CASC PROJECT

Computational Aspects of Statistical Confidentiality

August 28, 2002

Report on preparation of the data set and improvements on Sullivans algorithm

Ruth Brand
Sarah Giessing

Destatis
Wiesbaden

Deliverable No: 1.1-D1

Stand: 29.08.2002

Tests of the Applicability of Sullivan's Algorithm to Synthetic Data and Real Business Data in Official Statistics

1. Introduction

Empirical economic researchers criticize the lack of access to economic microdata of official statistics: The information potential of surveys conducted by the statistical offices at enterprises and local units is only partially exploited. In principle, statistical offices would like to provide business microdata to empirical economic researchers to expand the uses of these data. As major information providers, the statistical offices support any endeavours to make extensive use of their data holdings. According to the European (and also the German) law however, it is legal only to give access to microdata by way of providing scientists and researchers with 'de facto anonymised' microdata.

In the area of statistics on households and individuals this way of giving access to microdata has been pursued for several years. Where microdata on enterprises and local units are concerned, however, de facto anonymisation is considered to be more difficult. The de facto anonymisation of enterprise and local unit data requires sophisticated techniques but, even internationally, methods are not yet fully explored.

As an important step in the work on this topic so called data perturbation techniques are investigated. One group of these techniques that has been discussed for more than 20 years now is masking the date by adding noise. Several algorithms were developed that have different characteristics. The simplest algorithm consists of adding white noise to the data. More sophisticated methods use more or less complex transformations of the data and more complex error-matrices to improve the results. This paper gives an overview over the algorithm suggested by Sullivan.

From the theoretical point of view this algorithm is of special interest, because it preserves several properties of the data that are often used in empirical analysis. Furthermore it is the only algorithm that gives the opportunity to mask continuous and discrete variables in one step. This can be seen as a special advantage, because real data often consist of both types of variables.

In this report the algorithm and its extension in order to integrate partial masks is described and the results of tests on artificial and real data are discussed. The emphasis of this part of the report lies on the practical applicability of the algorithm to real data sets

This report is organised as follows. In section 2, criteria necessary for applying anonymisation methods in statistical agencies are described. Section 3 gives a detailed description of the algorithm and the extensions made. Empirical tests are shown in section 4. These allow some investigations into the properties of the algorithm in terms of analytical validity, level of protection and practical applicability.

2. Criteria for testing the usefulness of the algorithm

In order to examine the usefulness of a statistical disclosure method one has to consider three main aspects: analytical validity of the protected data, level of protection and practical applicability to real life data.

The first aspect "analytical validity" has often been discussed under the aspect which properties of the original data have to be preserved while applying statistical disclosure control methods. A review of the literature shows that it seems to be necessary to preserve the possibility to obtain unbiased or at least consistent estimates of central sample statistics for disseminating "useful" scientific use files, because empirical, sociological, or economic studies usually evaluate causal hypothesis by multivariate statistics or econometric analysis (see e.g. Brand 2000, Kim 1986, 1990, McGuckin/Nguyen 1990, McGuckin 1993, Winkler 1998). Estimates for these models are usually conducted on the basis of the unweighted sample. Hence it is assumed that the disseminated dataset contains no weights or similar information about the sample design.

Central sample statistics needed for applying most multivariate methods are the sample means and the sample covariance of the masked data. It should be possible to derive consistent estimates for the true means and covariance in terms of the masked variables in order to ensure that standard-techniques, especially OLS-regression-estimates, can be applied to the anonymised data. For those potential users should of the data using standard statistical software packages, it is very desirable that these central statistics can be obtained without correcting calculation formulas.
For examining analytical usefulness one has furthermore to decide whether the descriptive statistics of the original sample should be replicated exactly by the disseminated data set or whether it is sufficient to preserve them only in terms of expected values. In the literature it is often mentioned that values calculated for sample statistics of the anonymised data should be close to those of the original data to obtain similar results (McGuckin/Nguyen 1990). Furthermore it is desirable to preserve at least the univariate distributions because of their usefulness for a general description of the data and the deduction of analytical models.

Hence in this contribution the following aspects of analytical validity are considered in terms of similarity of:

- means and standard-deviations
- univariate distributions
- correlations and
- multivariate distribution of all variables .

This setting stands in line with the criterions used in the literature (see e.g. Kim 1986, 1990, Domingo-Ferrer/Torra 2001, Domingo-Ferrer/Mateo-Sanz 2001). Nevertheless the criteria chosen are not the only ones that can be relevant for anonymising real data. Other aspects may be the third and fourth moments of the multivariate distributions or preserving the conditional distributions of some of the variables while other variables are fixed.

The second aspect that has to be considered is whether the level of protection achieved is sufficient. The level of protection can be measured in different ways: First, on statistics that base on the number of unique records[1], and second on statistics for the number of true matches. The latter approach has to be applied if the data are masked by adding noise, because in this approach it is not intended to reduce the number of unique records. Instead the records will be modified in a way that combining them with additional information will not lead to a re-identification. For applying this method, it has to be defined first which matches are considered "true" and which matching procedures will be used (Probabilistic matching or distance based matching, see e.g. Domingo-Ferrer/Torra 2002).

[1] A record is unique, if it is the only one, which has the combination of values for the variables taken into account.

A detailed analysis of the level of protection is not intended here. Nevertheless, Sullivan's algorithm incorporates a distance based criterion for sufficiency of masking. A simple distance based matching algorithm is used that is integrated in the masking procedure. Subsequently, the vector of distances between every masked record and all original records (m_i) is calculated. It has the typical element: $m_{it} = d_{it}\Sigma_d^{-1}d_{it}$, where d_{it} denotes the vector of differences between records i and t: $d_{it} = z_i - z_t^*$, Σ_d the covariance-matrix of the distance-vectors, and z_i the vector of values for record i and z_i^* the vector of masked values for the i-th record (Sullivan 1989). The criterion chosen for sufficiency is that m_{ii} is not one of the two smallest distances in m_i. Otherwise the mask will be repeated (Sullivan 1989, p.70).
Other criteria for sufficiency of the mask, e.g. absolute distances, could be chosen as well. Nevertheless, for practical reasons this criterion is fixed in the algorithm.

Note, that calculating this distance criterion is very time consuming because the number of calculations needed increases quadratically with the number of records. Hence application of this check seriously affects the number of records that can efficiently be masked on a personal computer.

The third aspect relevant for the work in statistical agencies is whether practical applications can be performed in a reasonable amount of time and whether they need expert knowledge about the data and the Statistical disclosure control methods used. Whether or not statisticians, who are not statistical disclosure control experts, would be able in practice to use a certain masking algorithm may depend on the following issues: the number of parameters necessary for applying the algorithm, stability of the algorithm if unusual parameter values are chosen and/or in situations where data are not well prepared or not prepared properly. Complex data manipulations, which cannot be standardised, and extensive parameter specifications are time-consuming and require experienced users. Application of a method that requires a lot of parameters to be chosen, or is not very robust, may lead to higher costs than a - maybe less optimal - method that can be applied easily.

3. Description of the algorithm suggested by Sullivan

In this section an overview over the main methods of masking by adding noise is given. At first simple addition of random noise is described. On this basis the approach proposed by Sullivan
(Sullivan 1989, Sullivan/Fuller 1989, 1990, Fuller 1993) is discussed.

