ESSNET-SDC Deliverable
Report on Synthetic Data Files

Josep Domingo-Ferrer[1], Jörg Drechsler[2] and Silvia Polettini[3]

[1] Universitat Rovira i Virgili,
Dept. of Computer Engineering and Maths,
Av. Països Catalans 26, 43007 Tarragona, Catalonia
e-mail josep.domingo@urv.cat

[2] Institute for Employment Research, D-90478 Nürnberg, Germany,
e-mail joerg.drechsler@iab.de

[3] Università degli Studi di Napoli Federico II,
Dipartimento di Scienze Statistiche,
Via L. Rodinò 22, 80128 Napoli, Italy
e-mail spolettini@unina.it

January 6, 2009

Contents

Abstract

Publication of synthetic —*i.e.* simulated— data is an alternative to masking for statistical disclosure control of microdata. The idea is to randomly generate data with the constraint that certain statistics or internal relationships of the original dataset should be preserved. Several approaches for generating synthetic data files are described in this report. The pros and cons of synthetic data are discussed and some suggestions to Eurostat are made.

Chapter 1

Introduction

Publication of synthetic —*i.e.* simulated— data is an alternative to masking for statistical disclosure control of microdata. The idea is to randomly generate data with the constraint that certain statistics or internal relationships of the original dataset should be preserved.

We give in this introductory chapter an overview of the "historical" literature on synthetic data generation. Two of the more recent approaches are then reviewed in greater detail in Chapters 2 and 3.

1.1 A forerunner: data distortion by probability distribution

Data distortion by probability distribution was proposed in 1985 [24] and is not usually included in the category of synthetic data generation methods. However, its operating principle is to obtain a protected dataset by randomly drawing from the underlying distribution of the original dataset. Thus, it can be regarded as a forerunner of synthetic methods.

This method is suitable for both categorical and continuous attributes and consists of three steps:

1. Identify the density function underlying each of the confidential attributes in the dataset and estimate the parameters associated with that density function.

2. For each confidential attribute, generate a protected series by randomly drawing from the estimated density function.

3. Map the confidential series to the protected series and publish the protected series instead of the confidential ones.

In the identification and estimation stage, the original series of the confidential attribute (*e.g.* salary) is screened to determine which of a set of predetermined density functions fits the data best. Goodness of fit can be tested

by the Kolmogorov-Smirnov test. If several density functions are acceptable at a given significance level, selecting the one yielding the smallest value for the Kolmogorov-Smirnov statistics is recommended. If no density in the predetermined set fits the data, the frequency imposed distortion method can be used. With the latter method, the original series is divided into several intervals (somewhere between 8 and 20). The frequencies within the interval are counted for the original series, and become a guideline to generate the distorted series. By using a uniform random number generating subroutine, a distorted series is generated until its frequencies become the same as the frequencies of the original series. If the frequencies in some intervals overflow, they are simply discarded.

Once the best-fit density function has been selected, the generation stage feeds its estimated parameters to a random value generating routine to produce the distorted series.

Finally, the mapping and replacement stage is only needed if the distorted attributes are to be used jointly with other non-distorted attributes. Mapping consists of ranking the distorted series and the original series in the same order and replacing each element of the original series with the corresponding distorted element.

It must be stressed here that the approach described in [24] was for one attribute at a time. One could imagine a generalization of the method using multivariate density functions. However such a generalization: i) is not trivial, because it requires multivariate ranking-mapping; and ii) can lead to very poor fitting.

Example. A distribution fitting software [7] has been used on the original (ranked) data set 186, 693, 830, 1177, 1219, 1428, 1902, 1903, 2496, 3406. Continuous distributions tried were normal, triangular, exponential, lognormal, Weibull, uniform, beta, gamma, logistic, Pareto and extreme value; discrete distributions tried were binomial, Poisson, geometric and hypergeometric. The software allowed for three fitting criteria to be used: Kolmogorov-Smirnov, χ^2 and Anderson-Darling. According to the first criterion, the best fit happened for the extreme value distribution with modal and scale parameters 1105.78 and 732.43, respectively; the Kolmogorov statistic for this fit was 0.1138. Using the fitted distribution, the following (ranked) dataset was generated and used to replace the original one: 425.60, 660.97, 843.43, 855.76, 880.68, 895.73, 1086.25, 1102.57, 1485.37, 2035.34. □

1.2 Synthetic data by multiple imputation

In the early 1990s, [46] suggested creating an entirely synthetic dataset based on the original survey data and multiple imputation. Rubin's proposal was more completely developed in [35] and a simulation study is given in [36]. In [41] inference for multivariate estimands is discussed and in [40] and [11] applications are given.

We next sketch the operation of the original proposal by Rubin. Consider an original micro dataset X of size n records drawn from a much larger population

of N individuals, where there are background attributes A, non confidential attributes B and confidential attributes C. Background attributes are observed and available for all N individuals in the population, whereas B and C are only available for the n records in the sample X. The first step is to construct m multiply-imputed populations of N individuals, where m is the number of imputations. These populations consist of the n records in X and m matrices of (B, C) data for the $N - n$ non-sampled individuals. If the released data should contain no real data for (B, C), all N values can be imputed. The variability in the imputed values ensures, theoretically, that valid inferences can be obtained on the multiply-imputed population. A model for predicting (B, C) from A is used to multiply-impute (B, C) in the population. The choice of the model is a nontrivial matter. Once the multiply-imputed populations are available, a sample Z of n' records with the same structure as the original sample can be drawn from each population yielding m replicates of (B, C) values. The result are m multiply-imputed synthetic datasets.

More details on synthetic data based on multiple imputation are given in Chapter 3.

