543***	-0.017	0.003
Establishment size 20-199	0.684***	0.645***	0.701***	-0.039	0.016
Establishment size 200-499	1.352***	1.203***	1.343***	-0.149	-0.009
Establishment size 500-999	1.346***	1.340***	1.367***	-0.006	0.020
Establishment size 1000 +	1.955***	1.778***	1.776***	-0.177	-0.180
Share of qualified employees	0.787***	0.820***	0.785***	0.033	-0.002
State-of-the-art technical equipment	0.171***	0.168***	0.174***	-0.003	0.004
Collective wage agreement	0.255***	0.313***	0.268***	0.058	0.013
Apprenticeship training	0.490***	0.406***	0.507***	-0.084	0.017
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 [37]

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 datasets more closely, we see that the estimates from the partially synthetic datasets 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 datasets, because the industry code is part of the sampling design which is identical for the original and for the fully synthetic data.

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 [39]. 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 from the original and from the synthetic data 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 standard error of the estimate, because estimates with large standard errors 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 the standard error of the estimate. For more details on this method see [39]. Results for our regression example are presented in Table 2.

Table 2. Comparison of the average confidence interval overlap between the original dataset and the synthetic datasets

Exogenous variables	**CI overlap for the fully synthetic data**	**CI overlap for the partially synthetic data**
Redundancies expected	0.942	0.948
Many emp. expect. to be on maternity leave	0.955	0.861
High qualification need exp.	0.929	0.976
Appr. training reaction on skill shortages	0.843	0.996
Training reaction on skill shortages	0.913	0.982
Establishment size 20-199	0.768	0.902
Establishment size 200-499	0.417	0.932
Establishment size 500-999	0.944	0.953
Establishment size 1000 +	0.723	0.723
Share of qualified employees	0.880	0.993
State-of-the-art technical equipment	0.958	0.977
Collective wage agreement	0.682	0.927
Apprenticeship training	0.557	0.899
Average overlap	0.809	0.928

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 80.9% to 92.8%.

The advantages of the partially synthetic approach become even more obvious, if we compute the average number of employees for each of the 16 industries. Note that this analysis is based only on 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.

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)

	Aver. number of employees from org. data	Fully synthetic data	Partially synthetic data	CI overlap fully synthetic data	CI overlap part. synthetic data
Industry 1	71.47	84.41	82.36	0.591	0.784
Industry 2	839.11	754.91	852.88	0.688	0.872
Industry 3	681.07	633.67	593.10	0.593	0.801
Industry 4	642.86	644.39	649.64	0.657	0.971
Industry 5	174.46	194.72	187.38	0.592	0.819
Industry 6	108.89	121.89	120.69	0.637	0.769
Industry 7	117.08	120.31	119.61	0.657	0.878
Industry 8	548.67	446.59	512.99	0.614	0.890
Industry 9	700.70	674.80	713.39	0.687	0.963
Industry 10	546.97	474.26	487.68	0.628	0.871
Industry 11	118.64	111.67	130.98	0.679	0.792
Industry 12	424.31	365.74	425.21	0.626	0.919
Industry 13	516.74	546.13	551.92	0.731	0.800
Industry 14	128.09	144.24	158.99	0.686	0.665
Industry 15	161.98	164.18	238.09	0.890	0.637
Industry 16	510.84	455.42	439.33	0.611	0.566
Average overlap				0.660	0.812