3.1 Masking by Noise Addition: General Idea

Masking by adding noise was first tested extensively by Spruill (1983). The basic assumption is, that the observed values are realisations of continuous variables: $x_j \sim (\mu_{x_j}, \sigma_{x_j}{}^2)$ j = 1, ..., p. Adding noise means that the vector x_j is replaced by a vector z_j : $z_j = x_j + \varepsilon_j$, where $\varepsilon_j \sim N(0, \sigma_{\varepsilon_j}^2)$ denotes usually normal distributed errors with $Cov(\varepsilon_t, \varepsilon_l) = 0$ for all $t \neq l$ (white noise) generated independently from the original variables. Using matrices, this can be written as:

$$\mathbf{Z} = \mathbf{X} + \mathbf{E}$$

with $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{E} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{E}})$ and $\mathbf{Z}$ denoting the matrix of perturbed values: $\mathbf{Z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}_Z)$, with $\boldsymbol{\Sigma}_{\mathbf{Z}} = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{\mathbf{E}}$.
Usually it is assumed in the literature that variances of the ε_j are proportional to those of the original variables (Spruill 1983, Tendick 1988, Tendick 1991) : $\sigma_{z_j}^2 = \alpha\sigma_{x_j}^2$ with $\alpha > 0$. The parameter α denotes a positive constant used for varying the "amount of noise". Error variances proportional to the variance of the corresponding variable ensure that the relative error is identical for all variables. Furthermore, it is often implicitly assumed that the errors generated for different variables are independent, because the variables are described isolated in literature (Tendick 1991). Under these conditions we obtain for the masked data:

$$\mathbf{\Sigma_Z} = \mathbf{\Sigma} + \alpha \ \mathrm{diag}(\sigma_{x_1}^2, \sigma_{x_2}^2, \Lambda\ , \sigma_{x_p}^2) \quad , \quad \alpha > 0 \quad .$$

This method leads to useful results if relatively small errors are used (Spruill 1983) due to the characteristics of the error-distribution. Obviously, the sample means of the masked data are unbiased estimators for the expectancy values of the original variables due to the error-structure chosen; nevertheless this is not true for the variances and correlations, because the variances are inflated by factor (1+α). This implies that the sample variances of the masked data are asymptotically biased estimators for the variances of the original data.

It has therefore been suggested to generate an error matrix $\mathbf{E}^*$ under the restriction $\mathbf{\Sigma}_{\mathbf{E}^*} = \alpha\mathbf{\Sigma}$ (correlated noise), which implies $\mathbf{\Sigma_Z} = (1+\alpha)\mathbf{\Sigma}$. In this approach all elements of the covariance matrix of the perturbed data differ from those of the original data by factor: (1+α). Hence correlations of the original data can be estimated asymptotically unbiased by the sample correlations of the masked data (see e.g. Kim 1986, 1990). Kim (1990) shows that expected values and covariances of subpopulations can be estimated consistent as long as α is known, too. Furthermore it is possible to obtain consistent estimates of these parameters when only some of the variables used are masked (partial masks, see Kim (1990). These findings illustrate a general advantage of this method: Consistent estimates for several important statistics can be achieved as long as the value of parameter α is known. This means identical results can be achieved on the average, numerical equivalence of every estimate is not guaranteed.

A general disadvantage that can not be adjusted by 'external' information about the masking procedure is that the distributions of the masked variables can not be determined as long as the variables are not normal distributed. The reason for this is that the distribution of the sum of a normal distributed variable (the error-term added) and a not normal distributed variable is not known in general.

Although simple masking by adding correlated noise has some pleasant properties for analysis, it is usually not used due to its very low level of protection (see e.g. Kim/Winkler 1995, 1997, 2001, Tendick 1988, Tendick 1991). Therefore it is primarily a reference framework for studying general problems of adding noise to continuous variables.

Some modifications of the approach are discussed in the literature. The one most often referred to it the one proposed by Kim (1986). This algorithm is based on combining noise addition with linear transformations. The algorithm allows consistent estimates of means, standard-deviations, and correlations and regression analysis for the whole sample and non-random subsamples. Nevertheless, the univariate distributions are not maintained and even the range cannot be preserved. Hence a different approach proposed by Sullivan and Fuller (Sullivan 1989, Sullivan/Fuller 1989, 1990, Fuller 1993) will be discussed in the following section.

3.2 Masking by Noise Addition and Non-linear Transformations

Sullivan (1989) proposes an algorithm that combines masking by noise addition with non-linear transformations. The transformations allow for application to discrete variables and yield preservation of the univariate distributions.

The algorithm proposed by Sullivan consists of several steps (Sullivan/Fuller 1989, Fuller 1993):

1. Calculate the empirical distribution function for every variable,
2. Smooth the empirical distribution function,
3. Convert the smoothed empirical distribution function into a uniform random variable and this into a standard normal random variable,
4. Add noise to the standard normal variable,
5. Back-transformation to values of the distribution function,
6. Back-transformation to the original scale.

During the application of the algorithm a distance criterion is used. It assures that the distance between the masked variables obtained in step four and its "standard normal" counterpart is not one of the two smallest distances. The properties of this more complicated algorithm are determined by the application of all steps. Therefore they will be described in this section in more detail.

Steps 1 and 2 contain different transformations for continuous and discrete variables. For a continuous variable X the empirical distribution is calculated in step 1. In step 2, this distribution is transformed to the so called smoothed empirical distribution Sullivan (1989). Therefore the midpoints between the values X_i will be used as domain limits:

$$x_i = (X_i - X_{i-1})/2 \quad \text{for all} \quad i = 1, \Lambda\ , m \quad ,$$

with $X_0 = 2X_1 - X_2$ and $X_{m+1} = 2X_m - X_{m-1}$. Within these classes the smoothed empirical distribution is calculated:

$$\overline{F}_x(z) = \hat{F}(x_{i-1}) + \frac{\hat{F}(x_i) - \hat{F}(x_{i-1})}{x_i - x_{i-1}}(z - x_{i-1}) \qquad \text{for} \qquad z \in (x_{i-1}, x_i] \quad ,$$

where $\hat{F}(x_i)$ denotes the values of the empirical distribution function at the limits. This equation is calculated for every value of X: $p_i = \overline{F}_x(X_i)$. These are mapped into standard normal values by applying the quantile function of the standard normal distribution function (step 3): $Z_i = \Phi^{-1}(p_i)$.

These transformations are a standardisation for normal distributed variables. Hence correlations of the transformed variables are nearly identical to those of the original variables as long as the variables are jointly normal. If the observed variables are not normally distributed the correlations of the transformed variables differ substantially. The amount of this divergence depends on the differences between the empirical distribution of standardised values and the standard normal distribution.

In order to transform a discrete variable with k possible outcomes, the variable is first split in k-1 Bernoulli variables (Sullivan/Fuller 1990). Second the conditional covariance matrix of all Bernoulli-variables in the data set, given the continuous variables ($m_{dd.cc}$) is calculated:

$$m_{dd.cc} = m_{dd} - m_{cd}' m_{cc}^{-1} m_{cd} \quad ,$$

where m_{cc} denotes the covariance matrix of the continuous variables, m_{dd} the covariance matrix of the binary variables and m_{cd} the matrix of covariances between continuous and binary variables.