1.3 Synthetic data by bootstrap

Long ago, [13] proposed generating synthetic microdata by using bootstrap methods. Later, in [14] this approach was used for categorical data.

The bootstrap approach bears some similarity to the data distortion by probability distribution and the multiple-imputation methods described above. Given an original microdata set X with p attributes, the data protector computes its empirical p-variate cumulative distribution function (c.d.f.) F. Now, rather than distorting the original data to obtain masked data, the data protector alters (or "smoothes") the c.d.f. F to derive a similar c.d.f. F'. Finally, F' is sampled to obtain a synthetic microdata set Z.

1.4 Synthetic data by Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) appears in the literature as another method for generating multivariate synthetic datasets. In [19], the LHS updated technique of [15] was improved, but the proposed scheme is still time-intensive even for a moderate number of records. In [8], LHS is used along with a rank correlation refinement to reproduce both the univariate (*i.e.* mean and covariance) and multivariate structure (in the sense of rank correlation) of the original dataset. In a nutshell, LHS-based methods rely on iterative refinement, are time-intensive and their running time does not only depend on the number of values to be reproduced, but on the starting values as well.

1.5 Partially synthetic and hybrid microdata

Generating plausible synthetic values for all attributes in a database may be difficult in practice. Thus, several authors have considered mixing actual and synthetic data.

One approach has been to create multiply-imputed, partially synthetic datasets that contain a mix of actual and imputed (synthetic) values. The idea is to multiply-impute confidential values and release non-confidential values without perturbation. This approach was first applied to protect the Survey of Consumer Finances [22, 23]. In Abowd and Woodcock [3, 4], this technique was adopted to protect longitudinal linked data, that is, microdata that contain observations from two or more related time periods (consecutive years, etc.). Methods for valid inference on this kind of partial synthetic data were developed in [37] and a non-parametric method was presented in [38] to generate multiply-imputed, partially synthetic data.

Closely related to multiply imputed, partially synthetic microdata is model-based disclosure protection [17, 32]. In this approach, a set of confidential continuous outcome attributes is regressed on a disjoint set non-confidential attributes; then the fitted values are released for the confidential attributes instead of the original values.

A different approach called hybrid masking was proposed in [9]. The idea is to compute masked data as a combination of original and synthetic data. Such a combination allows better control than purely synthetic data over the individual characteristics of masked records. For hybrid masking to be feasible, a rule must be used to pair one original data record with one synthetic data record. An option suggested in [9] is to go through all original data records and pair each original record with the nearest synthetic record according to some distance. Once records have been paired, [9] suggest two possible ways for combining one original record X with one synthetic record X_s: additive combination and multiplicative combination. Additive combination yields

$$Z = \alpha X + (1 - \alpha) X_s$$

and multiplicative combination yields

$$Z = X^{\alpha} \cdot X_s^{(1-\alpha)}$$

where α is an input parameter in $[0, 1]$ and Z is the hybrid record. [9] present empirical results comparing the hybrid approach with rank swapping and microaggregation masking (the synthetic component of hybrid data is generated using Latin Hypercube Sampling [8]).

Another approach to combining original and synthetic microdata is proposed in [47]. The idea here is to first mask an original dataset using a masking method. Then a hill-climbing optimization heuristic is run which seeks to modify the masked data to preserve the first and second-order moments of the original dataset as much as possible without increasing the disclosure risk with respect to the initial masked data. The optimization heuristic can be modified to preserve

higher-order moments, but this significantly increases computation. Also, the optimization heuristic can use a random dataset as an initial dataset instead of a masked dataset; in this case, the output dataset is purely synthetic.

In Chapter 2 we give details on a hybrid approach [31] similar to the one in [9], but based on the IPSO synthetic generator [6].

1.6 Data shuffling

Data shuffling, suggested by Muralidhar and Sarathy (MS) [30], guarantees that marginal distributions in the released datasets will be exactly the same as the marginal distributions in the original data, since the original values are only reordered and no synthetic values are released. Clearly this also means an increased risk of attribute disclosure with respect to properly synthetic methods. Consider for instance the case when the intruder knows which unit in the dataset has the highest income: he or she will simply have to look for the highest income in the dataset to get exactly the true amount, and it does not matter that this income is now attached to another unit. The intruder will also be able to refine his or her guess if he or she knows that the income of a certain unit lies within a defined quantile of the data.

Another limitation of data shuffling can be that it requires ranking of the whole dataset. This can be computationally difficult for large datasets from Censuses with millions of records. Furthermore, MS so far only provided results for artificial datasets where the exact marginal distributions are known for all variables. Even though the method depends only on the rank order of the generated synthetic data that is reverse mapped on the original data, it is unclear how this rank order is affected, if the assumptions about the distributions of the underlying data are not fulfilled.

The most important drawback of this method is however that MS obtained a U.S. patent for their idea and thus the method cannot be implemented without an agreement from the authors. Since it is likely that NSIs would have to pay for this agreement, we do not consider this idea as a realistic alternative to other freely available approaches. We therefore do not discuss this approach any further in this report.

Chapter 2

Information preserving synthetic data

2.1 Information Preserving Statistical Obfuscation (IPSO)

Three variants of a procedure called Information Preserving Statistical Obfuscation (IPSO) are proposed in [6]. The basic form of IPSO will be called here IPSO-A. Informally, suppose that the dataset is divided in two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. A multiple regression of Y on X is computed and fitted Y'_A attributes are computed. Finally, attributes X and Y'_A are released by IPSO-A in place of X and Y.

In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients.

Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete dataset (y, x). If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). The second IPSO method, IPSO-B, modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* statistics $\hat{B}$ and $\hat{\Sigma}$, sufficient for the multivariate normal case, are preserved. This is done by the third IPSO method, IPSO-C.

