WP. 10
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
(Bilbao, Spain, 2-4 December 2009)

Topic (ii): Synthetic and hybrid data

SYNTHETIC DATASETS FOR THE GERMAN IAB ESTABLISHMENT PANEL

Invited Paper

Prepared by Jörg Drechsler, Institute for Employment Research, Germany

Synthetic Datasets for the German IAB Establishment Panel

Jörg Drechsler*

* Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg. (joerg.drechsler@iab.de)

Abstract. Disseminating microdata to the public that provide a high level of data utility while at the same time guaranteeing the confidentiality of the survey respondent is a difficult task. Generating multiply imputed synthetic datasets is an innovative statistical disclosure limitation technique with the potential of enabling the data disseminating agency to achieve this twofold goal. So far, the approach was successfully implemented only for a limited number of datasets in the U.S. In this paper we present the first successful implementation outside the U.S.: The generation of partially synthetic datasets for a German establishment survey, the IAB Establishment Panel. We will describe the synthesis, present our disclosure risk evaluations and provide some first results on the data utility of the generated datasets.

1 Introduction

In 1993 Rubin suggested for the first time to generate multiply imputed synthetic datasets. Specifically, he suggested to treat the survey variables for those units from the sampling frame that did not participate in the survey as missing data and multiply impute them according to the multiple imputation framework (Rubin, 1978). From these fully imputed populations simple random samples can be released to the public. These are called fully synthetic datasets.

However, since this approach generates imputed values for every variable for every unit in the population, a small bias in the imputation models can have severe negative consequences for the quality of the released data. For this reason, Little (1993) suggested to replace only sensitive variables and/or variables that bear a high risk of disclosure with imputed values. With this approach, now called generating partially synthetic datasets, it is not mandatory to replace all units for one variable. The replacement can be tailored only to the records at risk. It might be sufficient for example to replace the income only for units with a yearly income above 100,000 EUR to protect the data. This method guarantees that only those records that need to be protected are altered. Leaving unchanged values in the dataset will generally lead to higher data quality, but releasing unchanged values obviously poses a higher

risk of disclosure. Thus, a careful disclosure risk evaluation is necessary with this approach before any actual data release.

This paper describes the generation of partially synthetic datasets for one wave of the IAB Establishment Panel, a German establishment survey conducted annually by the German Institute for Employment Research (IAB). The actual release of the data is planned for the end of the year 2009.

The remainder of the paper is organized as follows: Section 2 describes how to obtain valid inferences from partially synthetic datasets when the original data is subject to nonresponse. Section 3 introduces the IAB Establishment Panel. Section 4 describes the generation of the partially synthetic datasets for the survey. Section 5 and 6 present some data utility and disclosure risk evaluations. The paper concludes with some final remarks.

2 Inference for partially synthetic datasets when the original data is subject to nonresponse

Most if not all surveys are subject to item nonresponse and even registers can contain missing values, if implausible values are set to missing during the data editing process. Since the generation of partially synthetic datasets is based on the ideas of multiple imputation, it is reasonable to use the approach to impute missing values and generate synthetic values simultaneously. However new combining rules for the variance estimator are necessary in this case:

Let Q be an estimand, such as a population mean or regression coefficient. Suppose that, given the original data, the analyst would estimate Q with some point estimator q and the variance of q with some estimator u. Let $q_i^{(l)}$ and $u_i^{(l)}$ be the values of q and u in synthetic dataset $D_i^{(l)}$, for $l = 1, ..., m$ and $i = 1, ..., r$. The analyst computes $q_i^{(l)}$ and $u_i^{(l)}$ by acting as if each $D_i^{(l)}$ is the genuine data. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} q_i^{(l)}/(mr) = \sum_{l=1}^{m} \bar{q}^{(l)}/m \quad (1)$$

$$\bar{b}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} (q_i^{(l)} - \bar{q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} b^{(l)}/m \quad (2)$$

