Working Paper No. 24
ENGLISH ONLY

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

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

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

MICRODATA DISCLOSURE BY RESAMPLING – EMPIRICAL FINDINGS FOR BUSINESS SURVEY DATA

Contributed Paper

Submitted by the Centre for European Economic Research (ZEW), Germany[1]

[1] Prepared by Sandra Gottschalk (gottschalk@zew.de).

Microdata Disclosure by Resampling - Empirical Findings for Business Survey Data

by
Sandra Gottschalk

Centre for European Economic Research (ZEW)

March 19, 2003

Preliminary Version

Abstract:

A problem which statistical offices and research institutes are faced with by releasing micro-data is the preservation of confidentiality. Traditional methods to avoid disclosure often destroy the structure of data, i.d., information loss is more or less high. In this paper I discuss an alternative technique of creating scientific-use-files, which reproduce the characteristics of the original data quite well. It is based on an idea of Fienberg (1997 und 1994) [4], [5] to estimate and resample from the empirical multivariate cumulative distribution function of the data to get synthetic data. The procedure creates datasets - the resample - which have the same charateristics as the original survey data. In this paper I present some applications of this method with (a) simulated data and (b) innovation survey data, the Mannheim Innovation Panel (MIP), and compare resampling with a traditional method of disclosure control, disturbance with multiplicative error, concerning confidentiality on the one hand and the usage of the disturbed data for different kinds of analyses on the other hand. Univariate and multivariate distributions can be better reproduced by resampling. Linear regression results can be reproduced quite well with perturbed data as well as with resamples. Anonymized data with multiplicative perturbed values better protect against re-identification as resampling.

Keywords: resampling, multiplicative data perturbation, Monte Carlo studies, business survey data

JEL Classification: C13, C15, C81

1 Introduction

Empirical research in economic and social science needs information about households and firms, which are collected by statistical offices and official or private research institutes in form of microdata. As computer capability and availability of statistical software increased in the last past years, empirical anaylses and thus demand for microdata have dynamically advanced. German law provides, that microdata of government statistics are only allowed to pass on for scientific use and if disclosure limitation is guaranteed in effect[1]. The same holds for survey-data, conducted by private or official research institutes, if confidentiality is promised to the respondents. Hence, a problem which statistical offices and research institutes are faced with by releasing micro-data is the preservation of confidentiality. Even business survey data are at risk, because disclosure is more likely than for personal data as additional information are easier obtainable and population size is smaller (see e.g. Brand, 2000 [1]). Traditional methods to avoid disclosure often destroy the structure of data, i.e., information loss is more or less high.

In this paper I discuss an alternative technique of creating scientific-use-files[2], resampling, which generates a synthetic microdata file with nearly the same characteristics as the original survey data. It is based on an idea of Fienberg (1997 und 1994) [4], [5] to estimate and resample from the empirical multivariate cumulative distribution function of data. As elements of the resample are only replicates and do not necessarily correspond to any of those individuals in the original sample survey, an identification of true values is not possible. Nevertheless one cannot rule out the possibility of disclosure, as synthetic datasets could be very similar to real characteristics of observations. Especially, extreme values are at risk.

The paper is structured as follows: In the first section I describe the idea of resampling and an easily constructed algorithm to create synthetic data, attributed to Devroye and Györfi (1985) [3] and Silverman (1986) [15]. Afterwards applications with simulated data (Section 3) and business innovation survey data (Section 4, see also Appendix) points out the properties of resamples. Confidentiality and utilizability are examined. In a second step I compare particularly resampling with a traditional method of disclosure control, disturbance with multiplicative error (see e.g. Hwang, 1986 [9]), concerning confidentiality on the one hand and the usage of the disturbed data for different kinds of analyses on the other hand.