In [27], a non-iterative method for generating continuous synthetic microdata

is proposed. In a single step of computation, the method exactly reproduces the means and the covariance matrix of the original dataset. The running time grows linearly with the number of records. This method can be regarded as a special case of the IPSO generator.

2.2 The sufficiency based approach: HybridIPSO

The method in this Section, which will be called HybridIPSO, was first presented in [31] and generates hybrid data in a way similar to proposal [9] explained in Section 1.5 above. The difference is that [9] used Latin hypercube sampling (Section 1.4) to generate synthetic data, whereas HybridIPSO described next relies on the aforementioned IPSO generator (Section 2.1).

We have written an implementation of HybridIPSO in "R" which is attached to this report. Some details on that implementation are given in the Annex.

We will first describe the theoretical properties of the method and then the empirical properties.

2.2.1 Theoretical properties

Let $\mathbf{X} = (X_1, \cdots, X_K)$ represent a set of K confidential attributes, let $\mathbf{S} = (S_1, \cdots, S_L)$ represent a set of L non-confidential attributes, and let $\mathbf{Y} = (Y_1, \cdots, Y_K)$ represent the set of K perturbed (*i.e.* hybrid) attributes. Let n represent the number of records in the dataset. Let $\mathbf{\Sigma_{XX}}$, $\mathbf{\Sigma_{SS}}$, and $\mathbf{\Sigma_{YY}}$ represent the covariance matrices of $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{Y}$, respectively. Let $\mathbf{\Sigma_{XS}}$ and $\mathbf{\Sigma_{YS}}$ represent the covariance between $\mathbf{X}$ and $\mathbf{S}$, and $\mathbf{Y}$ and $\mathbf{S}$, respectively. Let $\bar{\mathbf{X}}$, $\bar{\mathbf{S}}$ and $\bar{\mathbf{Y}}$ be the mean vector of $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{Y}$, respectively. Let α be a matrix of size $K \times K$ representing the multipliers of $\mathbf{X}$ and let β be a matrix of size $K \times L$ representing the multipliers of $\mathbf{S}$.

Using the above definition the hybrid values $\mathbf{y}_i$ are generated as:

$$\mathbf{y}_i = \gamma + \mathbf{x}_i\alpha^T + \mathbf{s}_i\beta^T + \mathbf{e}_i, \quad i = 1, \cdots, n \tag{2.1}$$

One requires HybridIPSO to preserve variances, covariances and means, that is,

$$\mathbf{\Sigma_{YY}} = \mathbf{\Sigma_{XX}}$$
$$\mathbf{\Sigma_{YS}} = \mathbf{\Sigma_{XS}}$$
$$\bar{\mathbf{Y}} = \bar{\mathbf{X}}$$

Based on the above preservation requirements, it turns out that:

$$\beta^T = \mathbf{\Sigma_{SS}^{-1}\Sigma_{SX}}(\mathbf{I} - \alpha^T) \tag{2.2}$$

$$\gamma = (\mathbf{I} - \alpha)\bar{\mathbf{X}} - \beta\bar{\bar{S}} \tag{2.3}$$

$$\mathbf{\Sigma_{ee}} = (\mathbf{\Sigma_{XX}} - \mathbf{\Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX}}) - \alpha(\mathbf{\Sigma_{XX}} - \mathbf{\Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX}})\alpha^T \tag{2.4}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Sigma_{ee}}$ is the covariance matrix of the noise terms $\mathbf{e}$.

Thus, α completely specifies the perturbation model shown in Equation (2.1). However, it must be checked that the selected α yields a positive definite covariance matrix for the error terms, that is, that the matrix obtained when such an α is input to Expression (2.4) is positive definite.

Matrix α represents the extent to which the masked data is a function of the original data. There are three possible options for specifying matrix α:

1. α is a diagonal matrix and all the diagonal elements are equal. This would represent a situation where all the confidential attributes are perturbed at the same level. In addition, the value of Y_i is a function only of X_i and does not depend on the value of X_j. Let α represent the value of the diagonal. In this case, it is easy to verify from Equation (2.4) that when $0 \leq \alpha \leq 1$, then $\mathbf{\Sigma_{ee}}$ will be positive definite.

2. α is a diagonal matrix and *not all* the diagonal elements are *equal*. This would represent a situation where the confidential variables are perturbed at different levels. As in the previous case, the perturbed values of a particular variable are a function of the original values of that particular confidential variable and do not depend on other confidential variables. However, in this case, after the specification of α, it is necessary to verify that the resulting $\mathbf{\Sigma_{ee}}$ is positive definite. If not, it may be necessary to respecify α so that the resulting $\mathbf{\Sigma_{ee}}$ is positive definite.

3. α is not a diagonal matrix. In this case, the perturbed values for a particular variable are a function of the original values of that confidential variable as well as the original values of other confidential variables. This is the most general of the specifications and also the most complicated. In [31], this third specification is considered as not entailing any advantages and is thereafter disregarded in the empirical work presented. This approach is only justified though, if only the mean vector and the covariance matrix should be preserved. For any other statistic computed for the confidential variables, the data utility will increase, if α is not a diagonal matrix.

2.2.2 Empirical properties based on applications

The theoretical properties in Section 2.2.1 above hold for a dataset of any size, any underlying distribution and any noise distribution. While these results require no empirical evaluation, a few empirical examples are given in [31] to illustrate the application of the approach proposed there. However, the examples given fail to address the consequences of this perturbation approach for any other statistics than the mean vector and the covariance matrix which are preserved by definition. As of this writing, no other results are available and no agency adopted this approach.