$$B_M = \sum_{l=1}^{m} (\bar{q}^{(l)} - \bar{q}_M)^2/(m-1) \quad (3)$$

$$\bar{u}_M = \sum_{i=1}^{m}\sum_{i=1}^{r} u_i^{(l)}/(mr) \ . \quad (4)$$

The analyst then can use $\bar{q}_M$ to estimate Q and $T_M = (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M$ to

estimate the variance of $\bar{q}_M$. When n is large, inferences for scalar Q can be based on t-distributions with degrees of freedom

$$\nu_M = \left(\frac{((1+1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\bar{b}_M/r)^2}{m(r-1)T_M^2} \right)^{-1} \tag{5}$$

The variance estimate T_M can become negative, since $\bar{b}_M/r$ is subtracted. In this case Reiter (2008) suggests to use the conservative variance estimator $T_M^{adj} = (1 + 1/m)B_m + \bar{u}_M$. This estimator is equivalent to the variance estimator for multiple imputation for missing data. Consequently the degrees of freedom are given by $\nu_M^{adj} = (m-1)(1+m\bar{u}_M/((m+1)B_M))^2$. Generally negative variances can be avoided by increasing m and r.

3 The IAB Establishment Panel

The IAB Establishment Panel is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, 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 include 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.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry. These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually - since 1996 with over 4,700 establishments in East Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. 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. For a detailed description of the dataset we refer to Fischer *et al.* (2008) or Kölling (2000).

4 Generating synthetic datasets from the multiply imputed IAB Establishment Panel

For brevity, we don't discuss the imputation of the missing values in this paper. We only note that we used own coding in R to generate $m = 5$ datasets. A detailed description of the imputation can be found in Drechsler (2009). For the synthesis, the first and crucial step is to decide which variables need to be synthesized and whether it is necessary to synthesize all records in the dataset. In our project we decided to synthesize a combination of key variables and sensitive variables. Obviously key variables like *establishment size*, *region* and *industry code* need to be protected, since a combination of the three variables would enable the intruder to identify most of the larger establishments, but we also synthesized the most sensitive variables in the dataset like *turnover* or *amount of subsidies received from the government*. Almost all numerical and some of the categorical variables are synthesized. It might have been sufficient to synthesize values only for the larger establishments since the sampling uncertainty and the similarities between establishments will make re-identification very difficult for small establishments. However, we decided to synthesize all records since, given the large amount of information contained in the dataset (close to 300 variables), all records are sampling uniques arguably even population uniques. Of course only a few variables in the dataset can be considered key variables, but once the dataset is released, a survey respondent might try to identify herself in the released dataset. Since the respondent knows all the answers she provided, it will be easy for her to find herself in the dataset. If she realizes that her record is included completely unchanged, she will feel that her privacy is at risk, even if an intruder that will not have the same background information will never be able to identify this respondent. To drive down this perceived risk we decided to synthesize all records in the dataset.

For the synthesis we use the SRMI approach (Raghunathan *et al.*, 2001) with linear regression models for the continuous variables and logit models for the binary variables. Since all records are replaced with imputed values in our synthesis, developing good models is essential. All variables that don't contain any structural missings are used as predictors in the imputation models in hopes of reducing problems from uncongeniality (Meng, 1994). For the synthesis we use several imputation models for every variable whenever possible. Different models are defined for West and East Germany and for different establishment size classes defined by quantiles. Depending on the number of observations that could be used for the modeling, we define up to 8 different regression models.

The standard approach for a model based imputation of categorical variables with many categories is the multinomal/Dirichlet approach (see for example Abowd *et al.* (2006)). The disadvantage of this approach is that covariates can not be incorporated in the model directly. In general, a different model is fit for a large number

of subcategories of the data defined by cross-classifying some of the covariates to preserve the conditional distributions in the defined classes. This approach is impractical if the number of observations in a survey is low, because there will not be enough observations to define suitable models in every subclass for which the marginal distribution should be preserved. For this reason we use CART models as suggested by Reiter (2005) to generate synthetic values for the categorical variables in our dataset.

