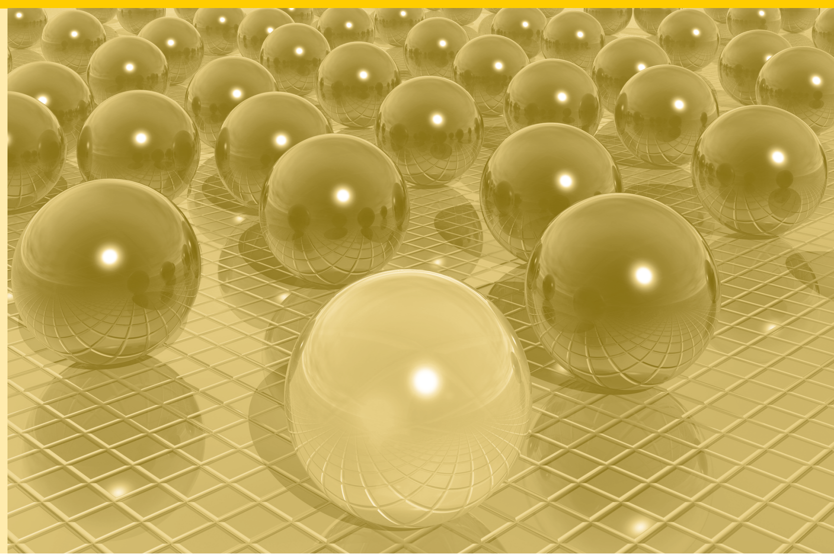
FDZ-Arbeitspapier
Nr. 37



Masking Micro Data with Stochastic Noise

Jörg Höhne, Julia Höninger

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 442006
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 9449-2873
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 9449-2876
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Mai 2011

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

FDZ-Arbeitspapier
Nr. 37

Masking Micro Data with Stochastic Noise

Jörg Höhne, Julia Höninger

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Masking Micro Data with Stochastic Noise

Jörg Höhne and Julia Höninger[1]

Abstract. Stochastic noise is a comparatively new method to anonymise micro data. It is classified as a data perturbation method, as compared to classical anonymisation methods. A new scheme has been developed combining mixture distributions of random noise and the application of the masking method on transformed variables. Adding noise to the logarithmized variables allows preserving correlations in the anonymous data. By combining these two ideas all variables can be masked with a high yet the same relative degree.

The suggested variant of stochastic noise is applied to a data set of all manufacturing firms in Germany. Assessing the quality of the anonymous data set demonstrated the good performance of the anonymisation routine with regards to utility and security even though the parameters are chosen to reach absolute anonymity.

Zusammenfassung. Stochastische Überlagerung ist ein relativ neues Anonymisierungsverfahren für Mikrodaten und gehört zu den datenverändernden Anonymisierungsverfahren. Eine neue Verfahrens-variante wurde entwickelt, welche Mischungsverteilungen und die Idee, das Anonymisierungs-verfahren auf transformierte Variablen anzuwenden, verbindet. Einen stochastischen Fehler auf die logarithmierten Variablen zu addieren ermöglicht es, die Korrelationen in den anonymen Daten zu erhalten. Die zwei Ideen werden in dem hier vorgeschlagenen Verfahren kombiniert und so können alle Variablen mit einem konstanten relativen Fehler anonymisiert werden.

Das neu entwickelte Verfahren wird in einer Beispielstudie auf Daten des Verarbeitenden Gewerbes in Deutschland angewendet. Obwohl die Anonymisierungsparameter so gewählt wurden, dass der Datensatz absolut anonymisiert wurde, zeigt das Anonymisierungsverfahren sehr gute Eigenschaften in Bezug auf Datensicherheit und Analysefähigkeit.

Keywords: stochastic noise, anonymisation, micro data, data perturbation.

JEL: C10, C81, C40, C52

[1] Amt für Statistik Berlin-Brandenburg, State Statistical Institute, Alt-Friedrichsfelde 60, 10315 Berlin, Germany. Email: Joerg.Hoehne@statistik-bbb.de and Julia.Hoeninger@statistik-bbb.de. This research is part of the research project “infinitE – an Informational Infrastructure for the E-Science Age”, funding by the German Federal Ministry of Education and Research is greatly appreciated.

1. Introduction

Disseminating micro data files is a comparatively new task for data providers. To comply with the required levels of statistical confidentiality anonymisation methods have to be applied before releasing any micro data files. Generally only eligible researchers can get access to scientific use files (SUF) where data are only slightly modified. The general public has access to public use files (PUF) that have undergone more drastic procedures of anonymisation.