Example 1

As a first example, consider the case where the data provider requests that all variables be perturbed at the same level and specifies that $\alpha = 0.90$. Table 3 in [31] provides the hybrid attributes (Y_1, Y_2) resulting from a dataset consisting of 25 records with 2 nonconfidential attributes (S_1, S_2) and 2 confidential attributes (X_1, X_2) with the following characteristics:

$$\mathbf{\Sigma_{XX}} = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix}$$

$$\mathbf{\Sigma_{SS}} = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$

$$\mathbf{\Sigma_{XS}} = \begin{bmatrix} 0.2 & 0.4 \\ -0.3 & -0.2 \end{bmatrix}$$

The mean vectors $\bar{\mathbf{X}}$ and $\bar{\mathbf{S}}$ were specified as $\mathbf{0}$ resulting in $\gamma = \mathbf{0}$. Based on the data provider's request, α is specified as:

$$\alpha = \begin{bmatrix} 0.900 & 0.000 \\ 0.000 & 0.900 \end{bmatrix} \tag{2.5}$$

In this case, one assumes that the data provider has requested that all variables be perturbed at the same level. Using Equation (2.2), one can compute β as

$$\beta^T = \begin{bmatrix} -0.006250 & 0.043750 \\ -0.028125 & -0.003125 \end{bmatrix}$$

The resulting covariance of the noise term can be computed using Equation (2.4) as

$$\mathbf{\Sigma_{ee}} = \begin{bmatrix} 0.159125 & 0.089063 \\ 0.089063 & 0.172782 \end{bmatrix}$$

It can be verified that $\mathbf{\Sigma_{ee}}$ above is positive definite. The specification of α in Equation (2.5) implies that the perturbed values (Y_1, Y_2) of the confidential attributes are heavily influenced by the original values (X_1, X_2) of the confidential attributes. The non-confidential attributes (S_1, S_2) play a very small role in the perturbation. The extent of the noise term is also relatively small, about 16% for the first and about 17% for the second confidential attribute. The results show that the perturbed values $\mathbf{Y}$ have the same mean vector and covariance matrix as $\mathbf{X}$. It can easily be verified that for many traditional parametric statistical analyses (such as confidence intervals and hypothesis testing for the mean, analysis of variance, regression analysis, multivariate analysis of variance, multivariate multiple regression, etc.) using $(\mathbf{Y}, \mathbf{S})$ in place of $(\mathbf{X}, \mathbf{S})$ will yield exactly the same results.

As a variation of this first example, in [31] it is assumed that the data provider wishes that the coefficient for the first attribute should be 0.9 and that for the second attribute should be 0.2. In this case, the data provider would

like a much higher level of perturbation for attribute X_2 than for attribute X_1. From this specification

$$\alpha = \begin{bmatrix} 0.900 & 0.000 \\ 0.000 & 0.200 \end{bmatrix}$$

The resulting covariance matrix for the noise term is given by

$$\mathbf{\Sigma_{ee}} = \begin{bmatrix} 0.1591 & 0.3844 \\ 0.3844 & 0.8730 \end{bmatrix}$$

It can be verified that the above covariance matrix is not positive definite and it would be necessary for the data provider to consider alternative specifications in this case. In order to maintain the sufficiency requirements, it is necessary that some restrictions be imposed on the selection of α. Extremely disparate specifications such as the one above are likely to create problems as illustrated above.

Example 2

In the second example of [31], one takes

$$\alpha = \begin{bmatrix} 0.800 & 0.000 \\ 0.000 & 0.300 \end{bmatrix}$$

In this case, the two attributes are perturbed at different levels, the first attribute being perturbed less than the second attribute. The resulting values for β and $\mathbf{\Sigma_{ee}}$ can be computed as follows

$$\beta^T = \begin{bmatrix} -0.01250 & -0.19687 \\ 0.08750 & -0.02187 \end{bmatrix}$$

$$\mathbf{\Sigma_{ee}} = \begin{bmatrix} 0.3015 & 0.3563 \\ 0.3563 & 0.8275 \end{bmatrix}$$

It can be verified that $\mathbf{\Sigma_{ee}}$ is positive definite. The results of applying this specification on the original dataset are also provided in Table 3 in [31] (under Example 2). As before, it is easy to verify that the mean vector and covariance matrix of the masked data $(\mathbf{Y}, \mathbf{S})$ are exactly the same as that of the original data $(\mathbf{X}, \mathbf{S})$. Consequently, for those types of statistical analyses for which the mean vector and covariance matrix are sufficient statistics, the results of the analysis using the masked data will yield the same results as the original data.

Unfortunately, as mentioned before, [31] fail to note, that their method only guarantees that the mean vector and covariance matrix of the masked data will be exactly the same as that of the original data. Any analysis that is not based on the first two moments of the distribution or considers subdomains of the data that are not considered when creating the synthetic data will be distorted. It would be more interesting to see the consequences of different levels of α on such analyses.

Example 3

The third example given in [31] shows the case where the perturbed values are generated as a function of only the non-confidential variables and the coefficients of the confidential variables are set to zero ($\alpha = \mathbf{0}$). This choice of α caused HybridIPSO to become exactly the IPSO procedure [6] described in Section 2.1 above: the method does no longer yield hybrid data, it now yields synthetic data. The results of this example are displayed in Table 3 of [31] (under Example 3).

The information loss resulting from the data is also presented in the table measured by the variance of the original and perturbed values. The measure clearly shows an increase in information loss measured using variance $(X - Y)$ as α approaches $\mathbf{0}$, because when $\alpha = \mathbf{0}$ the perturbed values are independent of the original values, which results in synthetic data. By contrast, as α approaches $\mathbf{I}$, the resulting information loss is very low. As observed earlier, the opposite is true for disclosure risk; as α approaches $\mathbf{0}$, disclosure risk decreases and as α approaches $\mathbf{I}$, the disclosure risk is maximal since all confidential values are released unchanged. Thus, the implementation of this procedure needs to be evaluated by considering the trade-off between information loss and disclosure risk.