2 The Idea of Resampling

To generate synthetic data with the same characteristics as an original survey data file, one has to estimate the density function and then sample from it. It will be impossible to exactly achieve it in practice, as full information about the true density of data is not available. On could apply a parametric approach in assuming a theoretical density function with unknown parameters like the normal distribution. The paramters have to be estimated with the data, e.g. means and variances. To sample from a theoretical

[1] Disclosure should not be possible without unusually high costs and waste of time and energy.

[2] In contrast to public-use-files, which should be totally anonymized, i.d. disclosure is not possible under no circumstances. Scientific-use-files guarantee only disclosure limitation in effect (see above), therefore such files are still exploitable for scientific use.

distribution function is then quite easily done. But in reality survey data will rarely follow a specific theoretical distribution.

Fienberg (1997) [4] proposes non-parametric and semi-parametric estimation methods like kernel density estimators or a Bayesian approach (see also Fienberg et al., 1996 [6]. The estimated cumulative distribution function will differ more or less from the real distribution of the data. Most of the survey data, official statistics as well as surveys from research institutes, undercover the real population. Therefore the sample distribution is merely an estimation of reality. Another source of bias is introduced by measurement error. Furthermore, techniques to estimate multivariate cumulative distribution functions have been only used for low-dimensional data until now. Even three dimensional relations are difficult to describe, and only if the sample size is large enough (e.g. XploRe, Härdle et al., 1991 [8]). One possibility to improve the estimation is to use a Bayesian method: one has to estimate the empirical distribution function and generate the full posterior distribution (dependent distribution). This approach takes into account regression-like relationships within the sample. "It provides a way of formalisung the process of learning from data to update beliefs in accord with recent notions of knowledge synthesis" (Congdon, 2001, p. 1 [2]). In the following a sample is drawn from the posterior distribution. Fienberg et al. (1996) [6] propose to sample from it using Rubin's multiple imputation technique (Rubin, 1993 [13] , 1987 [14]), which includes Bootstrap sampling.

Devroye and Görfi (1985) [3] and Silverman (1986) [15], who deal with nonparametric density estimation and simulation from density estimates, show how to drawn from density function without need to estimate it explicitly. The procedure can be used to create samples that have the underlying characteristics and structure of the real data, but spurious details, that have arisen from random effects, are oppressed. The algorithm for the univariate case is described in the following: suppose a continuous variable $X = X_1, X_2, ..., X_n$ and a kernel density function K with bandwidth h.

1. Draw observations X_Z of the data file X with replacement.
2. Compute k to have probability density function K.
3. Generate $Z = X_Z + hk$.

The kernel function can be simulated from an epanechnikov kernel, for example[3]:

$$K(x) \quad = \quad \frac{3}{4}(1 - x^2) \quad \text{for} \quad |x| \leq 1$$

A simple procedure to simulate from the rescaled Epanechnikov kernel is given by Devroye and Györfi (1985)[3]:

1. Compute three univariate random numbers ZV_1, ZV_2, ZV_3 within $[-1, 1]$.
2. Generate $k = ZV_2$, if $|ZV_3| \geq |ZV_2|$ and $|ZV_3| \geq |ZV_1|$, otherwise $k = ZV_3$.

[3]One can also think of the normal density.

The procedure resamples with replacement from the data and disturb the information in such a manner that the distribution of each variable is retained. The sample size of Z has to be large enough to approximate the distribution of the original data X. The choice of bandwidth h of the used kernel function is rather difficult, because it influences the goodness of fit to the original distribution on the one hand and the probability of disclosure on the other. A narrow bandwidth causes a better approximation of distribution but rises the probability of re-identification, as the resampled values - though synthetic - could be very similar to the original. But one also should consider if an intruder is interested in disturbed values.