Even when data producers permit access to micro data only on-site, there is a need for anonymisation methods. In most countries access is provided through so called Research Data Centres (RDC). The most common access options are safe work stations inside the statistical offices and remote data processing. Dummy data sets are needed on which researchers can test their programme syntax before submitting it to remote data processing. The government funded project “infinitE” (Brandt and Zwick 2009) wants to produce better dummy data sets, so called data structure files (DSF). Correct reproduction of syntactic and semantic structures would reduce the burden on both sides as fewer programs with fewer faults have to be executed to address the research question. On the other hand DSFs should have a very high level of anonymity.

This paper will describe an anonymisation method that can be applied to produce such dummy data sets. The production of SUFs is similar but the aims are different. SUFs are supposed to be suitable to produce robust results for publication. SUFs are only slightly masked. The method suggested in this paper is very different: data will be masked heavily and results are only to resemble the original results. Researchers will only test programmes on the dummy data sets produced but then run the programmes in remote execution on the original data afterwards.

The criteria that the newly produced dummy data sets have to fulfil have been compiled by the “infinitE” project group. Anonymisation methods will be compared according to data utility and security in this project. Any masking method which broadens categories is ruled out as syntax will not identically fit on the original data. Traditional methods are therefore not tested. The project group will compare synthetic data generated with multiple imputation and data manipulation methods such as stochastic noise.

The suggested variant of stochastic noise as well as multiple imputation can conserve all variables and categories as well as the approximate ranges of metric variables, means, variances and univariate distributions of the variables. To conserve covariances and correlations is important for regression results. The stochastic noise model is especially successful in conserving the quotients of two variables due to the mixture distribution employed. This is important in the panel setting, as not only shares in the cross section have to be preserved but also the time series properties, e.g. growth patterns. Furthermore, structural zeros and signs are conserved which is desirable to preserve logical conditions (e.g. entering and exiting firms in the panel, non-negative number of employees).

On the other hand, anonymisation has to secure that the data file is anonymous and individuals and their attributes cannot be disclosed. The desired degree of anonymity for the aspired dummy data sets is far beyond the usual levels. Data that is classified as absolutely anonymous can be provided for download on the web. Researchers can devise programmes even before a contract has been signed. So the parameters are chosen so high that every single value deviates on average 25% from its original value. As stochastic noise works only for continuous data another anonymisation method probably has to be applied to the categorical variables in order to produce an absolutely anonymous file of any real data set.

2. Description of the Methodology - General

The general idea of multiplicative stochastic noise is to multiply all data points of the original data X^O with a stochastic factor:

$$X^A = U \otimes X^O \tag{1}$$

where U denotes a matrix of random values and X^A the anonymous data set. To conserve the expected value $E(X^O) = E(X^A)$, the noise is chosen to have $E(U)=1$. Unfortunately, in a multiplicative setting it is not possible to conserve the variance-covariance matrix of the original data (Kim and Winkler 2003). New ideas for multiplicative noise conserve variance and covariance on a different quality level that leads to higher disturbed correlations (i.e. see Oganian and Karr 2010, p. 7)

Firm data are usually highly skewed and analyses are often conducted using log-values. Therefore consider perturbing the logarithmized values additively

$$X^P = \log(X^O) \tag{2}$$

$$X^R = X^P + U \tag{3}$$

which when taking the anonymised values X^R to the exponent e becomes

$$\begin{aligned} X^A &= \exp(X^R) \\ &= \exp(\log(X^O)+U) \\ &= X^O\exp(U) \end{aligned} \tag{4}$$

where:

$X^{O,\,P,\,R,\,A}$ – Data matrix of single values with ^{O} in original, ^{P} after preparation step, ^{R} after randomizing step, ^{A} anonymous

E(U)=0 – all values of random matrix U have expectation value 0

log(X), exp(X) – log and exponential function are applied element wise.

The noise U has to have E(U)=0 and it is possible to preserve the variance-covariance-matrix of the original logarithmized data $V(X^P)$ in the logarithmized anonymised data up to a certain constant d (Kim and Winkler 2003).

If
$$V(U) = dV(X^P)$$
then
$$V(X^R) = (1+d)V(X^P) \tag{5}$$

This does not yield exact equality of the correlation-matrices of the data sets transformed back to levels. The parameter d can be used to calibrate the anonymisation level.