We generate $r = 5$ datasets for every imputed dataset, i.e. $m * r = 25$ synthetic datasets will be released. Reiter (2008) elaborates on the number of imputations on stage one and two when using multiple imputation for nonresponse and disclosure control simultaneously. He suggests to set $m > r$, especially if the fraction of missing information is large, to reduce variance from estimating missing values. But this approach will increase the risk of negative variance estimates since $\bar{b}_M$ will increase relative to B_M.

In our dataset only 12 variables (out of more than 300) have missing rates above 5%. On the other hand, we always synthesize 100% of the records. In his simulations Reiter (2008) does not find a significant reduction in variance with increasing m compared to r for 100% synthesis paired with low missing rates. On the other hand, the risk of negative variance estimates increases significantly. With these results in mind, we decided to set $m = r$ in our case.

5 Measuring the data utility

We evaluate the data utility of the generated datasets by comparing analytic results achieved with the original (fully imputed) data[1] with results from the synthetic data. To provide realistic analyses, we use two regressions suggested by colleagues at the IAB, who regularly use the panel for applied analyses. The probit regression displayed in Table 1 is adapted from a regression originally based on a different wave of the establishment panel. The dependent variable indicates if an establishment employs part-time employees. The 19 explanatory variables include among others dummies for the establishment size, whether the establishment expects changes in the number of employees, and information on the personnel structure. Since there are still differences within Germany, the results are computed for West Germany (Table 1) and East Germany (Table omitted for brevity) separately. The regression clearly demonstrates the good data quality. All point estimates from the synthetic data are close to the point estimates from the original data and the confidence interval overlap (Karr *et al.*, 2006) in column four that measures how much the confidence intervals from the original and the synthetic data overlap is higher than 90% for most estimates with an average of 90%. We also report the z-scores for

[1]For convenience, we will refer to the dataset with all missing values multiply imputed as the original data from here on.

Table 1: Regression results from a probit regression of *part time-employees (yes/no)* on 19 explanatory variables in West Germany. For the CI length ratio the CI length of the original datasets is in the denominator.

	original data	synth. data	CI over-lap	z-score org.	z-score syn	CI length ratio
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
sh. of emp. with uni. degree	0.319	0.368	0.91	2.18	2.59	0.97
sh. of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
sh. of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
empl. in the last 6 mths	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 mths	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good/very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above coll. wage agr.	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

all regressions, because some researchers are concerned that synthetic datasets will provide valid results for the significant variables, but might provide less acurate results for variables with lower z-scores. From the results it is obvious that this is not true. We also note that the z-scores from the synthetic data are very close to the z-scores from the original data. This is an important result, since model selections are often based on significance levels. The last column reports the 95% confidence interval length ratio with the confidence interval length of the original data in the denominator. Since the multiple imputation procedure for generating synthetic datasets correctly reflects the uncertainty in the imputation models, it can happen that the confidence intervals from the synthetic datasets are much wider an thus less efficient than the confidence intervals from the original data. We find that the intervals are never increased by more than 6%. The results for East Germany (not reported) are more or less identical, with a slightly increased average confidence interval overlap across all estimates of 93%. Only for the variable *share of low qualified workers* we find that the confidence interval length is increased by 19% in the regression for East Germany.