Chapter 3

Synthetic datasets based on multiple imputation

Generating fully synthetic datasets by multiple imputation was originally proposed by [46]. The basic idea is to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public. Because all 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 do not 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 does not 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 does not 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 [25] 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 in the literature as generating partially synthetic datasets, has been adopted for some datasets in the US [23, 3, 4, 2].

3.1 Fully synthetic datasets

In 1993, [46] 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 there is no need to incorporate the design in the model. It is not necessary however to release simple random samples. If a complex design is used 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 [39]. 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 [45].

To understand the procedure of analyzing fully synthetic 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. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{3.1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{3.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i/m \ . \tag{3.3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_f = (1 + m^{-1})b_m - \bar{u}_m \tag{3.4}$$

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 $\nu_f = (m-1)(1-r_m^{-1})^2$, where $r_m = ((1 + m^{-1})b_m/\bar{u}_m)$. Derivations of these methods are presented in [35]. Extensions for multivariate Q are presented in [41].

A disadvantage of this variance estimate is that it can become negative. For that reason, [36] 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 $\delta = 1$ if $T_f < 0$, and $\delta = 0$ otherwise. Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic population.

3.1.1 Theoretical properties

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 datasets, it does not 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 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.

Regarding data utility, all values in the released fully synthetic datasets are generated from the distribution of Y given X. This means that in theory up to sampling error, the joint distribution of the original data do not differ from the joint distribution of the released data. Thus, any analysis performed on the released data will provide the same results as any analysis performed on the original data. However, the exact multivariate distribution of the data usually is unknown and any model for the joint distribution of the data will likely introduce some bias since the original distribution will rarely follow any multivariate standard distribution. For that reason, it is common to use an iterative algorithm called sequential regression multiple imputation (SRMI, [33]) that is based on the ideas of Gibbs sampling and avoids otherwise necessary assumptions about the joint distribution of the data. Imputations are generated variable by variable where the values for any variable Y_k are synthesized 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. Further refinements for the imputation of heavily skewed variables based on kernel density estimation are given in [49]. But as we said before, 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. 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. But if these variables do not 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 propagate, leading to a really problematic bias over the imputation process. Partially synthetic datasets, described in Section 3.2. can be a helpful solution to overcome these drawbacks since only records that are actually at risk are synthesized.

3.1.2 Empirical properties based on applications

As of this writing no agency actually released fully synthetic datasets. However, [40] and [11] evaluate this approach using real datasets. Reiter generates synthetic datasets for the US Current Population Survey. He computes more than 30 descriptive estimands for different subpopulations of the data and runs a linear and a logit regression on the datasets. He finds that results are encouraging for most estimates. The deviation between the true and the synthetic point estimates is seldom higher than 10% and for many estimands the synthetic 95% confidence interval covers the true estimate in more than 90% of the simulation runs. But he also notes that some descriptive estimates are biased and especially the results for the logit regression are rather poor. He points out that the biased estimates are a result of uncongenialty [28], i.e., the imputation model differs from the analysis model. He notes that if he modifies the imputation model, results from the synthetic data are similar to results from the original data. The bad results in the paper are published as a caveat that defining good imputation models is a very important step in the synthesis process. Reiter also shows that confidentiality would be guaranteed if the data would be considered for release.

[11] generate synthetic datasets for a German establishment survey, the IAB Establishment Panel. They compute several descriptive statistics and run a probit regression originally published in [50] based on the original data of the Establishment Panel. The results from the synthetic datasets are very close to the results from the original data and Zwick would have come to the same conclusions if he would have used the synthetic data instead of the original data. A detailed disclosure risk evaluation in the paper shows that again confidentiality would be guaranteed if the synthetic datasets would be released.

3.2 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 [25]. 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, but compared 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 $T_p = b_m/m + \bar{u}_m$. 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 needed here, since $\bar{u}$ already captures the variance of Q given the observed data. This is different in the missing data case, where $\bar{u}$ is the variance of Q given the completed data and $\bar{u} + b_m$ is the variance of Q given the observed data. Inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)(1+r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. Derivations of these methods are presented in [37]. Extensions for multivariate Q are presented in [41]. The variance estimate T_p can never be negative, so no adjustments are necessary for partially synthetic datasets.

3.2.1 Theoretical properties

Compared to fully synthetic data, partially synthetic datasets present a higher disclosure risk, 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 prove to be useful in terms of both data utility and disclosure risk, a release should be considered. Methods to calculate the disclosure risk for partially synthetic datasets are described in [44] and [12].

3.2.2 Empirical properties based on applications

Some agencies in the US released partially synthetic datasets in the last years. For example, in 2007 the U.S. Census Bureau released a partially synthetic,

public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of social security benefits information and dozens of other highly sensitive variables (www.sipp.census.gov/sipp/synth data.html). A description of the project with detailed discussion of the data utility and disclosure risk is given in [2]. The Census Bureau also created synthesized origin-destination matrices, i.e. where people live and work, available to the public as maps via the web (On The Map, http://lehdmap.did.census.gov/). In the Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, releasing a mixture of imputed values and the not replaced, collected values [23]. For the next release of public use files of the American Communities Survey, the Census Bureau also plans to protect the identities of people in group quarters (e.g., prisons, shelters) by replacing demographic data for people at high disclosure risk with imputations. Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Statistical agencies in Australia, Canada, Germany, and New Zealand ([18]) are also investigating the approach. [10] compare the partially and fully synthetic approach and give guidelines about which method agencies should pick for their datasets. Other applications of partially synthetic data are described in [3, 4, 1, 26, 29, 5, 43].