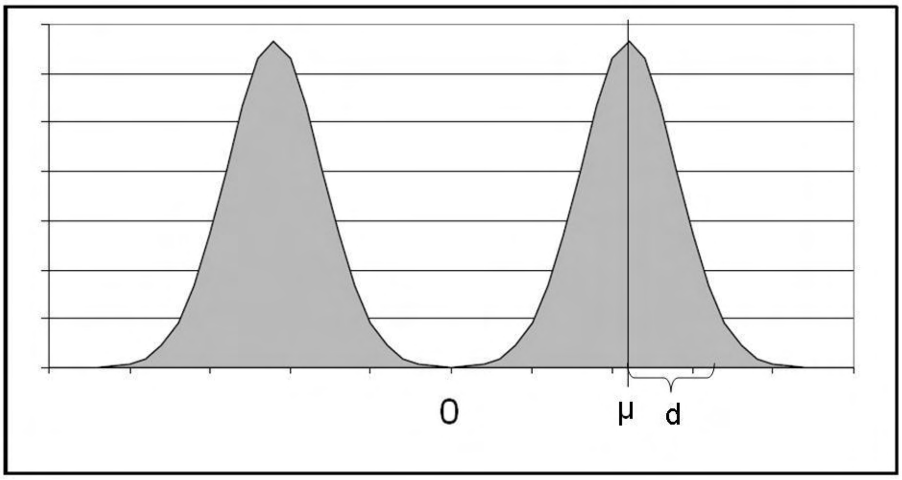
The anonymisation will therefore consist of three steps:

1. Logarithmic transformation on the data matrix. Take the logarithm element wise of positive values and the negative logarithm of absolute values of negative values. Especially when representing proportions the definition area should be conserved for logical reasons. In the most recent version of the program, no logarithms will be taken of proportions, i.e. values between -1 and 1 are not transformed.
2. Adding random noise to the transformed data. To achieve best results the random noise will be drawn from a mixture distribution and transformed to have a correlation matrix similar to the one of the transformed data matrix. As the anonymisation of the logarithmized data would lead to a positive bias in comparison to the original data, a controlled version (see chapter 4) is used to reduce the bias.
3. Transforming the dataset back to levels. The antilog of each data point is taken, respecting the rules for positive and negative original values.

3. Mixture distribution und proportionality of correlation

Random noise will be drawn from a mixture of multivariate normal distributions. The idea of using a mixture of multivariate normal distributions was first put forward by Roque (2000) and used by Yancey (2002). It allows to control the expectation value and still draw random numbers away from the expectation value. As can be seen in Figure 1, very few random numbers are close to 0, hence most data points will be shifted away from their true values and almost no data points are kept unaltered. Kim and Winkler (2003) used truncated distributions to control the maximum change.

Figure 1: Mixture of two normal distributions



Source: Höhne (2008).

A summary of characteristics of mixtures of distributions can be found in Ronning (2009). A mixture distribution with expectation 0 and covariance Σ is defined as:

$$f(x;0,\Sigma) = \sum_{j=1}^{k} w_j f(x_j;\mu_j,\Sigma_j) \tag{6}$$

where k is the number of distribution functions used, w_j the weight attributed to distribution function j and $f(x;\ \mu_j,\ \Sigma_j)$ is a multidimensional distribution function with mean μ_j and covariance Σ_j. Our suggested method uses a mixture of two symmetric normal distributions, i.e. two identical normal distributions shifted by their expectation value to the left and to the right of the expectation value 0 for the mixed distribution. The noise can then be written as

$$\mathrm{f}(\mathrm{x};0,\Sigma) = 0.5\,\mathrm{f}(\mathrm{x}_1;\mu_1\underline{1},\Sigma_1) + 0.5\,\mathrm{f}(\mathrm{x}_2;\mu_2\underline{1},\Sigma_2) \tag{7}$$

where the random matrices in each component are drawn from a multidimensional normal distribution with expectation μ_1, μ_2 and the variance-covariance matrices Σ_1, Σ_2. To get easily expectation 0 and the variance-covariance matrix Σ it should be assumed that $\mu_1=-\mu_2=\mu$ and $\Sigma_1=\Sigma_2$.

The added noise to the logarithmized variables is equivalent to a multiplicative error on the original values. If Σ is proportional to the original variance-covariance matrix like suggested in Kim and Winkler 2003 this results changes of the original variables depending on the variation of the variables. In our simultaneous tests with variables like i.e. employees and turnover it was not possible finding parameters with a sufficient anonymisation level and retaining the usability for both variables. So we change the covariance structure of the noise to be proportional to the original correlation matrix with a constant variance s^2 on the diagonal elements to maintain the correlation structure. That leads to a relative constant level of change for all variables.

The shift parameter μ is set once for the whole data set and can be chosen according to the requested level of anonymity. For one unit of observation i all values should either be enlarged or downsized. The expectation of the random vector μ*1 or –μ*1 will be the same in each element (for all characteristics). So for every characteristic j of object i a random value will be drawn from the same component of the mixture distribution.