A higher dimensional version of the algorithm can be constructed in using directional information in the data, such as the covariance-matrix of X. Therefore, the multivariate distribution can nearly be performed. Hence, Devroye and Györfi (1985) [3] modificate the third step of the algorithm for d-dimensions:

$$\begin{aligned} Z &= X_Z + h\kappa A \\ VCV^{-1} &= A'A \\ \kappa &= [k_1 \quad k_2 \quad \dots \quad k_d]. \end{aligned}$$

VCV is the covariance-matrix of the original variables $X_1, X_2, ..., X_d$, which is used as weight of the different kernels κ. To get first and second moment properties the same as those of the data the procedure can be transformed (Silverman (1986) [15]):

$$Z = \bar{X} + (X_Z - \bar{X} + hk)/(1 + h^2\sigma_k^2/\sigma_X^2)^{1/2},$$

where $\bar{X}$ are the sample mean of X and σ_X^2 and σ_k^2 the variances of X respectively k. These corrections prevent an overestimation of variances. Devroye and Györfi give some modified versions of the above algorithms for simulating from density estimations of various kinds, e.g. for variables which concentrate its mass on an intervall, like positive numbers.

The procedure presented above is only applicable for continuous variables. As most of the survey data contain discrete variables too, one has to find additional masking methods as there exist confidentiality problems concerning them. Especially, regional information and classifications of economic sectors could be meaningfully used for re-identification of individuals.

3 Simulations

To demonstrate the effects of the resampling procedure Monte Carlo simulations are very useful as regularities can be revealed (see e.g. Robert and Casella, 2002 [12]). The datasets contains 2000 observations, respectively, and the procedure is repeated 100 times.

Six variables are simulated by drawing from theoretical distribution functions, whereas one variable is a linear combination of three others. Four different kinds of resamples are constructed, which differ concerning bandwidth of the kernel and the usage of the covariance-matrix of the unmasked data as weights.

1. A1: A bandwidth is used, which leads to a good approximation of the kernel density, but therefore higher disclosure risk. The width specifies the halfwidth of the kernel, the width of the density window around each point.[4]

2. A2: The bandwidth of A1 is multiplied by factor 3.

3. B1: Is the same as A1, additionally the kernels are weighted with the covariance-matrix.

4. B2: Is the same as A2, additionally the kernels are weighted with the covariance-matrix.

For comparison, an anonymized version of the data is constructed by multiplying each variable with random numbers from univarate distribution functions within the intervall [0.5;1.5].

A measure of how much confidentiality is provided by the masking techniques, can be defined as follows (see e.g. Spruill, 1983 [16]):

1. Find the observation in the anonymized file that minimizes the sum of absolute or squared deviations for all common variables.

2. If the observation, which is found in 1., is the same as the one on which the masked file is based and only differs 20% from the original[5], a link is made.

3. The confidentiality criteria is then defined as the percentage of observations for which such a link cannot be made.

This proceeding should be distinguished from an estimation of the re-identification probability. The last takes into account the probability, that observations are in the additional database, which is used for disclosure, and measurement errors (see e.g. Brand, 2000 [1] for a discussion of the re-identification risk of business survey data).

In the simulations I use the sum of absolute deviations to find a link and assume two common variables. In table 1 confidentiality measures of the different kind of resamples and a dataset of masked variables with multiplicative error are shown.

Table 1: Monte Carlo Simulations - Confidentiality Measure (CM)

	Resample A1	Resample A2	Resample B1	Resample B2	Multiplicative Error
CM	86.6%	97.0%	85.0%	96.6%	97.1%

Confidentiality is preserved on a high level by each method. Multiplicative data perturbation and resampling with large bandwidth lead to lowest re-identification risk.

[4]Here a width proposed by the software package STATA is used. It would minimize the mean integrated square error if the data were Gaussian and a Gaussian kernel were used, so is not optimal in any global sense.

[5]If the values differ more than 20%, I presume that confidentiality is still satisfied as uncertainty of a re-identification is too high.

Tables 2 and 3 show the extent of information loss the different perturbation procedures add to the data. Deterioration is measured by the average relative absolute deviation from the original measuring unit, respectively.