The second regression is an ordered probit regression with the expected employment trend in three categories (increase, no change, decrease) as the dependent

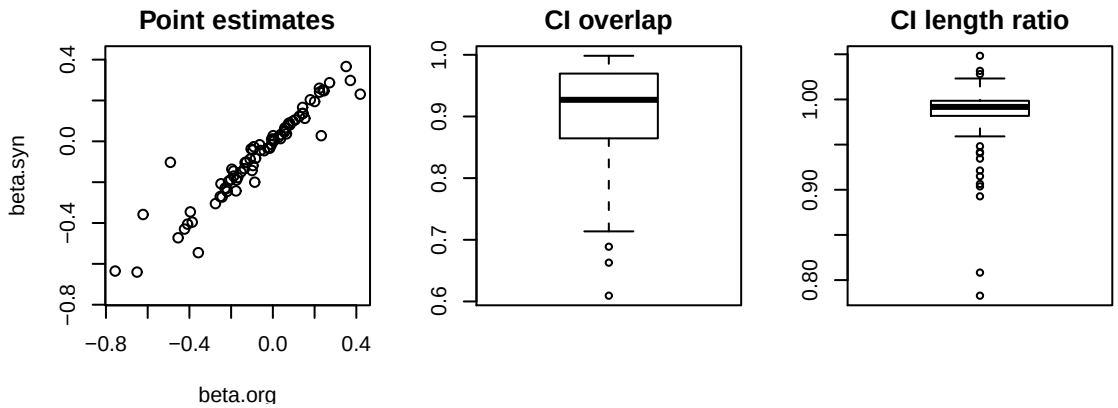

Figure 1: Ordered probit regression of *expected employment trend* on 39 explanatory variables and industry dummies.

variable. In the regression, we use 39 explanatory variables and the industry dummies as covariates. Again the analysis is computed for West Germany and East Germany separately. Figure 1 contains a plot of the original point estimates against the synthetic point estimates and a boxplot for the confidence interval overlap and the confidence interval length ratio. All graphs are based on the 78 estimates from the two regressions. Most of the point estimates in the first graph are close to the 45 degree line indicating that the point estimates from the synthetic data are very close to the point estimates from the original data. But even if the point estimates differ, we find that the data utility measured by the confidence interval overlap is high. The measure never drops below 61% and the median overlap is 92.7%. Thus, even though some estimates are a little off the 45 degree line, the results are close to the original results since these coefficients are estimated with a high standard error. The boxplot of the confidence interval length ratio indicates that we do not loose much efficiency by using the synthetic data instead of the original data. The confidence interval never increases by more than 5% compared to the original data.

Not all users of the data will be interested in multivariate regression analysis. For this reason we also included an evaluation of the data utility for descriptive statistics. The results that further underline the good quality of the datasets are omitted for brevity. A detailed technical report on the synthesis that includes all the results and also provides an example of the limitations of the generated synthetic datasets can be obtained from the author upon request.

6 Assessing the disclosure risk

It is unlikely that an intruder has detailed information about who participated in the survey, thus using the actual true data from the survey for the disclosure risk calculations is an unrealistic conservative scenario. For this reason we apply the disclosure risk measures described in Drechsler and Reiter (2008) that account for the additional uncertainty from sampling. These measures assume, a potential intruder has information about some target records from external databases and tries to identify some units in the survey by matching the target records to the released dataset using the external information. In our application, we are interested in two

measures: The *true match rate*, i.e. the percentage of target records that are matched correctly, and the *false match rate*, i.e. the percentage of times a declared single match is a false match. See Drechsler and Reiter (2008) for a detailed description of the disclosure risk measures.

To obtain a set of target records for which we assume the intruder has some knowledge from external databases that she uses to identify units in the survey, we sample new records from the sampling frame of the survey, the German Social Security Data (GSSD). We sample from this frame, using the same sampling design as for the IAB Establishment Panel: Stratification by establishment size, region and industry code. Merging the stratification matrix from the panel to the stratification matrix of the GSSD reveals that there are 14 stratification cells with positive entries in the panel matrix that are empty in the GSSD matrix. This is a result of the fact that some establishments don't provide answers only for their own entity. They erroneously provide the numbers for the whole concern they belong to instead. By doing so, the establishment might move to another stratification cell that is empty in the original sampling frame. We remove these 14 entries from the stratification matrix of the survey. For the same reason it is possible that some panel cells contain more records than the corresponding GSSD cell. If this happens, we sample all records in this GSSD cell. Overall this leads to a reduction from 15,644 establishments in the original data to 15,624 records in the target sample.