Third a matrix of standard normal random numbers F_{dc} is generated, with column vector f_{dct}:

$$f_{dct} = m_{cd}' m_{cc}^{-1} L_{cc}^2 Z_{ct} + m_{dd.cc}^{1/2} e_{d,ct} \quad ,$$

where Z_{ct} denotes the vector of transformed continuous variables stemming from observation t, $L_{cc}^2 = \mathrm{diag}(m_{cc})$ a matrix that has the sample variances on the diagonal and all off-diagonal elements zero and $e_{d,ct}$ a vector of standard normal random numbers.

Although the f_{dct} have nearly the same correlations as the original Bernoulli variables they do not depend on their true values. To link the original values of the Bernoulli variables with f_{dct} further transformations are needed. Therefore the values of the distribution function are determined $g_{dct,j} = \Phi(f_{dct,j})$ for all $j = 1, \Lambda\ , p_d$.

The $g_{dct,j}$ are realisations of uniform random variables. Nevertheless they do not depend on the Bernoulli variables either. Hence the random variable $h_{dct,j}$ is generated that depends on $g_{dct,j}$ and the original variable $x_{dt,j}$:

$$h_{dt,j} = \begin{cases} g_{dtc,j}(1-p_{oj}) & \text{if} \quad x_{dt,j}=0 \\ 1-p_{oj}(1-g_{dtc,j}) & \text{if} \quad x_{dt,j}=1 \end{cases} \qquad j=1,\text{K}\ , p_d \quad ,$$

where p_{oj} denotes the mean of the j-th Bernoulli-variable. In step 3 normalised ranks $R_{d1,j}, \Lambda\ R_{dn,j}$ are assigned starting with the smallest value of $h_{dt,j}$:

$$\widetilde{R}_{dt,j} = \frac{R_{dt,j} - 0.5}{n} \quad .$$

These are transformed in standard normal variables using the quantile function: $Z_{dt,j} = \Phi^{-1}(\widetilde{R}_{dt,j})$.
Combining these with the transformed continuous variables leads to the vector of standard normal variables for every observation: $Z_t = (Z_{ct}' Z_{dt}')$.

In the fourth step of the algorithm the transformed variables Z_t are masked by adding noise. This mask is in principle identical to the simple mask described in section 3.1. Let Z denote the matrix of transformed standard normal variables with row vectors Z_t and the matrix of errors $U^* \sim N(0, M_{ZZ})$. Then the matrix of masked variables is defined as:

$$Z^* = Z + U^* = Z + \sqrt{\alpha}\, UT_{ZZ} \qquad \text{for} \qquad \alpha > 0,$$

where $U \sim N(0, I_{pp})$ and T_{ZZ} denotes a decomposition of the correlation matrix of the transformed data: $T_{ZZ}' T_{ZZ} = P_{ZZ}$. Since the elements of Z and U^* are normally distributed, the masked values Z_t^* are normal: $Z_t^* \sim N(0, M_{Z^*Z^*})$, with $M_{Z^*Z^*} = (1+\alpha) M_{ZZ}$.

In this step of the algorithm partial masks can be integrated easily by setting all errors U for the variables that should not be masked to zero. Hence the "masked values" Z^* are identical to the transformed original values Z for those variables that are not masked. The same idea can be used if some observations should not be masked. This way of including partial masks leads to the covariance matrix:

$$M_{Z^*Z^*} = \begin{pmatrix} (1+\alpha) M_{Z_m Z_m} & \sqrt{1+\alpha}\, M_{Z_o Z_m} \\ \sqrt{1+\alpha}\, M_{Z_m Z_o} & M_{Z_o Z_o} \end{pmatrix} \quad ,$$

where Z_m denotes the matrix of variables included in the mask and Z_o denotes the matrix of variables that should not b masked[2].

Subsequently, for any masked observation a vector of Mahalanobis-distances between the masked
and all original observations is calculated (see section 2) The criterion chosen for sufficiency is that the distance between the original record and it's masked counterpart is not one of the two smallest distances. Otherwise the mask will be repeated.

For manipulating the errors two subroutines are proposed: If the error is "slightly" to small, this means if the distance of the true pair is the second smallest and if the value of the distance is larger than a predetermined value the errors will be multiplied by a constant. Otherwise the error will be generated completely new.

[2] An alternative for integrating partial masks is using the GDAP-Method of Muralidhar, Sarsa and Sarathy (1999).

In steps 5 and 6 of the algorithm, the masked values are transformed back to the original scale. For each variable Z_j^*, $j = 1,\mathrm{K}\ , p$ a vector of normalised ranks D_j^* is calculated. For this purpose a vector of ranks R_j^* is calculated with elements of Z_j^* in ascending order and divided by the number of observations n.

This 'empirical distribution' is modified for back transformation, because its values depend solely on the sample size. Hence the errors are standardised:

$$u_{tj}^+ = \frac{u_{tj}^* \sqrt{\frac{1}{n-1}} \sum_{t=1}^{n} u_{tj}^{2*}}{n-1} \quad ,$$

mapped into the domain (0;1) and added to the ranks:

$$D_{tj}^* = \frac{R_{tj}^* - \varphi(u_{tj}^+)}{n} = \frac{R_{tj}^*}{n} - \xi_{tj} \qquad , \qquad t = 1,\mathrm{K}\ , n \quad ,$$

where $\varphi(\bullet)$ denotes the function that maps the u_{tj}^+ to values between zero and one. A good choice for $\varphi(\bullet)$ is the standard normal distribution, because u_{tj}^+ is normally distributed which leads to a uniform distributed ξ_{tj}. If the errors for some variables are set to zero (partial masks) u_{tj}^+ is zero. Hence the normalised ranks of this "masked variables" are identical to the ranks respectively the smoothed empirical distribution of the original variables.

The final back transformation differs depending on whether a variable is continuous or discrete.
For continuous variables the inverse of the smoothed empirical distribution $\overline{F}_X$ is used:

$$X_{ctj}^* = x_{i-1,j} + \frac{D_{tj}^* - \hat{F}(x_{i-1,j})}{\hat{F}(x_{i-1,j})}(x_{i.j} - x_{i-1,j}) \quad , \qquad \text{for} \qquad D_{tj}^* \in (\hat{F}(x_{i-1,j}), \hat{F}(x_{i,j})] \quad .$$

For back-transforming the binary variables the corresponding transformation equation is inverted:

$$X_{dtj}^* = \begin{cases} 0 & \text{if} \quad D_{tj}^* \in (0, 1-p_{oj}) \\ 1 & \text{if} \quad D_{tj}^* \in (1-p_{oj}, 1) \end{cases} \qquad \text{for} \qquad t = 1,\mathrm{K}\ , n \qquad j = 1,\mathrm{K}\ , p_d \quad .$$

This back transformations ensure that the univariate distributions will be preserved approximately. Therefore sample means and variances are similar to those of the original data.

The correlations, however, differ due to numerical differences and/or not jointly normally distributed original variables and if partial masks are performed. Furthermore the correlations between X and X^* differ between the variables, this means that the "relative amount of noise" differs. To adjust for these drawbacks Sullivan (1989, pp. 76) proposes two iterations. First, the cross correlations between the variables and their masked counterparts are adjusted to a robust average. Second, differences between the elements of the correlation matrices are minimised.