For the point estimates, we get mixed results. Nine of the sixteen estimates are closer to the true value for the partially synthetic datasets. But again it is important to note that the estimates for the fully synthetic datasets are based on exact marginal distribution for the industry since industry is one of the sampling indicators and we don’t synthesize industry for the new samples. So we would actually expect a better performance for the fully synthetic approach. Furthermore, inferences obtained from partially synthetic datasets are closer to the true inferences most of the times. The average confidence interval overlap for partially synthetic datasets is 81.2% whereas the overlap for the fully synthetic datasets is only 66.0%. There is a lot of variability in the establishment size within each industry. This means that the variance estimate u_i for the mean estimator in every synthetic dataset will be high, too. Since we subtract $\bar{u}_m$(the average over all u_i) in our final estimate for the variance of $\bar{q}_m$ for fully synthetic datasets, we get a negative variance estimate for all 16 estimates displayed in Table 3. Thus, we need adjustments for the two stage variance estimate that are similar to the adjustments described in Section 2.2. Following [36] a conservative but always positive variance estimate is given by $T^*_{2step} = T_{2step} + \bar{u}_m$. Using this adjusted estimate leads to overly wide confidence intervals and by this to a lower overlap with the true confidence intervals. No adjustments are necessary for the partially synthetic variance estimate, again an advantage for partially synthetic datasets.

Of course, partially synthetic datasets should always provide results that are at least as good as the ones from the fully synthetic dataset for analyses that are based solely on variables left unchanged in the partially synthetic data. So, in terms of data utility, partially synthetic datasets will outperform fully synthetic datasets in most cases. Furthermore, there might be instances where defining imputation 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 datasets come at the price of an increased disclosure risk that should be discussed in the following Section.

4.2 Disclosure risk

In general, the disclosure risk[8] for the fully synthetic data is very low, since all values are synthetic values. Still, it is not zero: For most establishment surveys the probability of inclusion depends on the size of the establishment and sometimes can be close to 1 for the largest establishments. This means that the additional protection offered in the fully synthetic approach by drawing new samples from the sampling frame that is very effective for household surveys can be very modest for larger establishments. A possible intruder can be confident that large establishments in the released synthetic data represent establishments that were also included in the original survey. For a detailed discussion of the potential disclosure risk for fully synthetic datasets from this survey, see [38].

Besides this actual risk of disclosure the perceived risk of disclosure also needs to be considered. The released data 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. Nevertheless the disclosure risk in general will be very low since the imputation models would have to be almost perfect and the intruder faces the problem that he never knows (i) if the imputed values are anywhere near the true values and (ii) if the target record is included in one of the different synthetic samples.

The disclosure risk is higher for partially synthetic datasets especially if the intruder knows that some unit participated in the survey, since true values remain in the dataset and imputed values are generated only for the survey participants and not for the whole population. So for partially synthetic datasets 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 datasets. Only if the datasets proof to be useful both in terms of data utility and in terms of disclosure risk, a release should be considered.

[8] There are different possibilities of data disclosure. In general a breach of confidentiality (disclosure of information) means that additional information on an observed unit in a given dataset is gained. One case of disclosure is re-identification. For a detailed discussion about identification see [40-42].
A re-identification of firms is only possible if the following conditions are fulfilled:
- The information on the firm is known by the intruder (additional information).
- The firm of interest is in the data.
- The two datasets have some variables in common.
- It is possible to combine the variables in common so that an unequivocal match results.
- The intruder is sure (at least subjectively) that the connection is correct [7].

To evaluate disclosure risks for the partially synthetic datasets in our study, we compute probabilities of identification by following the approach of [41]. Related approaches are described by [42,44-47], but only [41] in combination with [44] is directly applicable to partially synthetic datasets, thus we will focus on their ideas in the remainder of this Section. Roughly, in this approach we mimic the behavior of an ill-intentioned user of the released data who possesses the true values of the quasi-identifiers for selected target records (or even the entire database). To illustrate, suppose the malicious user has a vector of information, *t*, on a particular target unit in the population which may or may not correspond to a unit in the *m* released simulated datasets, $D = \{ D^{(1)}, . . . D^{(m)}\}$. Let t_0 be the unique identifier (e.g., establishment name) of the target, and let d_{j0} be the (not released) unique identifier for record *j* in *D*, where $j = 1,\ldots, s$. Let *M* be any information released about the simulation models. The malicious user's goal is to match unit *j* in *D* to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in D$. Let *J* be a random variable that equals *j* when $d_{j0} = t_0$ for $j \in D$ and equals $s + 1$ when $d_{j0} = t_0$ for some $j \notin D$. The malicious user thus seeks to calculate the $Pr(J = j|t,D,M)$ for $j = 1, . . . , s + 1$. He or she then would decide whether or not any of the identification probabilities for $j = 1,\ldots , s$ are large enough to declare an identification. Let Y_{rep} be the synthesized records in *Y*. Because the malicious user does not know the actual values for Y_{rep}, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record in *D* we compute