Merging the GSSD and the IAB Establishment Panel using the establishment identification number, we find that 1,360 units from the panel are not included in the GSSD.[2] As a consequence, these records will never appear in the target sample. Since more than 93% of these records are establishments with less than 100 employees, only 4 of them have between 1,000-5,000 employees and non has more than 5,000 employees, we are not concerned that we underestimate the disclosure risk by excluding these records from the target sample.

We find that 917 records from the target sample are also included in the original sample. Table 2 displays the percentage of records from the original dataset that are also included in the target sample for different establishment size classes. As expected, this probability increases with the establishment size. For establishments with less than 100 employees the probability is always less then 10% whereas large establishments with more than 5,000 employees are included in both samples with a probability close to 40%.

For the disclosure scenario we assume, the intruder has information on region, industry code (in 17 categories) and establishment size (measured by the number of employees covered by social security) for her target records and uses this information to identify units in the survey. We further assume that she would consider any record

[2]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.

Table 2: Probabilities to be included in the target sample and in the original sample depending on establishment size.

establishment size class	probability(%)
1-4 employees	0.91
5-9 employees	1.62
10-19 employees	2.87
20-49 employees	4.10
50-99 employees	6.55
100-199 employees	11.39
200-499 employees	16.69
500-999 employees	20.48
1000-4999 employees	31.89
>=5000 employees	39.39

in the synthetic datasets a potential match for a specific target record, if it fulfills two criteria: First, the record's synthetic industry code and region exactly matches the target's true industry code and region. Second, the record's synthetic number of employees lies within a defined interval around the target's number of employees. To define these intervals, we divide the number of employees by the 10 stratification classes for establishment size and calculate the standard deviation within each size class. The interval is $t_e \pm \sqrt{sd_s}$, where t_e is the target's true value and sd_s is standard deviation of the size class in which the true value falls. We investigated several other intervals, e.g. using the standard deviation directly or defining the intervals by 10-20 establishment size classes instead of using the stratification classes. However, we found that the criteria above led to the highest risk of disclosure.

6.1 Log-linear modeling to estimate the number of matches in the population

In general, the intruder will not know the number of records $F_{\mathbf{t}}$ that fulfill the matching criteria in the population to estimate the matching probabilities according to Drechsler and Reiter (2008). One way to estimate the population counts from the released samples was suggested by Elamir and Skinner (2006). We apply this approach to our data assuming that the population counts follow an all-two-way-interactions log-linear model. To simplify the computation, we use the original sample to fit the log-linear model instead of fitting a log-linear model to each synthetic dataset separately. Arguably, this will slightly increase the estimated risk, but we don't expect much difference in the results. A detailed description of the approach and of the evaluation of the quality of the log-linear model can be found in the technical report of the synthesis.

6.2 Results from the disclosure risk evaluations

Summary statistics of the estimated risk are presented in Table 3. Notice that using $\hat{F}_{\mathbf{t}}$ –the estimated population count– instead of $F_{\mathbf{t}}$ gives almost similar results. In both cases, we find that the disclosure risk is very low. Overall less than 1% of the records in the target sample are matched correctly and the false match rate is

Table 3: Disclosure risk summaries for the synthetic establishment panel wave 2007.

	$mean(\hat{F}_{\mathbf{t}})$	$mean(F_{\mathbf{t}})$
true match rate (%)	0.97	0.96
false match rate (%)	98.75	98.76