Because expectancy values of original and masked variables are identical, adjusting cross correlations is based on:

$$X_{tj}^* = \rho_{x_i x_j^*} + \xi_{tj} \qquad ,$$

where $\rho_{x_i x_j^*}$ denotes the correlation between X_i and X_j^* , and ξ_j an error term independent of X_i, $\xi_j \sim (0, \sigma_{\xi_j}^2)$, $Cov(X_i \xi_j) = 0$ with $\sigma_{\xi_j}^2 = \sigma_{X_j}^2 (1 - \rho_{x_i x_j^*})$.
This leads to:

$$\rho_{x_i x_j^*} = 1 - \frac{\sigma_{\xi_j}^2}{\sigma_{X_i}^2} \quad .$$

It is reasonable to assume that the correlation between the original variables and their masked counterparts is positive, this means $\sigma_{\xi_j}^2 < \sigma_{X_j}^2$. Therefore the correlation increases with decreasing $\sigma_{\xi_j}^2$. Hence a modification of the error terms can be used for adjusting the correlations:
The target correlation chosen is a robust average of the correlations $\bar{\rho}$ estimated by the means of the cross-correlations not using their extreme values.

For determining the amount of variation of variances a simple linear approximation is used. A transformation matrix B_{aa} is defined with typical element:

$$b_{aaii} = \begin{cases} \dfrac{1 - \bar{\rho}}{1 - 0.5(\bar{\rho} + r_{X_i X_{ij}^*})} & \text{if} \quad i = j \\ 0 & \text{else} \end{cases} \quad .$$

and new standard normal masked values are calculated by: $Z_t^* = Z_t + u_t^* B_{aa}$.

These are transformed back to the original scale. This adjustment is done iterative until the observed cross correlations differ from the desired cross correlations less than a specified amount or until a predetermined number of iterations is exceeded.

As mentioned above correlations, of the masked and the original data differ usually. Sullivan therefore proposes (1989, pp. 78) a second iterative adjustment in order to make the off diagonal elements of the correlation matrices nearly identical. The basic idea is again to use a linear transformation for adjusting the error terms. Sullivan (1989, p. 78) proposes to modify them subsequently, starting with the variable for which the sum of the differences between the correlations in the original data and the masked data is maximal.

For modifying the errors, a linear combination H_1^* of the transformed original variable chosen (Z_1) and the masked values of the other variables (Z_j^*) is calculated:

$$H_1^+ = b_0 Z_1 + \sum_{i=1}^{p} b_j Z_j^* = Z^{+\prime} b$$

with $b_0 = 1 - b_1$. The system of equations defining the desired properties is $r(G_1^*, X_1) = \kappa$ and $r(G_1^*, X_1^*) = r_{x_1, x_1^*}$, where G_1^* denotes the back transformed variable corresponding to H_1^* and κ the arithmetic mean of the cross correlations between X_1 and the masked variables X_2^*, K, X_p^*: $\kappa = 1/(p-1) \Sigma_{j=2}^{p} r_{X_1 X_j^*}$.

The coefficients are calculated by solving:

$$Z^{+\prime} b = \rho^+ \quad ,$$

where Σ_{Z^+} denotes the correlation matrix of $Z^+ = (Z_1, Z_2^*, \mathrm{K}\ , Z_p^*)$ and $\rho^{+'} = (\kappa, r_{X_1X_2,\mathrm{K}\ ,r_{X_1X_p}})$. The newly calculated values of H_1^+ are transformed back to the original scale by steps 5 and 6 of the algorithm. This approximation can be repeated iteratively until a convergence criterion for the correlations is achieved or a predetermined number of iterations is exceeded.

The algorithm described above allows masking of binary and continuous variables. For discrete variables with more than two categories a final back transformation has to be applied. Let Z_t^* be a vector defined as

$$Z_{tj}^* = \begin{cases} X_{dt1}^* & \text{for} \quad i = 1 \\ 1 - \Sigma_{i=1}^{i-1} Z_{tj}^* X_{dtj}^* & \text{for} \quad i = 2, \mathrm{K}\ , k-1 \end{cases} .$$

Then the elements of the masked variable X_d^* are defined as $X_{dt}^* = i$, if $Z_{ti}^* = 1$.

So, the algorithm proposed by Sullivan (1989), (Fuller 1993) combines transformations with adding noise. The transformations chosen assure that the univariate distributions are preserved approximately. They are not linear and quite complex in comparison to the basic idea and the ones used by Kim (1986, 1990)[3]. Additionally iterative procedures are used for correcting differences in correlations induced by transformations and mask. Due to these corrections it is not guaranteed that all variables have the same level of protection.

3.3 Properties of the algorithm: Theoretical considerations for analytical usefulness

In order to investigate into the properties of the algorithm it is useful to analyse whether the masked variables allow an unbiased estimate of the first two moments of the unmasked variables as long as only binary discrete variables are masked. In (Brand 2000) it is shown that the sample means of the masked data are unbiased estimates for the expected values of the original data. Furthermore, sample variances of masked binary variables are unbiased estimates for the variances of the original binary variables. This is a very interesting property of the algorithm, because most other algorithms proposed for masking through noise addition can only be applied to the continuous variables.

Nevertheless the variances of continuous variables increase (Brand 2000). The increase is higher with increasing sample size and larger differences between the ordered observed values in the original data. This result shows that the sample variance calculated by the masked variables is a biased estimate for the sample variance of the original data. The variation in the original data is overestimated due to the additional variation within the 'classes' - this means the differences between ranked observed values - used for transformation.

Due to the complex structure of the algorithm a more detailed analysis of the properties is not useful. The main reason for this is that the distribution of the errors on the original scale can not be obtained explicitly. Therefore, only general inferences for regression estimates are possible on the basis of an errors-in-variables-model (Fuller 1993). Additionally, analysis of non random subsamples leads usually to misleading results, because correlations of subsamples are not preserved.

Furthermore linear and non linear dependencies on the level of records can not be fully preserved. Hence relevant variables for analysis that base on combinations of values (e.g. binary variables, indices) have to be derived before the algorithm is applied. This is due to the fact that the algorithm preserves only the univariate distributions and the correlations. All other properties of the data may be distorted by the algorithm (see Brand 2000). Hence regression analysis based on any variables constructed by transformations of the masked variables may be misleading.

[3] For a comparison see e.g. Brand (2002).

Summarising, the algorithm proposed by Sullivan (Sullivan 1989, Sullivan/Fuller 1989, Fuller 1993) can be characterised as a complex method that combines adding noise with non-linear transformations. Univariate distributions of the variables are maintained approximately. Variances of continuous variables, however, increase in small samples due to the structure of the transformations. The algorithm ensures that the correlation structure of the original variables is preserved by iterative adjustment procedures.

4. Empirical Results

4.1 Implementation of the algorithm

The algorithm described above has been implemented as a GAUSS-Routine. The GAUSS Mathematical and Statistical System is a flexible, fast matrix programming language designed for computationally intensive tasks. GAUSS routines may be integrated in other programs by an interface. The current implementation allows partial masks for continuous variables as long as no binary variables are planned to be masked. Partial masks for binary variables are possible as long as all continuous variables will be masked.