The two parameters μ and d can be used to adjust the protection level. Higher shift parameter μ and variation within the components of the distribution make disclosure more difficult. The structure of mixture distributions is good as the shifting of -μ or +μ allows a smaller variation within the components to reach the same total variation and hence degree of anonymisation. The random error from the components of the mixture distribution are necessary to protect further characteristics of one firm if an attacker does know one characteristic of the firm. He cannot simply calculate the shift parameter and use the ratio of values to deduct the other values of this firm. While the mixed normal distribution is symmetric the retransformed variables where multiplied by a not symmetric lognormal distributed noise (see figure 2). The parameters of the lognormal distribution can no longer be interpreted that easily like in nontransformed models. Therefore, it is better to control the anonymisation level by μ (exp(μ)) and the overall standard deviation s (exp(s)) of the mixed distribution. For the user of the dataset it is easier to understand are measures of relative change of original values which can be calculated from parameters μ and s.

For the construction of the mixed distribution we use the algorithm

- o Chose μ as the shift parameter, chose s^2 as the variation of the mixture distribution.
- o s is usually greater than μ (but close to it) so that very few observations are close to the original value
- o The variance-covariance matrix of the components can be calculated as

$$\Sigma_1 = s\Sigma - \mu^2 \underline{1}\underline{1}^T \tag{8}$$

  where Σ_1 is the variance-covariance-matrix of each component in the mixture and has to be positive definite, therefore μ cannot be chosen too big in relation to s.
- o As a data matrix X^P of only continuous variables and with dimension n*m is to be masked, the mixed distribution random matrix can now be constructed by two random matrices of the size (n/2)*m drawn from the distribution $f(x_i, 0, \Sigma_1)$. A bleaching method is applied (Brand 2002). Then add to all elements of the first matrix μ and to the second matrix -μ. A random mix of the rows of the two matrices leads to the mixed distribution.

This random noise from a mixed distribution leads to an unbiased expectation in the anonymous logarithmic values and the same correlation pattern as in the logarithm of the original values. Small deviations in tests result from missing values in the dataset. Missing values will remain missing, have not been transformed and no random noise was added. The structure of missing values is important for logical dependencies in the dataset, i.e. in panel data all variables should be missing if the firm did not exist in the year. The variance in the anonymised dataset is higher then in the original data. The bias in the variance depends on the variance of the random values which on the other hand influences the anonymisation level. For the retransformed values results also a higher variance and a bias in the expectation like it is shown in Kim and Winkler (2003, p.7) or Ronning (2009, pp.11-12).

To get a high anonymisation level, a high standard deviation s and a high shift μ of the two components in the mixed distribution are needed. In order to get a minimal bias in the data structure file small values for s and μ would be required. To solve the contradiction we introduce the controlled anonymisation (similar to Ronning et al. (2005) and Nayak, Sinha and Zayatz (2010)).

4. Controlled Anonymisation

Using a lognormal distribution as multiplicative error term would yield a bias in the anonymous data. This has been demonstrated in Kim and Winkler (2003) and Roque (2000) (as mentioned above). The expectation of a lognormal distribution is:

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{9}$$

When transforming the masked data back to levels, the noise in the presented anonymisation technique can be regarded as drawn from a mixture distribution of two lognormal distributions. Hence the expectation of the noise can be written as follows. Inserting the parameters used in the practical study (see section 5) would lead to an overestimation of antilog variable means by 6.5%.

$$\begin{aligned} E(U) &= 0.5\exp\left(\mu + \frac{d^2}{2}\right) + 0.5\exp\left(-\mu + \frac{d^2}{2}\right) \\ &= 0.5\exp\left(0.25 + \frac{0.065025}{2}\right) + 0.5\exp\left(-0.25 + \frac{0.065025}{2}\right) = 1.065 \end{aligned} \tag{10}$$

Therefore the anonymisation method has been amended so that the data is not only unbiased in logarithms but also that the properties of the antilogarithmized variables are best preserved. A correction of the mean and variance like presented in Kim (1986) or Oganian and Karr (2010) is not used, because it leads to systematic errors in the values. Correction of variance as proportional reduction of the distance between the value to the mean of the variable, results in systematic increasing of small units and a decreasing of big units. That leads to systematically biased values in the groups. Good results for antilogarithmized variables can be achieved by interrupting the stochastic process and introducing a control mechanism. To reduce the error due to masking, the procedure tries to balance enlarging and diminishing effects systematically.

Some sorting algorithm is applied that groups the dataset into pairs of two similar objects. This method is familiar from microaggregation. The idea of nearest point next (NPN) has been suggested in Domingo-Ferrer et al. (2006). The choice of the distance measure determines which dimension is to be optimized. At present the best conservation of variable means in the antilog data is implemented. The Euclidian distance is an adequate measure to assess the “distance” of two objects as a dominating variable is misleading in the panel context where there are entering and exiting firms. If every pair of objects is combined with a pair of random vectors (one each from the two components of the mixed distribution), then a compensation effect is in place. One object will be enlarged and the other one diminished. The level of compensation depends on the distance between the two objects. More similar objects have a higher compensation effect.