$$\Pr(J = j \mid t,D,M) = \int \Pr(J = j \mid t,D,Y_{rep},M)\Pr(Y_{rep} \mid t,D,M)dY_{rep} \qquad (1)$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|t,D,M)$. First, sample a value of Y_{rep} from $Pr(Y_{rep}|t,D,M)$. Let Y_{new} represent one set of simulated values. Second, compute $Pr(J = j|t,D, Y_{rep} = Y_{new},M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y_{new} are collected values. This two-step process is iterated *R* times, where ideally *R* is large, and (1) is estimated as the average of the resultant *R* values of $Pr(J = j|t,D, Y_{rep} = Y_{new},M)$. When *M* has no information, the malicious user can treat the simulated values as plausible draws of Y_{rep}.

Following [41], we quantify disclosure risk with summaries of these identification probabilities. It is reasonable to assume that the malicious user selects as a match for *t* the record *j* with the highest value of $Pr(J = j|t,D,M)$, if a unique maximum exists. We consider two disclosure risk measures: the *expected match risk* and the *true match risk*. To calculate these measures, we need some further definitions. Let c_j be the number of records in the dataset with the highest match probability for the target t_j for $j = 1,\ldots, s$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The expected match risk can now be defined as $\sum_j (1/c_j)I_j$. When $I_j = 1$ and $c_j > 1$, the contribution of unit j to the expected match risk reflects the intruder randomly guessing at the correct match from the c_j candidates. The true match risk equals $\sum_j K_j$.

We assume that the intruder knows which establishments are included in the survey and their true values for the number of employees and industry. This is a conservative scenario but gives, in some sense, an upper bound on the risk for this level of intruder knowledge. Intruders might also know other variables on the file, in which case the agency may need to synthesize them as well. The intruder computes probabilities using the approach outlined above. We assume that the agency does not reveal the synthesis model to the public, so that the only information in M is that employee size and industry were synthesized. For a given target t, records from each $D^{(i)}$ must meet two criteria to be possible matches. First, the record's synthetic industry code exactly matches the target's true industry code. Second, the record's synthetic number of employees lies within an agency-defined interval around the target's true number of employees. Acting as the agency, we define the interval as follows. We divide the transformed (true) number of employees into twenty quantiles and calculate the standard deviation of the number of employees within each quantile. The interval is $t_e \pm sd_s$, where t_e is the target's true value and sd_s is the standard deviation of the quantile in which the true value falls. When there are no synthetic records that fulfill both matching criteria, the intruder matches only on the industry code. We use 20 quantiles because this is the largest number of groups that guarantees at least some variation within each group. Using a larger number of quantiles results in groups with only one value of employment, which forces exact matching for targets in those quantiles. On the other hand, using a small number of quantiles does not differentiate adequately between small and large establishments. For small establishments, we want the potential matches to deviate only slightly from the original values. For large establishments, we accept higher deviations.

Given this matching scenario the expected match risk and the true match risk both would be 139, i.e. the intruder would get 139 true correct single matches. There is no obvious common pattern for the identified records. Neither for the region nor for the industry the distribution of the identified records differs significantly from the distribution in the underlying data. The identified records contain very small and very large establishments. However, as one might expect, the actual risk of disclosure depends on establishment size. While only 1.38% of the establishments with less than 100 employees are identified, this rate increases to 1.87% for establishments with 100-1,000 employees and to 5.21% for establishments with more than 1,000 employees. Considering the fact that the intruder matches on 7,332 records and never knows which of the 7,330 single matches he obtains actually are correct matches the risk is very moderate. Especially since these measures are based on the very conservative assumptions that (i) the intruder knows who participated in the survey and (ii) has exact information on the industry code and the establishment size for all the survey participants. If the agency deems the risk of disclosure still too high, it might broaden the industry codes or suppress this information completely in the released file. Another possibility would be to use less detailed models for the large establishments to ensure a higher level of perturbation for these records.