98.8%. We evaluate the disclosure risk in different establishment size classes and find that the percentage of true matches increase with the establishment size, but never exceeds 7%. We also investigate, if the risks increase, if the intruder only matches, when the average match probability exceeds a predefined threshold γ. Table 4 lists the false match rate and the number of true matches for different threshold values using $F_{\mathbf{t}}$ (there is almost no difference in the results if we use $\hat{F}_{\mathbf{t}}$ instead). The false match rates continually decrease to almost 80% at $\gamma \leq 0.5$. Further reducing γ leads to no improvements in terms of the false match rate. Only for $\gamma \leq 0.1$ the rate drops to 66.7%. At the same time, the true match rate continuously decreases until no true match is found at a threshold of $\gamma = 0$. Since the intruder never knows, which matches actually are true matches, these results indicate that the data seems to be well protected at least under the given assumptions about the information an intruder can gather in her target data. However, large establishments are still at risk in these datasets, because they are identifiable using only the establishment size. For this reason, we protect the largest datasets by drawing from a variance inflated imputation model. A detailed discussion of this extra step is beyond the scope of this paper. Again we refer to the technical report for a complete description of the synthesis.

7 Concluding remarks

Generating multiply imputed synthetic datasets is a promising SDC method. With this approach the user doesn't have to learn complicated adjustments that might differ depending on the kind of analysis the user wants to perform. Furthermore, it is possible with synthetic datasets to account for many real data problems like skip patterns and logical constraints. Most standard SDL techniques can not deal with these problems. Besides, it is very easy to address missing data problems and confidentiality problems at the same time when generating partially synthetic datasets. Since both problems can be handled by multiple imputation, it is reasonable to impute missing values first and then generate synthetic datasets from the multiply imputed datasets. This will actually increase the value of the generated datasets since the fully imputed, not synthesized datasets could be used by other researchers inside the agency that otherwise might not be able to adjust their analyses to account for the missing values properly.

Still, it would be misleading to praise the synthetic data approach as the panacea for data dissemination. It is simply impossible to generate a dataset with any kind of statistical disclosure limitation technique that provides valid results for any potential analysis while at the same time guaranteeing 100% disclosure protection. The synthetic data reflect only those relationships included in the data generation models. If these models do not include some important relationships found in the

Table 4: False and true match rate for different levels of γ.

γ	false match rate	true match rate
1	98.76	0.96
0.9	94.42	0.62
0.8	91.47	0.38
0.7	88.72	0.24
0.6	84.57	0.17
0.5	81.91	0.10
0.4	84.62	0.05
0.3	82.14	0.03
0.2	85.71	0.01
0.1	66.67	0.01
0.0	-	0

original data, the analyst will not detect these relationships in the released data. Therefore it is important that data disseminating agencies also release information about the synthesis models that were used for the generation of the data. With this information the analyst can decide, if the relationship of interest will be reflected in the synthetic data or if he or she will have to apply for access to the original data in the research data center.

The interest in synthetic data is ever growing and many seemingly insurmountable obstacles have been overcome in the last few years. There are still some efforts necessary to make the concept a universal, widely accepted, and easy to implement approach, but the first releases of partially synthetic datasets in the U.S. and in Germany demonstrate that the approach successfully managed the critical step from a pure theoretical concept to practical implementation. Nevertheless, plenty of room remains for future research in this area that will further improve the feasibility of this approach. With the continuous proliferation of publicly available databases and improvements in record linkage technologies releasing synthetic datasets might soon be the only reasonable strategy to balance the trade-off between disclosure risk and data utility when disseminating data collected under the pledge of privacy to the public.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). 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.

Drechsler, J. (2009). Far from normal – multiple imputation of missing values in a German establishment survey. In *Proceedings of the UN/ECE Work Session on Statistical Data Editing and Imputation*, Available at `http://www.unece.org/stats/documents/ece/ces/ge.44/2009/wp.21.e.pdf`.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic

data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 227–238. New York: Springer-Verlag.

Elamir, E. and Skinner, C. J. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics* **22**, 525–539.

Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2008). The IAB Establishment Panel - from sample to survey to projection. Tech. rep., FDZ-Methodenreport, No. 1 (2008).

Karr, A. F., Kohnen, 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* **60**, 224–232.

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

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

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

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Reiter, J. P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters* **78**, 15–20.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 20–34.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.