The algorithm of controlled stochastic noise is as follows:

1. The two data records in the original data file (hence superindex o) to be masked next are found by determining which data record i has the largest distance to the centroid (vector of means) of the remaining data file.

$$\max_{i}\left(\sum_{k=1}^{m}\left(\frac{x_{i,k}^{o}-\overline{x_{k}^{o}}}{\overline{x_{k}^{o}}}\right)^{2}\right) \tag{11}$$

2. Data record j with the smallest distance to the data record i, that has been selected in step 1, is looked for by optimizing:

$$\min_{j}\left(\sum_{k=1}^{m}\left(\frac{x_{j,k}^{o}-x_{i,k}^{o}}{\overline{x_{k}^{o}}}\right)^{2}\right) \tag{12}$$

3. The two data records i and j are positioned to the end of the data set that is still to be worked on (position n and n-1).

4. For the two selected data records of the original data material x_n and x_{n-1} and given vectors of random numbers u_n and u_{n-1}, the one of the two sketched masking variants v_1 and v_2 is chosen that minimizes the summarized relative error in the means of the antilog variables:

$$\min_{v1,v2}\left(\sum_{k=1}^{m}\left(N\sum_{p=1}^{N}\left(x_{p,k}^{a}-x_{p,k}^{o}\right)\Big/\sum_{p=1}^{N}x_{p,k}^{o}\right)^{2}\right)$$

$$\text{where}: \tag{13}$$

$$v_1: \quad x_{n,k}^{a}=x_{n,k}^{o}e^{u_{n,k}} \quad and \quad x_{n-1,k}^{a}=x_{n-1,k}^{o}e^{u_{n-1,k}}$$

$$v_2: \quad x_{n,k}^{a}=x_{n,k}^{o}e^{u_{n-1,k}} \quad and \quad x_{n-1,k}^{a}=x_{n-1,k}^{o}e^{u_{n,k}}$$

5. The two units are then stored in the material as masked. The anonymisation will be concentrated on the remaining data set (n=n-2) and will continue analogously as steps 1 through 5 are repeated until all data records are masked.

As the original data set is treated in descending order of size, not only the systematic controlled compensation effect is in place but the “errors introduced by masking” are getting smaller. At the end of the anonymisation process, the difference in first and second moments between original and masked data will be minimal. A similar algorithm was introduced in Ronning et al. (2005, p.67).

5. Application

The described version of stochastic noise has been applied to a data set that can be accessed by the scientific community through the Research Data Centres (RDC) of the statistical offices in Germany. For testing the anonymisation method, the monthly report of establishments in the manufacturing sector was chosen. All establishments with 20 or more employees have to report to this statistic. A panel for the years 1999 to 2002 was used. The statistic contains 26 variables such as number of employees, value of sales, domestic and foreign sales. More details of this statistic can be found in the quality report by the Federal Statistical Office (Statistisches Bundesamt 2009). The unbalanced panel contains data on almost 60 000 entities. The variables are highly skewed.

Table 1: Parameters used for stochastic noise

Variant	1	2	3
Shift parameter μ	0.25	0.25	0.25
Total variation in random noise s	0.255	0.265	0.27

Source: Research Data Centre of the Statistical Offices in Germany, Monthly Report for Establishments from Manufacturing and Mining, 1999-2002, own calculations.

Three different parameter combinations have been used to generate three anonymous data sets that were included in a comparison of anonymisation methods. Table 1 gives an overview of the parameters used. In the spirit of a quality measures approach the quality of the anonymised data has been assessed. Shift parameter μ=0.25 was used, because a change of values by more than 15%

was defined as useless values for data attackers. Most single values are masked to a secure anonymisation level.

5.1 Data Utility

a) Comparing variable averages and variations

Due to the controlled noise, means in variables can be conserved quite well though not perfectly. This holds for all three parameter combinations. The average relative deviation of variable means is 1.07% while the biggest deviation in a variable mean is 4%. The data set masked with the highest total variation in the noise has a slightly higher deviation than data masked with little variation. Relative deviations in standard deviations are higher than had been accomplished in means. On average, standard deviations of the anonymous variables are 5% higher than in the original data.

b) Comparison of correlation matrices

Correlations in the logarithmized anonymous data can be conserved very well by the suggested anonymisation technique. Absolute deviations of all pairwise correlations are on average around 0.006. Preserving the correlation in the log values leads also to an approximate conservation of the correlation pattern in the values transformed back to levels. The average absolute deviation over all variables in the data set is around 0.02 for all three parameter combinations. Looking only at 3 selected variables that will probably be employed in most studies using this data set and that feature far less missing values, the correlations between these variables is preserved better. The deviations are smaller by the factor 1/4. The exhibited difference in correlations is therefore in part indebted to the structure of missing values.

c) Regression results with anonymised data

Reproducing research questions addressed in the literature with data masked by the suggested variant of stochastic noise yields very promising results. When looking for the determinants of enterprise growth (Strotmann 2002), the following model can be estimated with enterprise growth as dependent variable, α_i as unit of observation specific effect and x_{it} representing a vector of explanatory variables (Biewen 2010):

$$growth_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \tag{9}$$

Estimating a random and a fixed effects model yields the results displayed in Table 2. Estimating both models with original and anonymous data yields nearly the same results. The two variants with low noise variation perform better as coefficients have always the same sign, the same order of magnitude and in all but one case the same level of significance. Higher noise variation leads to a change in signs of one coefficient.

