**WP.11**
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

# COMPARING FULLY AND PARTIALLY SYNTHETIC DATA SETS FOR STATISTICAL DISCLOSURE CONTROL IN THE GERMAN IAB ESTABLISHMENT PANEL

## Supporting Paper

Prepared by Jörg Drechsler and Stefan Bender, Institute for Employment Research (IAB) and
Susanne Rässler, Otto-Friedrich-University Bamberg, Germany

# 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
[**] Otto-Friedrich-University Bamberg, Department of Statistics and Econometrics, Feldkirchenstraße 21, 96045 Bamberg, Germany, susanne.raessler@sowi.uni-bamberg.de

**Abstract:** 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. Changing only some variables in general will lead to higher analytical validity. However, the disclosure risk will also increase for partially synthetic data sets since true values remain in the data. Thus, agencies willing to release synthetic data sets 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.

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. The results are as expected: In both cases the analytical validity of the synthetic data is high with partially synthetic data sets outperforming fully synthetic data sets in terms of data utility. But this advantage comes at the price of a higher disclosure risk for the partially synthetic data.

## 1 Introduction

A new approach for statistical disclosure control 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. However, the quality of this method strongly depends on the accuracy of the model used to impute the "missing" values.

To overcome this problem, 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 an establishment survey of the German Institute for Employment Research (IAB) and discuss advantages and disadvantages for both methods in terms of data utility and disclosure risk.

# 2 Application of the Two Synthetic Data Approaches to the IAB Establishment Panel

## 2.1 The IAB Establishment Panel

The IAB Establishment Panel is based on the employment statistics aggregated via the establishment number as of 30 June of each year. Consequently the panel only includes establishments with at least one employee covered by social security. For the imputation of the IAB Establishment Panel, we use additional information from the German Social Security Data (GSSD). The basis of the GSSD is the integrated notification procedure for the health, pension and unemployment insurances. We use the establishment identification number to match the selected establishment characteristics aggregated from the employment register with the IAB Establishment Panel.

## 2.2 Generating Fully Synthetic Data Sets

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. After the imputation procedure, all original observations from the Establishment Panel are omitted and only the imputed values are kept for analysis.

## 2.3 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. 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.

# 3 Comparison Between the Original and the Imputed Data Sets

## 3.1 Data Utility

To create the fully synthetic data sets we draw ten new samples from the German Social Security Data (GSSD) 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, 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 comparison 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. 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. 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.

| **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 | | |

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

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 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 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.

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 data sets become clearly evident with an increase of the average overlap from 80.8% to 92.6%.

| Exogenous variables | CI overlap for the fully synthetic data | CI overlap for the partially synthetic 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 |

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

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 15 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.

| Exogenous variables | Coefficients from org. data | Fully synhetic data | Partially synthetic data | CI overlap fully synthetic data | CI overlap part. synthetic 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 |

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)

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

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

### 3.2 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 are 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 should 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.

## 4 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 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.

## References

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

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

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

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

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

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

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

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