Tests were conducted on the basis of three different data sets. Firstly, some tests were performed on simulated correlated normal distributed data and no binary variables added. Secondly, tests were undertaken with the simulated data and binary variables added. The binary variables are clearly correlated with the continuous variables. A third test data set was used that has properties similar to real business microdata surveys. Business data are usually characterised by extremely skewed distributions, not very high correlated with the exception of the variables that depend strongly on the size of the unit.

Table 1: Parameters: Specification of interfaces for the current version:

Parameter	**Description**
Multiplier for variances (α)	Relative amount of noise; real value > 0
Constant for sufficiency criterion additionally used to difference criterion	Absolute value of minimum difference, real value > 0
Constant for multiplying noise, if completely new generation is not required	Constant for multiplying errors slightly too small, real value > 0
Tolerance criterion for adjusting cross correlations	maximum value for differences in cross correlations between the variables and their masked counterparts, real value between zero and one
Tolerance criterion for adjusting correlations of original and masked data	maximum value for differences between the correlations of original and masked variables", real value between zero and one

To apply the algorithm, at least five parameters have to be set (table 1). The multiplier for the variances of the transformed variables, the constant for multiplying the errors and the distance which determines whether errors will be generated completely new are necessary for masking itself. They determine the details of masking. Hence, they influence the adjustment routines and the number of not sufficiently masked observations. The tolerance for the adjusting cross-correlations between the original variables and their masked counterpart determine the maximum differences in the amount of noise on the level of the original variables. This leads to a "similar level of protection" for all variables. The last parameter, the tolerance level for adjusting correlations of original and masked data, determines the maximum difference between the correlation-matrices of original and masked data.

Some examples for application of the algorithm proposed by Sullivan can be found in the literature (Sullivan 1989, Sullivan/Fuller 1990, Fuller 1993, Brand 2000, 2002, Brand/Bender/Kohaut 1999) that illustrate the properties. These examples confirm that sample means and variances are approximately preserved for different distributions of the original variables. Differences in the correlations occur if the number of iterations in the procedures adjusting the correlations is limited to a few or if the distributions of the original variables are extremely skewed. It should be stressed that estimates in subpopulations are biased due to the structure of transformations and an explicit adjustment formula is not known (Sullivan 1989, Sullivan/Fuller 1990). Furthermore, for Sullivan's

algorithm it can be shown that regression analysis with limited dependent variables can be misleading, because the multivariate distribution is not preserved (Brand 2000). The same is true for all methods discussed above if likelihood estimations require an explicit determination of the error distribution.

4.2 Results for test data

In the previous sections the masking algorithm suggested by Sullivan has been described and its analytical properties were discussed. In this section, numerical examples are presented and the results are compared with the ones found in literature. For this purpose, a normally distributed test data set and test data that base on real data were used. These test data is one of the datasets chosen for comparing the different perturbation methods tested in the CASC-project, the so called "Tarragona" data[4]. These data are stemming from 834 companies of the Tarragona area in Spain.

4.2.1 Results for Simulated Data

Firstly, normally distributed test data with 500 records and six columns were generated for illustrating the technical applicability of the algorithm. The descriptive statistics of these data can be found in table A1 and table A2 (means, standard-deviations, correlations). The algorithm was applied to the whole data set, as well as to subsamples of 250 records with 10 replications, each with a relative amount of noise used for masking the standard normal data (α) of 0.5, a minimum distance for using the simplified modification of errors of 0.5, and a multiplier for the errors of 0.5.

The results can be found in table A3 – A6. Table A3 shows that the algorithm works properly, if it is applied to these data in the sense that means and standard-deviations are similar and correlations are preserved with differences lower than the predetermined amount (0.01). According to the internal distance criterion about 19% of 250 records respectively 16% of 500 records have not been sufficiently masked.
Table A4 gives the results if variable six is not masked, this means if partial masks are applied to variables one to six. The results show that masks can be performed successfully with the same parameter values than in the first experiment. The number of not sufficiently masked records increases to 25% if 250 records are used and to 20% for the whole test data. Other experiments – not reported in the tables – indicate that the masking algorithm fails if more variables are excluded from the mask, because the correlation adjustment procedures fail.

As explained above, successfully masking of binary variables requires that they are clearly correlated with the continuous variables. Hence, the binary variable generated for testing this part of the algorithm is correlated with the continuous variables (table A1). The results for 250 and 500 records show that the algorithm works successfully in most replications (Table A5). When including binary variables, it seems that the algorithm works more stable with relatively large maximum differences allowed after adjusting cross correlations and correlations. This is due to the indirect inclusion of the variables in the algorithm.

If an additional continuous variable in included (table A6) that has only small variation and is highly correlated with the binary variable, adjustment of correlations fails most often if a maximum tolerances levels of 0.01 is chosen. If tolerance levels are set to higher values (0.05 for both adjustment procedures) the algorithm works properly. According to the internal distance criterion the number of not sufficiently masked records increases to 25%.

The results shown above stand in line with the findings in Sullivan (1989) and Brand (2000) indicating that the algorithm works properly for synthetic data as long as the binary variables are correlated with the continuous ones.

Tests with synthetic data that are characterised by extremely skewed distributions and high proportions of the records with zero values are shown in Brand (2002). Here, it is shown that the algorithm is applicable to these data. The differences in the univariate distributions especially in the standard-deviations are higher than for the standard normal data. Furthermore, minima, maxima and the proportion of values close to zero are not too far from the values in the original data due to the transformations chosen. Adjustment of cross correlations

[4] For a detailed description see Domingo-Ferrer/Mateo-Sanz 1998.

terminated at its limit, and the iterations used for equalising the correlation matrices were terminated at their limit for one of the variables after being successfully used for four other variables. Although the distributions of the original variables are extremely skewed and therefore the transformed standard normal variables have a correlation structure that is different from the original data the adjustment procedures worked properly.

For investigating systematically into the properties of the algorithm if real business data are used tests with the so called “Tarragona”-data were undertaken. These data are one of the data sets chosen for comparing the different SDC-Methods in the CASC-project. The variables included in this data set are shown in table A7. They are typical for business surveys in the sense that the distributions are skewed, the cover a wide range and the distributions are highly skewed.

The algorithm failed, when it wass applied to the whole data set. Thus, the data had to be manipulated before the algorithm was applied. One possible choice of modifying the data would be to apply those transformations usual in the process of data analysis (tables A9 – A11). Another option would be to split the data into more homogeneous subsamples, e.g. by size (see e.g. Brand 2000, Brand/Bender/Kohaut 1999).

In our experiment, we first choose transformations based on some simple considerations. With the prepared data set (table A12) the algorithm proposed by Sullivan gave the following results: In ten replications of the experiment the algorithm failed four times, if the whole data set was to be masked, and three times, if the transformed variable six was not taken into account. At least in one of the cases the algorithm failed, when the maximum difference between the correlations of the original and the masked data exceeded 0.45. These results are not satisfactorily. An inspection of the detailed results shows that for variables X1 and X3 adjustment of the correlation failed for Variable X1 often. The reason for this is that the underlying assumptions concerning the correlation of variables are not fulfilled. For both variables, the calculated linear combinations do not lead to a sufficient increase in similarity between the correlation matrices of the original and the masked data.
Unfortunately, we are unable to give a clear mathematical reason for this.