Table 2: Determinants of enterprise growth, fixed and random effects regression.

Regressors	**Coefficients Original**	**Coefficients Anonymous**		
		s=0.255	**s=0.265**	**s=0.27**
RE-estimation:				
log. employment	-0,0760***	-0,0652***	-0,0536***	0,0494***
(log. employment)²	0,0154***	0,0124***	0,0085***	0,0070***
(log. employment)³	-0,0010***	-0,0008***	-0,0005***	-0,0004***
Dummy 2000	-0,0125***	-0,0126***	-0,0124***	-0,0127***
Dummy 2001	-0,0608***	-0,0618***	-0,0619***	-0,0622***
Dummy West	-0,0077***	-0,0077***	-0,0072***	-0,0071***
MBU Dummy+	-0,0285***	-0,0284***	-0,0280***	-0,0276***
log. Herfindahl	0,0026***	0,0021***	0,0029***	0,0034***
log. export ratio	0,0162***	0,0163***	0,0182***	0,0189***
Constant	0,1788***	0,1661***	0,1711***	0,1750***
***FE-estimation*:**				
log. employment	-1,1268***	-1,0689***	-0,9665***	-0,9419***
(log. employment	0,1176***	0,0948***	0,0438***	0,0278***
(log. employment)³	-0,0055***	-0,0051***	-0,0023***	-0,0015***
Dummy 2000	-0,0047***	-0,0040**	-0,0028***	-0,0014***
Dummy 2001	-0,0558***	-0,0561***	-0,0548*	-0,0529***
log. Herfindahl	-0,0170***	-0,0168***	-0,0199***	-0,0198***
log. export ratio	-0,0485***	-0,0364***	-0,0357***	-0,0533***
Constant	2,7883***	2,9461***	3,1920***	-0,9419***

*Notes: significance at *** p<0.01,** p<0.05, * p<0.1*
+MBU Dummy: Unit is part of an enterprise comprising several units.

Source: Research Data Centre of the Statistical Offices in Germany Monthly Report for Establishments from Manufacturing and Mining, 1999-2002, calculations by Biewen 2010.

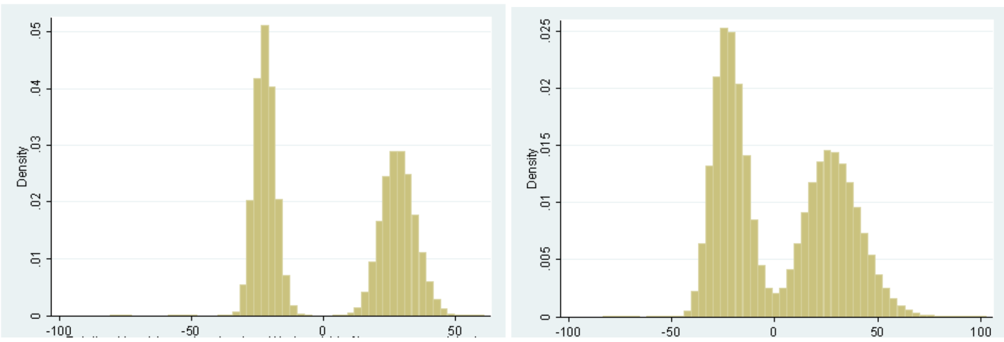
Thus it can be stated that the presented data set treated by the proposed method of stochastic noise would be suited well as a data structure file (DSF) in remote data processing. Researchers devising programme syntax could start to specify econometric models with the DSF correctly.

5.2 Data security

Detailed matching experiments are conducted by a partner in the “infinitE” project group (Brandt and Zwick 2009). Due to constraints in space they are not presented. Another good measure for data security are the relative deviations of single values from the original data. High deviations lead to a small usability for data attackers independent from success of data matches. Looking at the relative deviations of the data points is relevant to measure the anonymity of the data set. Due to the mixture distribution, most random values of the noise are „far away“ from zero. However, Figure 2 shows clearly the effect the total variation parameter s has. The higher the variation, the more stochastic values drawn from the distribution components are in absolute value close to 0. One could classify all data points that have been changed by less than 15% as critical. The variant with s=0.255 has only 2.76% of all single anonymous data points in the range of ±15%. For the variant s=0.27 there are 16.86% of all relative deviations of single data points in this critical interval. The smaller total noise variation performs better in this respect. On the other hand, ratios are protected less, i.e. if an attacker knows one value of a unit he would get closer estimating other values for the same unit by using the ratio.