Chapter 4

Pros and cons of the different approaches

In this report we have discussed data protection methods based on simulation; in general, this approach requires the definition of suitable probability distributions from which to generate synthetic data. Among the many devisable models, we have described in more detail the simulation model proposed by MS for its property of preserving some selected statistical aggregates, and the MI approach for its property of allowing for uncertainty about unknown characteristics of the data generating model.

We now discuss the disclosure protection and data utility offered by these methods, and the relative advantages/disadvantages envisaged in their implementation.

We remark that, as discussed, choice has to be made, between a full simulation or a partial simulation; among the methods presented, we stress that the one proposed by Muralidhar and Sarathy (MS) is designed to produce partially synthetic data, as by design categorical (non confidential) attributes are released unchanged.

Common to any strategy of synthetic data release is the appealing characteristic that, at a first glance, this approach seems to circumvent the re-identification problem: since published records are invented and do not derive from any original record, it might be concluded that no individual can complain from having been re-identified. At a closer look this advantage is less clear, especially when partially synthetic data are released. If, by chance, a published synthetic record matches a particular citizen's non confidential attributes (age, marital status, place of residence, etc.) and confidential attributes (salary, mortgage, etc.), re-identification using the non confidential attributes is easy and that citizen may feel that his confidential attributes have been unduly revealed. In that case, the citizen is unlikely to be happy with or even understand the explanation that the record was synthetically generated. Here lies one of the disadvantages of the approaches (like the one by MS) that only perturb con-

tinuous variables, therefore releasing the non confidential variables unchanged. Strategies based on the definition of joint distribution comprising of both categorical and numerical variables, as is the case for MI models mentioned in section 3, allow synthesizing both types of variables. So if the data releasing agency wants to prevent any form of disclosure (attribute or identity disclosure) the MI approach as well as other simulation strategies based on the definition of a suitable joint distribution may be preferable as they allow to perturb categorical variables, too. For fully synthetic datasets the actual disclosure risk is further reduced, since the synthetic data is generated for new samples from the population and the intruder never knows, if a unit in the released data was actually included in the original data. Partially synthetic datasets on the other hand have the advantage that the synthesis can be tailored specifically to the records at risk. For some datasets it might only be necessary to synthesize certain subsets of the dataset, e.g., the income for survey participants with an income above Euro 100.000. Obviously, the decision which records will remain unchanged is a delicate task and a careful disclosure risk evaluation is necessary in this context.

On the other hand, as with any perturbation method, limited data utility is a problem of synthetic data. Only the statistical properties explicitly captured by the model used by the data protector are preserved. A logical question at this point is why not directly publish the statistics one wants to preserve or simply the parameters of the imputation model rather than release a synthetic micro dataset. Possible defenses against this argument are:

- Synthetic data are normally generated by using more information on the original data than is specified in the model whose preservation is guaranteed by the data protector releasing the synthetic data.
- As a consequence of the above, synthetic data may offer utility beyond the models they exactly preserve.
- It is impossible to anticipate all possible statistics an analyst might be interested in. So access to the micro dataset should be granted.
- Not all users of a public use file will have a sound background in statistics. Some of the users might only be interested in some descriptive statistics and are happy, if they know the right commands in SPSS to get what they want. They will not be able to generate the results if only the parameters are provided.
- The imputation models in most applications can be very complex, because different models are fitted for every variable and often for different subsets of the dataset. This might lead to hundreds of parameters just for one variable. Thus, it is much more convenient even for the skilled user of the data to have the synthesized dataset available.
- The most important reason for not releasing the parameters is that the parameters themselves could be disclosive in some occasions. For that

reason, only some general statements about the generation of the public use file should be released. For example, these general statements could provide information, which variables where included in the imputation model, but not the exact parameters. So the user can judge if her analysis would be covered by the imputation model, but she will not be able to use the parameters to disclose any confidential information.

The sufficiency based approach has the advantage that the mean vector and the variance covariance matrix are preserved exactly, whereas the MI based methods preserve these statistics only in expectation. But on the other hand it is completely unclear what the effects of the sufficiency based approach are on any statistic that does not solely rely on these two statistics. Furthermore, the flexibility of choosing different levels of α for different confidential variables are limited, since only specific combinations of α will guarantee that the matrix of the error terms is positive semi-definite. Especially if there is a large number of confidential variables, finding a legitimate matrix for α that also satisfies the requested different levels of perturbation can be difficult.

As mentioned, the MS approach is designed to reproduce exactly some specified characteristics of the sample. This allows to address the issue of data validity at least partially. At the same time however other characteristics may be lost. Other simulation models such as MI address this problem through the definition of a more comprehensive imputation model. In this case a general limitation is that, in order for the method to produce analytically valid data, goodness of fit of the imputation model must be addressed.

One particular aspect of data validity is that the estimated model should be able to reproduce logical constraints among variables, something that not all methods discussed fulfil. Methods -such as MI- that try to estimate the joint distribution function and not only certain observed statistics may, if carefully designed, achieve this goal and better maintain the multivariate association structure, even though their performance in preserving some specific statistics would often be lower than for methods aiming explicitly to maintain these statistics. On the other hand, the imputation models may be completely general and are therefore not confined to linear regression models. Indeed in order to obtain probabilistic models with a good fit to the data, most often nonlinear techniques such as logistic and multinomial logistic regressions, and even nonparametric models such as kernel density estimators and CART have been introduced in the model building stage. The versatility of the approach is especially useful for public use files, where the goal is to preserve more than just the first two moments of the distribution, e.g., maintain interaction and nonlinear effects.