As an alternative, the agency might consider releasing fully synthetic datasets instead. So far there are no comparable disclosure risk measures for fully synthetic datasets in the literature, but as discussed above the disclosure risks should be further reduced compared to partially synthetic datasets.

5 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 datasets. While fully synthetic datasets will never contain any originally observed values, original values are replaced only for key identifiers and/or sensitive values in partially synthetic datasets. Since imputed values can be generated for the whole population with fully synthetic datasets, but only for the survey respondents with partially synthetic datasets, knowing that a certain unit participated in a survey will lead to greater risk for the partially synthetic datasets.

Nevertheless, partially synthetic datasets 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 dataset. The quality of the synthetic datasets 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 datasets 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 for most 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 datasets might not be zero, the disclosure risk will definitely be higher if true values remain in the dataset 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 imputed 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 datasets. 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 datasets, since these datasets 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 datasets and include a detailed disclosure risk study in their evaluation of the quality of the datasets considered for release.

References

[1] D.B. Rubin: Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 9, pages 462-468, 1993

[2] R.J.A. Little: Statistical Analysis of Masked Data, Journal of Official Statistics, Vol. 9, pages 407-426, 1993

[3] L. Willenborg and T. de Waal: Elements of Statistical Disclosure Control. Springer-Verlag, New York, 2001

[4] J.M. Abowd and J. Lane: New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. Privacy in Statistical Databases. LNCS, vol. 3050, Springer Verlag, New York, pages 282-289, 2004

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

[6] R. Brand: Masking through Noise Addition. In: Domingo-Ferrer, J.(ed.) Inference Control in Statistical Databases. LNCS, vol. 2316. Springer Verlag, Berlin Heidelberg, pages 97-116, 2002

[7] R. Brand, S. Bender, and S. Kohaut: Possibilities for the Creation of a Scientific-Use File for the IAB-Establishment-Panel. Statistical Office of the European Communities (Ed.): Statistical Data Confidentiality - Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality held in Thessaloniki in March 1999, Eurogramme, pages 57-74, 1999

[8] S. Gottschalk: Unternehmensdaten zwischen Datenschutz und Analysepotenzial. ZEW Wirtschaftsanalysen, Bd. 76, Nomos Verlag, Baden Baden, 2005

[9] G. Ronning and M. Rosemann: 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, pages 207-216, 2006

[10] G. Ronning, M. Rosemann, and H. Strotmann: Post-Randomization under Test: Estimation of a Probit Model. Journal of Economics and Statistics, Vol. 225, pages 544-566, 2005

[11] M. Rosemann: Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW, 2006

[12] P. Knoche: Factual Anonymity of Microdata from Household and Person-related Surveys - The release of Microdata Files for Scientific Purposes, Proceedings of the International Symposium on Statistical Confidentiality, Dublin, Eurostat, pages 407 – 413, 1993

[13] G. Paass: Disclosure risk and disclosure avoidance for microdata, Journal of Business and Economic Statistics, 6, pages 487-500, 1988

[14] U. Blien, W. Müller, and H. Wirth: Identification Risk for Microdata stemming from Official Statistics, Statistica Neerlandica, 46, 1, pages 69-82, 1992

[15] W. Müller, U. Blien, P. Knoche, H. Wirth, and others: Die Faktische Anonymität von Mikrodaten, in: Statistisches Bundesamt (ed.), Schriftenreihe Forum der Bundesstatistk, Band 19, Stuttgart: Metzler-Poeschel, 1991

[16] W. Müller, U. Blien, and H. Wirth: Identification Risks of Microdata, Evidence from experimental studies, Sociological Methods & Research, 24, 2, pages 131-157, 1995