In a second experiment, we chose another transformation: First, the natural logarithm of X1 was chosen in order to increase the variation of this variable. Variable X3 was modified in the way that the ratio of operating profit to net profit was substituted by the ratio to sales (table A13). The reason for this was that operating profit and net profit are economically less dependent than operating profit and sales. The descriptive statistics of this data set can be found in tables A14 – A15.

For these data, masking experiments were conducted for a full mask and a partial mask in which variable X6 was excluded. The results are presented by table A16. With respect to preservation of means, standard-deviations and correlations the algorithm worked properly although the maximum of the differences for the standard deviations wass quite high. Adjustment of cross-correlation worked properly while adjusting correlations between the original variables and their masked counterpart failed very often. For this data set about 11% of the records were not sufficiently masked on the average. This proportion increased to 16% if partial masking was used in which variable X6 was excluded. Compared to the former experiments described above, the number of not sufficiently masked records was significantly lower for the “Tarragona-data”. This result indicates that for real data the level of protection differs substantially from the one of the synthetic data.

The proportion of not sufficiently masked records is somewhat higher than the proportion of insufficiently masked records that is usually used as threshold for sufficient protection of a whole data set. Nevertheless, it has to be kept in mind that the results presented are based on a comparison of original and masked records. For investigating into the number of records that are insufficiently masked in a real data set, experiments with additional information stemming from external data sets would be necessary. A final evaluation of the level of protection is beyond the scope of this report.

4.2.2 Further Aspects of Practical Applicability in Statistical Offices

For applying the method to real data some further aspects have to be taken into account that have an impact on the practical applicability of an algorithm in a statistical agency with a reasonable amount of time and manpower. The first aspect is: does it require expert knowledge on the data and on statistical disclosure control methods

used to apply the algorithm? The second aspect relates to the computing time: is it possible to apply the algorithm to huge real life data sets in reasonable time?

For practical applicability reasons all tests were performed with a fixed data set. Only the tolerance levels for the differences in correlations were varied. The results show that variations of these parameters influence the results of the algorithm. The applications lead to sufficient results more often if the tolerances are set higher. Additional test results not shown here indicate that extreme parameter values for the mask especially for the relative amount of noise can lead to abortion of the algorithm for normal distributed test data too.

The results for the "Tarragona-data" show that small modifications in preparing the data can lead to substantially different results for masking. This modifications have to be done manually because no clear "ratios" or benchmarks can be determined theoretically. Hence detailed knowledge of the data and the algorithm is necessary for an successful application. Furthermore the meaning of the parameters is not always clear; the parameter values chosen determine the results indirectly, because applying the adjustment routines will lead to modifications in the errors and hence the level of protection chosen for the masking procedure. Hence applying the routine requires an experienced user who has applied the algorithm several times before generating the "final" mask.

The second aspect is the computation time needed. For the modified Tarragona-Data computation computing time usage is about 14 minutes for one replication on an Personal Computer (Pentium II, ca. 120 MB RAM). If the adjustment routines do not lead to an useful result in a small number of iterations the computation time increases for the "Tarragona-data" by a few minutes. If more records are used, computation time increases rapidly, because the number of calculations for determining the distances increases by factor n (n = number of observations) if one record is added to the data. Hence datasets with more than 1500 records cannot be masked in practice. Larger data sets must be split into groups before the masking algorithm is applied.

To come to a conclusion, the algorithm proposed by Sullivan can be applied in practice to real and synthetic data. Partial masks can be performed too in a way that correlations between masked and unmasked variables are preserved. Nevertheless, it does not work stable if the underlying structure of the data incorporates strong dependencies between some variables or if the distributions are extremely skewed and cover a wide range. Hence the data have to be manipulated manually by statistical experts in advance. The algorithm is quite time-consuming due to the internal distance criterion. This leads to an upper limit of ca. 1500 for the number of records that can be masked in one step. Therefore, larger datasets have to be split up in advance of masking them.

5. Summary

In this paper, we have been looking into masking through noise addition. A detailed description of the algorithm proposed by Sullivan is given, an extension in order to integrate partial masks is proposed and its applicability to real data sets is examined. Criteria relevant for examining the applicability are analytical usefulness, level of protection and practical applicability. The latter aspects do not only include technical stability and computing time requirement. It has to be born in mind too, that complex data manipulations are required, which cannot be standardised. Also, it is not easy to understand the parameters to be chosen, thus setting parameters leads to time-consuming trial and error processes, and require experienced users. Thus, it is expected that an application of Sullivan's algorithm is probably more 'costly' than using another (maybe less optimal) method, easier to be applied.

Sullivan's algorithm is characterised by a mixture of non-linear transformations and noise addition. The algorithm is of special interest, because it preserves univariate distributions and correlations of the data. Furthermore, it is the only algorithm that gives the opportunity to mask continuous and discrete variables on the basis of noise addition in one step. Nevertheless, variables necessary for analysis generally must be calculated in advance of masking the data (so they have to be known in advance of masking). Non-random subsamples of masked data are analytically valid only if the data was split into these groups before the algorithm was applied.

The results of the empirical tests show that the algorithm can be applied in practice to real and synthetic data. Partial masks can be performed as well, in a way that correlations between masked and unmasked variables are maintained. Nevertheless the algorithm does not work stable when there are strong dependencies between some

variables or if the distributions are extremely skewed and cover a wide range. Hence, the data have to be prepared manually by statistical experts. Furthermore the parameters have no clear effect on the results, especially the parameters for adjusting the cross correlations. The algorithm is quite time-consuming due to the internal distance criterion. This leads to an upper limit of ca. 1500 for the number of records that can be masked in one step. Larger datasets have to be split up in advance of masking them.

All in all, the results indicate that an algorithm as complex as the one proposed by Sullivan can only be applied by experts. Every application is very time-consuming and requires expert knowledge on the data and the algorithm. Especially for data sets that require complex data manipulations an application at statistical agencies is expected to be rather too expensive. Other routines, like microaggregation (see e.g. Domingo-Ferrer/Mateo-Sanz 2001) seem to be more promising in that case. Still, it is a very valuable framework that can be used as a reference for further research in the field of statistical disclosure methods.