As stochastic noise can only be applied to metric data, anonymisation by stochastic noise should be combined with another anonymisation method that treats categorical variables.

Figure 2: Relative deviations in single data points.



s=0.255 s=0.27

Source: Research Data Centre of the Statistical Offices in Germany, Monthly Report for Establishments from Manufacturing and Mining ring, 1999-2002, own calculations.

6. Conclusion

This paper presented a special variant of stochastic noise. The data transformed in logs is masked with an additive mixed normal distributed noise that has almost the same correlation matrix as the original log data. This is comparable to a multiplicative noise drawn from a mixed lognormal distribution. A controlled masking method is applied to reduce the bias resulting from mixed lognormal noise. As noise is drawn from a mixture distribution, few data points are changed by a small percent-tage, most data elements have been changed decisively. The two parameters, the shift parameter μ and the overall variation in the noise s^2, can be used to calibrate the anonymisation effect.

Controlled stochastic noise to log data has been proven to fulfil many criteria set up for reproducing the structure of original data while providing a high degree of anonymity. As advantages of the method one can enumerate its capability in approximately preserving means for the whole population, little deviations for subpopulations result from the stochastic draw of the masking errors. Correlations are conserved for the anonymous data in logs and approximately for the anonymous data. Signs and structural zeros are conserved by this anonymisation method.

As the programmed version of stochastic noise is applied to data in wide format and all values for one observation unit are either enlarged or scaled down, ratios and growth rates are very well replicated. These ratios are systematically preserved but only in an approximate way due to anonymisation requirements. A perfect ratio would constitute a disclosure risk. The anonymous data performed well in panel analysis executed by partners within the “infinitE” project (Biewen 2010). Analysis of gross job flows disaggregated for regions and industries led to results that differed from results with original data by very little. In regressions looking for determining factors of establishment growth the same variables proved to be significant and the coefficients had the same sign and order of magnitude.

This variant of stochastic noise is currently tested whether it can be used to produce data structure files that preserve the syntactic and semantic structure of data sets used in remote data processing and that have, at the same time, a high level of confidentiality and can thus be offered for free download on the homepages of the Research Data Centres of the statistical offices in Germany.

Stochastic noise does not yet yield results exactly as wanted. Parameter calibrations are still necessary to comply with the requirements for absolute anonymity. Yet to reach absolute anonymity with the panel of manufacturing firms, as a next step an anonymisation method for categorical variables has to be tested.

References

Biewen, E. (2010): Erster Verfahrensvergleich von Multipler Imputation und Stochastischer Überlagerung. Internal Working Paper presented at a Meeting of the “infinitE” Project Group. Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

Brand, R. (2002): Masking through Noise Addition. In: Domingo-Ferrer, J. (Eds.): Inference Control in Statistical Data Bases – From Theory to Practice, LNCS 2316, Springer, p. 97-116.

Brandt, M. and M. Zwick (2009): Improvement of the Informational Infrastructure – on the Way to Remote Data Access in Germany. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, Spain, 2-4 December 2009, Working Paper No. 16.

Domingo-Ferrer, J., A. Martínez-Bellesté, J. M. Mateo-Sanz and F. Sebé (2006): Efficient Multivariate Data-Oriented Microaggregation, The International Journal on Very Large Data Bases, Volume 15 (4), Springer, p.355-369.

Höhne, J. (2008): Anonymisierungsverfahren für Paneldaten. In: Wirtschafts- und Sozialstatistisches Archiv (2008) Bd. 2, Springer, p. 259-275.

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

Kim, J. J. und W. E. Winkler (2003): Multiplicative Noise for Masking Continuous Data. Research Report Series (Statistics #2003-01), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233.

Nayak, T. P., B. Sinha and L. Zayatz (2010): Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection. Research Report Series #2010-05, Statistical Research Division, U.S. Census Bureau, Washington, D.C. 20233.

Oganian, A. and A. F. Karr (2010): Masking Methods that Preserve Positivity Constraints in Microdata. Journal of Statistical Planning and Inference, In Press, Corrected Proof, Available online 25 May 2010.

Ronning G., R. Sturm, J. Höhne, R. Lenz, M. Rosemann, M. Scheffler and D. Vorgrimmler (2005): Handbuch zur Anonymisierung wirtschaftsstatischer Mikrodaten. In: Statistisches Bundesamt (Hrsg.) Statistik und Wissenschaft Bd. 4, Wiesbaden.