[17] J.M. Abowd and S.D. Woodcock: Disclosure limitation in longitudinal linked data. Confidentiality, Disclosure, and Data Access: Theory and Prac-

tical Applications for Statistical Agencies. North-Holland, Amsterdam, pages 215-277, 2001

[18] J.M. Abowd, and S.D. Woodcock: Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. Privacy in Statistical Databases. LNCS, vol. 3050, Springer Verlag, New York, pages 290-297, 2004

[19] J.M. Abowd, M. Stinson, and G. Benedetto: Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at http://www.bls.census.gov/sipp/synth data.html., 2006

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

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

[22] D.B. Rubin: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York, 1987

[23] D.B. Rubin: The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys. The American Statistician, Vol. 58, pages 298-302, 2004

[24] R.J.A. Little, R.J.A., D.B. Rubin, D.B.: Statistical Analysis With Missing Data. John Wiley & Sons, Hoboken, 2002

[25] D.B. Rubin, D.B., N. Schenker, N.: Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. Journal of the American Statistical Association, Vol. 81, pages 366-374, 1986

[26] J. Barnard and D.B. Rubin: Small-sample Degrees of Freedom With Multiple Imputation. Biometrika, Vol. 86, pages 948-955, 1999

[27] J.P. Reiter: Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodology Vol. 30, pages 235 – 242, 2004

[28] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin: Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 19, pages 1-16, 2003

[29] J.P Reiter: Satisfying Disclosure Restrictions with Synthetic Datasets. Journal of Official Statistics, Vol. 18, pages 531-544, 2002

[30] J.P Reiter: Inference for partially synthetic, public use microdatasets. Survey Methodology, Vol. 29, pages 181-188, 2003

[31] G. Fischer, F. Janik, D. Müller, and A. Schmucker: The IAB Establishment Panel – from Sample to Survey to Projection, FDZ-Methodenreport, No. 1, 2008

[32] A. Kölling: The IAB-Establishment Panel. Journal of Applied Social Science Studies, Vol. 120, pages 291-300, 2000

[33] T.E. Raghunathan, P. Solenberger, J. van Hoewyk: IVEware: Imputation and Variance Estimation Software, Available at: http://www.isr.umich.edu/src/smp/ive/, 2002

[34] T.E. Raghunathan, J.M. Lepkowski, J. van Hoewyk, and P. Solenberger: A multivariate technique for multiply imputing missing values using a series of regression models. Surv. Methodol. 27, pages 85-96, 2001

[35] X.-L. Meng: Multiple-imputation inferences with uncongenial sources of input (disc:P558-573). Statistical Science 9, pages 538–558, 1994

[36] J.P Reiter and J. Drechsler: Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. Tech. rep., IAB Discussion Paper, No.20, 2007

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

[38] J. Drechsler, A. Dundler, S. Bender, S. Rässler, T. Zwick: A new approach for disclosure control in the IAB establishment panel—multiple imputation for a better data access, Advances in Statistical Analysis, 2008 (to appear)

[39] A.F. Karr, C.N Kohen, A. Oganian, J.P. Reiter, and A.P. Sanil: A framework for evaluating the utility of data altered to protect confidentiality, The American Statistician, Vol. 60, pages 224 – 232, 2006

[40] J.G. Bethlehem, W. J. Keller and J. Pannekoek: Disclosure control of microdata. Journal of the American Statistical Association 85, pages 38-45, 1990

[41] J.P. Reiter: Estimating risks of identification disclosure in microdata, Journal of the American Statistical Association, 100, pages 1103-1112, 2005

[42] C.J. Skinner: The probability of identification: applying ideas from forensic statistics to disclosure risk assessment. Journal of the Royal Statistical Society A, 170, Part. 1, pages 195-212, 2007

[43] R. Mitra and J.P. Reiter: Adjusting survey weights when altering identifying design variables via synthetic data. In: Domingo-Ferrer, J., Franconi, L.