Annex: Tables

Table A1: Descriptive statistics of normal distributed test data (500 records)

Variable	Mean	Standard deviation
X1	20.83	11.84
X2	11.24	16.83
X3	13.34	41.31
X4	30.20	21.87
X5	49.91	3.243
X6	10.77	15.75
X7	45.04	2.301
B1	0.5180	0.5002

Table A2: Correlations of normal distributed test data (500 records)

	X1	X2	X3	X4	X5	X6	X7	B1
X1	1							
X2	0.7047	1						
X3	0.02368	0.02263	1					
X4	-0.09703	-0.08173	0.1909	1				
X5	-0.03850	-0.07756	0.1003	0.6936	1			
X6	-0.02899	0.04875	0.08803	0.2961	0.05274	1		
X7	-0.07573	-0.06918	0.1791	0.7964	0.6980	0.2523	1	
B1	-0.03693	-0.03667	0.06881	0.6500	0.5653	0.2019	0.8113	1

Table A3: Normal distributed test data: Results for Sullivan' Algorithm

Variables: X1 – X6

Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5, Constant for multiplying noise, if completely new generation is not required: 0.5,

Tolerance criterion for adjusting cross correlations: 0.01, Tolerance criterion for adjusting correlations between original and masked data: 0.01

Number of replications: 10

Differences in means and standard deviations	250 records	500 records
Maximum difference in absolute means	0.06	0.02
Mean of differences in absolute means	0.003	0.001
Maximum difference in standard deviations	0.70	0.59
Mean of differences in standard deviations	0.21	0.14
Maximum difference in correlations	< 0.01	< 0.01

Number of not sufficiently masked records	250 records			500 records		
	Mean	Maximum	Minimum	Mean	Maximum	Minimum
1. mask	77,0	93	65	119,2	135	107
2. mask	48,1	54	42	81,2	92	71
3. mask and adjustment of correlations	48,1 (19%)	54	42	81,2 (16%)	92	71

Table A4: Normal distributed test data: Results for Sullivan' Algorithm

Variables X1 – X6, partial mask X6 excluded
Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5,
Constant for multiplying noise, if completely new generation is not required: 0.5,
Tolerance criterion for adjusting cross correlations: 0.01, Tolerance criterion for adjusting correlations between original and masked data: 0.01
Number of replications: 10

Differences in means and standard deviations	250 records	500 records
Maximum difference in means	0.02	0.02
Mean of differences in means	0.004	0.001
Maximum difference in standard deviations	0.68	0.52
Mean of differences in standard deviations	0.14	0.09
Maximum difference in correlations	< 0.01	< 0.01

Number of not sufficiently masked records	250 records			500 records		
	mean	Maximum	Minimum	mean	Maximum	Minimum
1. mask	116,0	127	105	154,3	167	139
2. mask	63,3	70	57	99,8	108	86
3. mask and adjustment of correlations	63,3 (25%)	70	57	99,8 (20%)	108	86

Table A5: Normal distributed test data: Results for Sullivan' Algorithm

Variables X1 – X6, B1 (one binary variable included)
Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5,
Constant for multiplying noise, if completely new generation is not required: 0.5,
Tolerance criterion for adjusting cross correlations: 0.01, Tolerance criterion for adjusting correlations between original and masked data: 0.01
Number of replications: 10

Differences in means and standard deviations	250 records	500 records
Maximum difference in means	0,05	0,04
Mean of differences in means	0,001	0,002
Maximum difference in standard deviations	0,46	0,43
Mean of differences in standard deviations	0,13	0,08
Maximum difference in correlations	< 0,01	0,03
Number of replications in which adjustment of cross correlations is terminated at 100 iterations	2	1
Number of replications in which adjustment of correlations is terminated at 1000 iterations	0	1

Number of not sufficiently masked records	250 records			500 records		
	mean	Maximum	Minimum	Mean	Maximum	Minimum
1. mask	100,0	108	93	154,3	167	139
2. mask	57,5	63	63	99,8	108	86
3. mask and adjustment of correlations	57,5 (23%)	48	48	99,8 (20%)	108	86

Table A6: Normal distributed test data: Results for Sullivan' Algorithm

Variables X1 – X7, B1 (one binary variable included), 250 records
Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5,
Constant for multiplying noise, if completely new generation is not required: 0.5
Number of replications: 10

	Adjusting Cross Correlations and	Adjusting Cross Correlations with a	Adjusting Cross Correlations and

	Correlations with a tolerance of 0.01	tolerance of 0.05 and Correlations with a tolerance of 0.03	Correlations with a tolerance of 0.05
Maximum difference in means	0.04	0.05	0.05
Mean of differences in means	0.002	0.001	0.0008
Maximum difference in standard deviations	0.35	0.43	0.42
Mean of differences in standard deviations	0.09	0.09	0.09
Maximum difference in correlations	0.06	0.43	< 0.05
Number of replications in which adjustment of cross correlations is terminated at 100 iterations	7	0	0
Number of replications in which adjustment of correlations is terminated at 1000 iterations	5	3	0

Number of not sufficiently masked records	Adjusting Cross Correlations and Correlations with a tolerance of 0.01			Adjusting Cross Correlations with a tolerance of 0.05 and Correlations with a tolerance of 0.03			Adjusting Cross Correlations and Correlations with a tolerance of 0.05		
	Mean	Maximum	Minimum	Mean	Maximum	Minimum	Mean	Maximum	Minimum
1. mask	121,1	132	111	117,7	128	106	117,5	124	114
2. mask	61,5	70	54	61,4	63	57	62,1	67	54
3. mask and adjustment of correlations	61,5 (25%)	70	54	61,4 (25%)	63	57	62,1 (25%)	67	54

Table A7: Tarragona Data: Descriptive Statistics of selected variables (Number of records: 834)

	Mean	Standard deviation
Sales	546958.28	1155792.7
Labour Costs	74447.959	135407.60
Depreciation	10855.103	27193.581
Operating Profit	27622.765	88221.548
Financial Outcome	-8366.2482	25690.361
Gross Profit	21243.484	80287.944
Net Profit	14133.801	56155.537

Table A8: Tarragona Data: Correlations of selected variables

	Sales	Labour Costs	Depreciation	Operating Profit	Financial Outcome	Gross Profit	Net Profit
Sales	1						
Labour Costs	0.6494	1					
Depreciation	0.6227	0.5819	1				
Operating Profit	0.7435	0.5620	0.7361	1			
Financial Outcome	-0.5197	-0.3803	-0.4883	-0.4406	1		
Gross Profit	0.6767	0.5156	0.6813	0.9441	-0.2049	1	
Net Profit	0.6504	0.4904	0.6621	0.9332	-0.2170	0.9799	1

Table A9: Tarragona-Data: Description of variables

X1	Labour costs divided by sales
X2	Depreciation divided by net profit
X3	Operating profit divided by net profit
X4	Financial outcome divided by sales
X5	Gross profit divided by operating profit
X6	Net profit divided by gross profit

Table A10: Tarragona Data: Descriptive Statistics of transformed data (Number of records: 826*)

	Mean	Standard deviation
X1	0.2079	0.2411
X2	4.099	61.16
X3	5.935	58.28
X4	-0.02632	0.1336
X5	0.8451	5.438
X6	0.7635	0.6952

* The number of records decreases in comparison to the original data set due to not defined divisions.