Furthermore, model based imputation procedures such as MI offer more flexibility if certain constraints need to be preserved in the data. For example non negativity constraints and convex constraints like *total number of employees* $\geq$ *number of part time employees* can be directly incorporated at the model building stage. This is not possible for the Muralidhar-Sarathy synthetic data approaches, so datasets generated with these method will generally fail these constraints.

As a consequence of its greater flexibility, a practical limitation of the model based imputation approach is that the model building stage is a crucial and difficult step of the protection process, and requires the allocation of large resources in terms of time and personnel. A thorough analysis of the data to be released and its expected uses and a careful model selection would allow the statistical agency to define a simulation distribution that may be considered as a general-purpose model. In principle, such a model should be sufficiently general to allow most users to conduct their own analyses almost as if they were using the original data. It is clear however that not all the analyses, especially those on subgroups, may be preserved.

Synthetic microdata would be even further justified if valid analyses could be obtained on a number of subdomains, *i.e.* similar results were obtained in a number of subsets of the original dataset and the corresponding subsets of the synthetic dataset. The Muralidhar-Sarathy approaches can preserve the mean vector and the variance covariance matrix for predefined subsets of the data. However, if the analyst defines other subsets than the predefined subsets, say the average income for survey respondents with an income above Euro 110.000 instead of above Euro 100.000, nothing will be preserved. Partially synthetic or hybrid microdata are more likely to succeed in staying useful for any kind of subdomain analysis. However, when using partially synthetic or hybrid microdata, we loose the attractive feature of purely synthetic data that the number of records in the protected (synthetic) dataset is independent of the number of records in the original dataset. But the disclosure risk will also increase for data shuffling, if predefined subdomain means are to be preserved, since the shuffling algorithm has to be applied to the different subdomains separately and units are only reordered within these subdomains. This may increase the disclosure risk significantly if the values within the subdomains do not differ very much.

Another important question that might be raised and that has not been addressed so far is whether a model that may be selected as a best fitting one under the original data maintains a positive probability of being selected under the simulated data.

Lastly, as [42] points out, the MI approach as well as other model-based simulation methods can be relatively transparent to the public analyst. Metadata about the imputation models can be released and the analyst can judge based on this information if the analysis he or she seeks to perform will give valid results with the synthetic data. For the MS approaches it is very difficult to decide, how much a particular analysis has been distorted.

Chapter 5

Suggestions for Eurostat

Simple methods for synthetic data generation such as those described in Chapters 1 and 2 are rather limited as to the models and statistics they can exactly preserve.

If more flexibility is desired the methods in Chapter 3 can be very useful. The basic MI concepts for continuous and categorical data are straightforward to apply.

As surveys most often collect information on a very large number of variables, irrespective of the approach to protection by simulation (partially or fully synthetic data) it is necessary to define and estimate a multivariate probability distribution. This is a difficult task, and in general cannot be assisted by automatic model selection tools as they are only rarely available. Some experiments have been performed using flexible model selection techniques such as CART ([38]) or Bayesian networks [16]; these techniques however have some drawbacks, with respect to goodness of fit and/or computational tractability/times. In practice, experience has shown that a careful definition of conditional distributions is often a more efficient way to define the multivariate distribution while allowing for constraints and differential relationships among variables within subgroups and to actually simulate from the joint distribution. However the particular order chosen to model conditional distributions may have an impact on the quality of the simulated data, especially in cases when the model does not represent the data adequately. A practical suggestion is to isolate variables that show strong mutual relationships and start from the definition of models for the most relevant survey variables.

In cases where partially synthetic data are released, variables that are not related to the ones that must be simulated can be discarded; this aspect may offer some practical advantages at the model building stage.

For the release of partially synthetic data, it might be deemed appropriate to simulate only a subset of the variables, possibly for a subset of the units. In order to decide which variables should be simulated and which subset of units should undergo partial simulation, a careful risk analysis must be performed. A detailed account of the aspects to take into consideration is reported in [20]. As described

in [20] and [21], high risk units may be highlighted based on density concepts, and considering outlyingness and the possibility of spontaneous identification of some records. The existence of published information about some of the survey variables must also be considered: in case these variables are categorical (e.g. the economic activity) this may lead to the definition of substrata on which the risk is assessed; in case the publicly available variables are continuous, it might be decided to leave them unchanged and to simulate a certain number of other variables. A comprehensive account of the practical implementation of partial imputation referring to data from the Community Innovation Survey (CIS) is contained in [20].

An enhancement of the multiple imputation software IVEware [34] that also allows generating synthetic datasets is planned for 2009. This will enable agencies to generate synthetic datasets even if they are not very familiar with the technical details. However, if the structure of the dataset is very complex including all sorts of skip patterns and constraints, more complex methods are needed, such as CART-based methods ([38]). Wielding such complex methods requires more expertise in SDC than is now available in most European National Statistical Offices, especially the smaller ones. Indeed, a specific R+D project may be needed to apply such complex methods to a specific survey.

On the other hand [48] showed the severe drawbacks of most statistical disclosure control methods that are used mainly because of their ease of implementation. These methods (e.g. single variable swapping, rounding, rank swapping, single-variable microaggregation, and some forms of multivariate microaggregation) fail in terms of data utility and/or disclosure risk even in very simple settings. And what is the use of spending time generating a public use file that nobody will use for its known very limited data utility?

Therefore we suggest that Eurostat, like other accredited agencies such as the US Census Bureau, should consider funding research and development of synthetic or hybrid datasets for their most important surveys. Investment would also be required to train European national statistical offices in the understanding and use of these often complex methods.