(eds.) Privacy in Statistical Databases 2006, LNCS, vol. 4302, Springer, Heidelberg, pages 177-188, 2006

[44] J.P. Reiter and R. Mitra: Estimating Risks of Identification Disclosure in Partially Synthetic Data, Journal of Privacy and Confidentiality, forthcoming

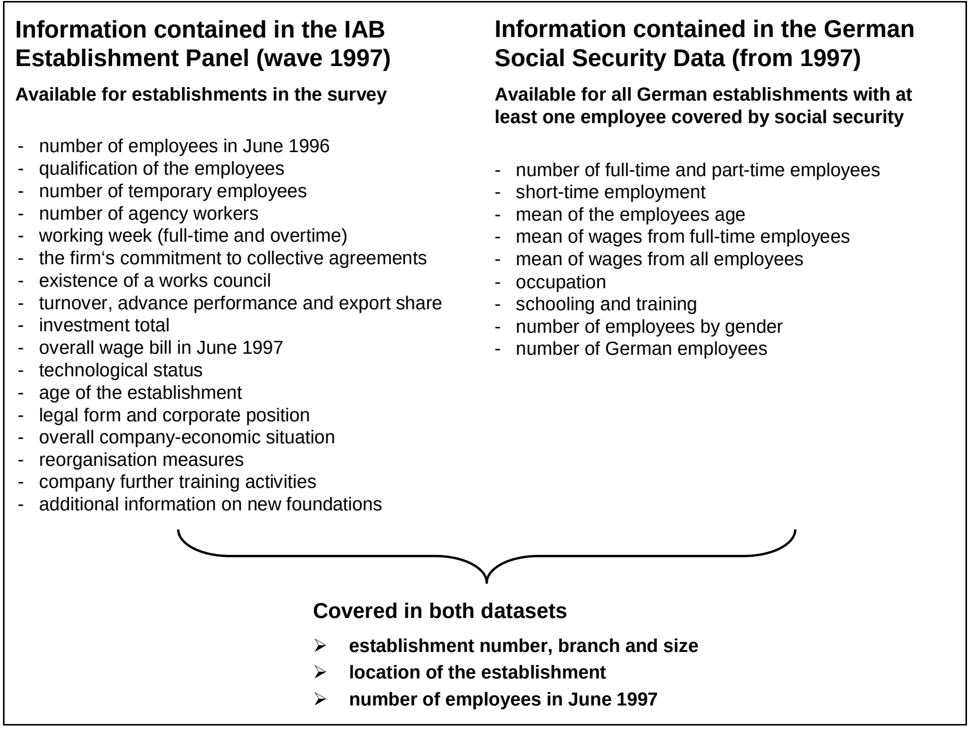
[45] G.T. Duncan and D. Lambert: The Risk of disclosure for microdata. Journal of Business and Economic Statistics 7, pages 207-217, 1989

[46] S.E. Fienberg, U.E. Makov, and A.P. Sanil: A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data, Journal of Official Statistics, Vol. 13, pages 75-89, 1997

[47] J. Drechsler and J.P. Reiter: Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey, Tech. Rep., Institute for Employment Research, 2008

Appendix

Figure 2. Data comparison



Information contained in the IAB Establishment Panel (wave 1997)

Available for establishments in the survey

- number of employees in June 1996
- qualification of the employees
- number of temporary employees
- number of agency workers
- working week (full-time and overtime)
- the firm's commitment to collective agreements
- existence of a works council
- turnover, advance performance and export share
- investment total
- overall wage bill in June 1997
- technological status
- age of the establishment
- legal form and corporate position
- overall company-economic situation
- reorganisation measures
- company further training activities
- additional information on new foundations

Information contained in the German Social Security Data (from 1997)

Available for all German establishments with at least one employee covered by social security

- number of full-time and part-time employees
- short-time employment
- mean of the employees age
- mean of wages from full-time employees
- mean of wages from all employees
- occupation
- schooling and training
- number of employees by gender
- number of German employees

Covered in both datasets

- **establishment number, branch and size**
- **location of the establishment**
- **number of employees in June 1997**