Table A11: Tarragona Data: Correlations of transformed data

	X1	X2	X3	X4	X5	X6
X1	1					
X2	0.001764	1				
X3	0.04088	0.5557	1			
X4	-0.08780	-0.007419	0.01576	1		
X5	0.006259	-0.008023	-0.01094	-0.01891	1	
X6	0.01087	-0.01085	-0.07738	-0.05589	-0.02345	1

Table A12: Tarragona data: Results for Sullivan' Algorithm

Variables X1 – X6

Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5, Constant for multiplying noise, if completely new generation is not required: 0.5

Number of replications: 10

	Full mask, Adjusting Cross Correlations with a tolerance of 0.05 and Correlations with a tolerance of 0.03	Partial mask, Adjusting Cross Correlations with a tolerance of 0.05 and Correlations with a tolerance of 0.03
Maximum difference in means	1,08	1,30
Mean of differences in means	0,16	0,18
Maximum difference in standard deviations	23,16	23,37
Mean of differences in standard deviations	3,38	3,72
Maximum difference in correlations	0,47	0,47
Number of replications in which adjustment of cross correlations is terminated at 100 iterations	4	4
Number of replications in which adjustment of correlations is terminated at 1000 iterations	4	3

Number of not sufficiently masked records	Full mask			Partial mask		
	Mean	Maximum	Minimum	Mean	Maximum	Minimum
1. mask	133,6	157	113	199,8	220	175
2. mask	94,2	106	106	140,4	156	127
3. mask and adjustment of correlations	94,2 (11%)	84	84	140,4 (17%)	156	127

Table A13: Tarragona-Data: Description of transformed variables

X1*	Natural logarithm of Labour costs divided by sales
X2	Depreciation divided by net profit
X3*	Operating profit divided by sales
X4	Financial outcome divided by sales
X5	Gross profit divided by operating profit
X6	Net profit divided by gross profit

Table A14: Tarragona Data: Descriptive Statistics of transformed data (Number of records 818*)

	Mean	Standard deviation
X1*	-1.953	0.9429
X2	4.172	61.52
X3*	0.02720	0.2825
X4	-0.02575	0.1335
X5	0.8226	5.423
X6	0.7636	0.6988

* The number of records decreases in comparison to table 9 due to taking logarithms of variable X1 (records with an absolute value over 1,000,000 for variable X1* were excluded).

Table A15: Tarragona Data: Correlations of transformed data

	X1*	X2	X3*	X4	X5	X6

X1*	1					
X2	0.002207	1				
X3*	-0.1626	0.006587	1			
X4	-0.09465	-0.007616	0.06827	1		
X5	0.01945	-0.007912	-0.006473	-0.01633	1	
X6	0.008331	-0.01065	-0.01132	-0.05488	-0.02271	1

Table A16: Tarragona data: Results for Sullivan' Algorithm

Variables X1* – X6

Multiplier for variances: 0.5, Constant for sufficiency criterion additionally used to difference criterion: 0.5, Constant for multiplying noise, if completely new generation is not required: 0.5

Number of replications: 10

	Full mask, Adjusting Cross Correlations with a tolerance of 0.05 and Correlations with a tolerance of 0.03	Partial mask, Adjusting Cross Correlations with a tolerance of 0.05 and Correlations with a tolerance of 0.03
Maximum difference in means	1,25	1,17
Mean of differences in means	0,15	0,12
Maximum difference in standard deviations	23,94	24,37
Mean of differences in standard deviations	3,48	2,77
Maximum difference in correlations	0,03	0,03
Number of replications in which adjustment of cross correlations is terminated at 100 iterations	7	8
Number of replications in which adjustment of correlations is terminated at 1000 iterations	0	0

Number of not sufficiently masked records	Full mask			Partial mask		
	Mean	Maximum	Minimum	Mean	Maximum	Minimum
1. mask	129,9	141	112	193,0	208	168
2. mask	88,9	101	75	133,5	142	117
3. mask and adjustment of correlations	88,9 (11%)	101	75	133,5 (16%)	142	117

Literature

Brand, R. (2000): Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung BeitrAB 237, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Brand, R. (2002): Masking through Noise Addition, in: Domingo-Ferrer, J. (ed.), Inference Control in Statistical Databases – From Theory to Practice, Lecture Notes in Computer Science, Heidelberg, New York, Springer.

Brand, R., Bender, S. and S. Kohaut (1999): Possibilities for the Creation of a Scientific Use File for the IAB-Establishment-Panel, Statistical Data Confidentiality, Proceedings of the Joint Eurostat/UN-ECE Work session on Statistical Data Confidentiality in March 1999.

Domingo-Ferrer, J. and J.M. Mateo-Sanz (2001): Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Skopje (The former Yugoslav Republic of Macedonia), 14-16 March 2001.

Domingo-Ferrer, J. and J.M. Mateo-Sanz (1998): A Comparative Study of Microaggregation Methods, Qüestiio, 22/3, pp. 105 – 112.

Domingo-Ferrer, J. and V. Torra (2001): A Quantitative Comparison of Disclosure Control Methods for Microdata. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M. and L.V. Zayatz (ed.): Confidentiality, Disclosure and Data Access – Theory and Practical Applications for Statistical Agencies, North-Holland, pp.111 – 134.

Domingo-Ferrer, J. and V. Torra (2002): Validating distance-based record linkage with probabilistic-based one, Lecture Notes in Artificial Intelligence, Heidelberg, New York, Springer, to appear.

Fuller, W.A. (1993): Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics 9, pp. 383 – 406.

Kim, J.J. (1986): A Method for Limiting Disclosure in Microdata based on Random Noise and Transformation, Proceedings of the Section on Survey Research Methods 1986, American Statistical Association, pp. 303 – 308.

Kim, J.J. (1990): Subpopulation Estimation for the Masked Data, Proceedings of the Section on Survey Research Methods 1990, American Statistical Association, pp. 456 – 461.

Kim, J.J. and W.E. Winkler (1995): Masking Microdata Files, Proceedings of the Section on Survey Research Methods 1995, American Statistical Association, pp. 114 – 119.

Kim, J.J. and W.E. Winkler (1997): Masking Microdata Files, Statistical Research Division RR97/03, US Bureau of the Census, Washington, DC.

Kim, J.J. and W.E. Winkler (2001): Multiplicative Noise for Masking Continuous Data, unpublished manuscript.

McGuckin, R. H. and S.V. Nguyen (1990), Public Use Microdata: Disclosure and Usefulness, Journal of Economic and Social Development 16, pp. 19 – 39.

McGuckin, R. H. (1993): Analytic Use of Economic Microdata: A Model for Researcher Access with Confidentiality Protection, Proceedings of the International Seminar on Statistical Confidentiality, 08. -- 10. Sept. 1992, Dublin, Ireland; Eurostat, pp. 83 – 97.

Muralidhar, K., Parsa, R. and R. Sarathy (1999): A General Additive Data Perturbation Method for Database Security: Management Science, 45, pp. 1399 – 1415.

Spruill, N.L (1983) The Confidentiality and Analytic Usefulness of Masked Business Microdata, Proceedings of the Section on Survey Research Methods 1983, American Statistical Association, pp. 602 -- 610.

Sullivan, G.R. (1989): The Use of Added Error to Avoid Disclosure in Microdata Releases, unpublished PhD-Thesis, Iowa State University.

Sullivan, G.R. and W.A. Fuller (1989): The Use of Measurement Error to Avoid Disclosure, Proceedings of the Section on Survey Research Methods 1989, American Statistical Association, pp. 802 -- 807.

Sullivan, G.R. and W.A. Fuller (1990): Construction of Masking Error for Categorical Variables, Proceedings of the Section on Survey Research Methods 1990, American Statistical Association, pp. 453 – 439.

Tendick, P. (1988): Bias Avoidance and Measures of Confidentiality for the Noise Addition Method of Database Disclosure Control, unpublished PhD-Thesis, University of California, Davis.

Tendick, P. (1991): Optimal Noise Addition for Preserving Confidentiality in Multivariate Data, Journal of Statistical Planning and Inference 27, pp. 341 -- 353.

Winkler, W.E. (1998): Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Statistical Data Protection 1998, Lisbon, Portugal.