Ronning, G. (2009): Stochastische Überlagerung mit Hilfe der Mischungsverteilung. Schätzung linearer (Panel-)Modelle auf Basis anonymisierter Daten [Version 49]. IAW Discussion Paper 48.

Roque, G. M. (2000): Masking Microdata Files with Mixtures of Multivariate Normal Distributions. Dissertation submitted at University of California Riverside, unpublished.

Statistisches Bundesamt (2009): Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht, März 2009, available at: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Qualitaetsberichte/VerarbeitendesGewerbeIndustrie/Monatsteineerden,property=file.pdf.

Strotmann, H. (2002): Arbeitsplatzdynamik in der baden-würtembergischen Industrie – eine Analyse mit amtlichen Betriebspaneldaten, Hohenheimer Volkswirtschaftliche Schriften, Band 39, Peter Lang Verlag, Frankfurt a.M.

Yancey, W. E. (2002): Working Paper for Mixture Model Additive Noise for Microdata Masking. Research Report Series (Statistics #2002-03), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233.

Bisher sind in der Reihe folgende FDZ-Arbeitspapiere erschienen:

Arbeitspapier Nr. 36: Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, Alexander Vogel, April 2011

Arbeitspapier Nr. 35: Temporary agency work and firm performance, S. Nielen/ A. Schiersch, April 2011

Arbeitspapier Nr. 34: Harmonisation of statistical confidentiality in the Federal Republic of Germany, M. Brandt/ A. Crößmann/ C. Gürke, März 2011

Arbeitspapier Nr. 33: Remote Access. Eine Welt ohne Mikrodaten ??, G. Ronning/ P. Bleninger/ J. Drechsler/ C. Gürke, Februar 2011

Arbeitspapier Nr. 32: Compiling a Harmonized Database from Germany´s 1978 to 2003 Sample Surveys of Income and Expenditure. T. Bönke/ C. Schröder/ C. Werdt, Mai 2010

Arbeitspapier Nr. 31: The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009

Arbeitspapier Nr. 30: Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten - Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/ M. Zwick, Januar 2008

Arbeitspapier Nr. 29: Reichtum in Niedersachsen und anderen Bundesländern -Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008

Arbeitspapier Nr. 28: Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009

Arbeitspapier Nr. 27: Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008

Arbeitspapier Nr. 26: Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/ K. Kohn/ A. C. Lembcke, August 2008

Arbeitspapier Nr. 25: German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008

Arbeitspapier Nr. 24: Zum Einkommensreichtum Älterer in Deutschland – Neue Reichtumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/ J. Merz, Februar 2008

Arbeitspapier Nr. 23: Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/ D. Oberschachtsiek/ R. Pohl, November 2007

Arbeitspapier Nr. 22: Stichprobendaten von Versicherten der gesetzlichen Kranken-versicherung - Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007

Arbeitspapier Nr. 21: KombiFid - Kombinierte Firmendaten für Deutschland, S. Bender/ J. Wagner/ M. Zwick, November 2007

Arbeitspapier Nr. 20: Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007

Arbeitspapier Nr. 18: Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007

Arbeitspapier Nr. 17: Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/R. Lenz, Dezember 2006

Arbeitspapier Nr. 16: Die europäische Union - Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006

Arbeitspapier Nr. 15: Keeping in Touch - A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/ L. Osberg, November 2006

Arbeitspapier Nr. 14: Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/ C. Gräb/ S. Kriete-Dodds, November 2006

Arbeitspapier Nr. 13: Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006

Arbeitspapier Nr. 12: Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006

Arbeitspapier Nr. 11: Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

Arbeitspapier Nr. 10: Konstruktion und Bewertung eines ökonomischen Einkommens aus der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/ F. Neher/ C. Schröder, August 2006

Arbeitspapier Nr. 9: Anonymising business micro data - results of a German project, R. Lenz/ M. Rosemann/ D. Vorgrimler/ R. Sturm, Juni 2006

Arbeitspapier Nr. 8: Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/H.-P. Hafner/ D. Schmidt, Juni 2006

Arbeitspapier Nr. 7: A standard for the release of microdata, R. Lenz/ D. Vorgrimler/ M. Scheffler, Juni 2006

Arbeitspapier Nr. 6: Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006

Arbeitspapier Nr. 5: De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/ M. Zwick, Oktober 2005

Arbeitspapier Nr. 4: Matching German turnover tax statistics, R. Lenz/ D. Vorgrimler, Juni 2005

Arbeitspapier Nr. 3: The research data centres of the Federal Statistical Office and the statistical offices of the Länder,
S. Zühlke/ M. Zwick/ S. Scharnhorst/ T. Wende, März 2005

Arbeitspapier Nr. 2: Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/ M. Zwick, April 2005

Arbeitspapier Nr. 1: Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005

Statistische Ämter des Bundes und der Länder,
FDZ-Arbeitspapier Nr. 37, Masking Micro Data with Stochastic Noise, 2011