ANNEX. Implementation of Muralidhar-Sarathy's hybrid IPSO generator

The HybridIPSO method described in Section 2.2 was implemented by Úrsula González-Nicolás (Universitat Rovira i Virgili) in "R". The implementation is available in the standard format of an "R" package and is attached to this document as a zipfile.

Bibliography

[1] Abowd, J. M. and Lane, J. I. New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 290–297, Berlin Heidelberg, 2004. Springer.

[2] Abowd, J.M., Stinson, M. Benedetto, G. Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program, 2006

[3] J. M. Abowd and S. D. Woodcock. Disclosure limitation in longitudinal linked tables. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 215–278, Amsterdam, 2001. North-Holland.

[4] J. M. Abowd and S. D. Woodcock. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 290–297, Berlin Heidelberg, 2004. Springer.

[5] An, D. and Little, R. Multiple imputation: an alternative to top coding for statistical disclosure control. Journal of the Royal Statistical Society, Series A 170: 923–940, 2007

[6] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.

[7] Crystal.Ball. http://www.cbpro.com/, 2004.

[8] R. Dandekar, M. Cohen, and N. Kirkendall. Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 245–253, Berlin Heidelberg, 2002. Springer.

[9] R. Dandekar, J. Domingo-Ferrer, and F. Sebé. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.

[10] Drechsler, J., Bender, S., Rässler, S. Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, 1(3):105–130, 2008.

[11] Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. A new approach for statistical disclosure control in the IAB establishment panel-Multiple imputation for a better data access, Advances in Statistical Analysis, 92(4):439–458

[12] J. Drechsler and J. Reiter. Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data. In J. Domingo-Ferrer and Y. Saygin, editors, *Privacy in Statistical Databases*, volume 5262 of *Lecture Notes in Computer Science*, pages 227–238, Berlin Heidelberg, 2006. Springer.

[13] S. E. Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report 611, Carnegie Mellon University Department of Statistics, 1994.

[14] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14(4):485–502, 1998.

[15] A. Florian. An efficient sampling scheme: updated Latin hypercube sampling. *Probabilistic Engineering Mechanics*, 7(2):123–130, 1992.

[16] Franconi, L. and Polettini, S. Some experiences at Istat on data simulation. Proceedings of the 56th Session of the International Statistical Institute, Lisboa, Portugal, 1, Invited Paper Meetings, IPM22, 2007.

[17] L. Franconi and J. Stander. A model based method for disclosure limitation of business microdata. *Journal of the Royal Statistical Society D - Statistician*, 51:1–11, 2002.

[18] Graham, P. and Penny, R. Multiply imputed synthetic data files. Tech. rep., University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm, 2005

[19] D. E. Huntington and C. S. Lyrintzis. Improvements to and limitations of Latin hypercube sampling. *Probabilistic Engineering Mechanics*, 13(4):245–253, 1998.

[20] Ichim, D. Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment. Documenti Istat, 2, available at http://www.istat.it/dati/pubbsci/documenti/Documenti/doc_2007/2007_2.pdf 2007

[21] Ichim, D. Disclosure control of business microdata: a density-based approach, 2009, to appear.

[22] A. B. Kennickell. Multiple imputation and disclosure control: the case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*, pages 248–267, Washington DC, 1999. National Academy Press.

[23] A. B. Kennickell. Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In J. Domingo-Ferrer, editor, *Statistical Data Protection*, pages 248–267, Luxembourg, 1999. Office for Official Publications of the European Communities.

[24] C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems*, 10:395–411, 1985.

[25] Little, R.J.A. Statistical Analysis of Masked Data. *Journal of Official Statistics* **9**, 407-426, 1993

[26] Little, R. J. A., Liu, F., and Raghunathan, T. E. Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, 141–152. New York: John Wiley & Sons, 2004

[27] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 298–306, Berlin Heidelberg, 2004. Springer.

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

[29] Mitra, R. and Reiter, J. P. Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., Privacy in Statistical Databases, 177–188. New York: Springer-Verlag, 2006

[30] K. Muralidhar and R. Sarathy. Data shuffling: a new masking approach for numerical data. *Management Science*, 52(5):658–670, 2006.

[31] K. Muralidhar and R. Sarathy. Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy*, 1(1):17–33, 2008. http://www.tdp.cat/issues/tdp.a005a08.pdf.

[32] S. Polettini, L. Franconi, and J. Stander. Model based disclosure protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 83–96, Berlin Heidelberg, 2002. Springer.

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

[34] 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

[35] T. J. Raghunathan, J. P. Reiter, and D. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003.

[36] J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544, 2002.

[37] J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–188, 2003.

[38] J. P. Reiter. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3):441–462, 2005.

[39] Reiter, J.P. Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology* **30**:235–242, 2004

[40] J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168:185–205, 2005.

[41] J. P. Reiter. Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377, 2005.

[42] J. P. Reiter. Letter to the editor. *Journal of Official Statistics*, 24(2):319-321, 2008.

[43] Reiter, J.P., and Drechsler, J. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality, *Statistica Sinica*, to appear

[44] Reiter, J.P., and Mitra, R. Estimating risks of identification disclosure in partially synthetic data, *Journal of Privacy and Confidentiality*, to appear

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

[46] D. B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[47] F. Sebé, J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 163–171, Berlin Heidelberg, 2002. Springer.

[48] Winker, W. E. Examples of Easy-to-implement, Widely Used Methods of Masking for which Analytic Properties are not Justified. Tech. Rep., U.S. Census Bureau Research Report Series, No. 2007-21

[49] Woodcock, S. D. and Benedetto, G. Distribution-Preserving Statistical Disclosure Limitation. Available at SSRN: http://ssrn.com/abstract=931535, 2007

[50] Zwick, T. Continuing vocational training forms and establishment productivity in Germany. German Economic Review 6 (2): 155–184, 2005.