Table 2: Monte Carlo Simulations - Average Relative Absolute Deviation in %

Method	Means	Variances	Correlations	Rank-Correlations	Covariances
Resample A1	0.25	0.30	13.06	28.59	13.06
Resample A2	0.74	0.91	39.48	72.06	39.39
Resample B1	0.27	0.29	13.20	46.86	13.18
Resample B2	0.79	0.84	63.75	204.42	63.72
Multiplcative Error	1.22	12.11	68.88	53.52	72.72

Univariate statistics are reproduced quite well by each anonymization method, but multiplicative perturbation distort variance of variables around 12%. The multivariate distributions could not be reproduced if the variables are multiplicated by errors. Resampling preserve correlations and covariances much better (about 13% bias) if the lower bandwidth is chosen, the error clearly increases with bandwidth. Rank correlations also considerably differ from the original datasets, resampling without weights performs the best, with a deviation about 29%.

Table 3 give an impression of the effects on econometric parameter estimation by the different kinds of anonymization.

Table 3: Monte Carlo Simulations - OLS-Regression Results

Variable	Original	Resample A1	Resample A2	Resample B1	Resample B2	Multiplicative Error
Var 1	0.497	0.496	0.469	0.484	0.847	0.448
(t-stat.)	(31.36)	(28.76)	(18.58)	(25.49)	(11.11)	(14.63)
Var 2	1.000	0.991	0.933	0.981	0.847	0.893
(t-stat.)	(104.94)	(95.04)	(61.15)	(85.38)	(42.15)	(48.89)
Var 3	0.202	0.198	0.190	0.193	0.150	0.180
(t-stat.)	(9.03)	(8.08)	(5.32)	(7.16)	(3.18)	(4.21)
Const.	0.703	0.724	0.864	0.756	1.161	0.966
(t-stat.)	(15.41)	(14.46)	(11.84)	(13.7)	(12.13)	(10.73)
R^2	0.92	0.91	0.81	0.89	0.66	0.72

Though parameter estimates with anonymized data significantly differ from the origi-

nal, results are meaningfully retained: The coefficients remain significant unequal to zero as well as signs do not change. Only goodness of fit clearly varies between the original estimation and regression analysis involving multiplicative errors and with weighted resamples and larger bandwidth, a lower part of variance can be explained. But interpretation of the results would not change. Hence, econometric analyses with perturbed variables are still possible.

4 Empirical Application - An Example

In a second step, anonymized data are constructed by using real data - the Mannheim Innovation Panel (MIP)[6] in the manufacturing sector from 2001. Four quantitative variables are chosen - "sales", "number of employees", "investment per sales" and "R&D expenditure per sales". A regression equation is specified using three sector dummy variables indicating innovative branches: chemical industry, manufacture of medical, precision and optical instruments (MPO), and metal-processing and -working industries. As resampling and multiplicative data perturbation are only applicable to continuous data the original values of the sector dummies are taken.

The quantitative variables in this experiment are censured to the left, i.d. have only positive values. The resampling procedure used here does not consider these restrictions. Therefore a few of the synthetic observations in the resample have negative signs. Test have shown that this fact does not matter for a lot of descriptive and regression analyses. But variables which have a markable number of values that are zero - like R&D expenditure - are difficult to reproduce as the share of zeros could not maintained. Some modifications of the resampling procedure are necessary. These will be subject of further work.

In the following, results of the different anonymization processes are shown. Table 4 lists the share of links could not be made with the original data.

Table 4: MIP - Confidentiality Measure (CM)

	Resample A1	Resample A2	Resample B1	Resample B2	Multiplicative Error
CM	58.17%	61.58%	57.25%	57.25%	84.59%

In contrast to the Monte Carlo simulations above, the anonymized datasets involve higher re-identification risks for the individuals. Multiplicative perturbation seems to best protect the data with a confidentiality measure about 85%.

Information loss due to the different anonymization methods is described in tables 5 and 6. The resamples only remarkably differ from the original data concerning rank-

[6]The scientific-use-file of the MIP is freely available for purely non-commercial basic research. The applied anonymization methods are described in Gottschalk, 2001 [7]. Here, each continuous variable is multiplied by different random numbers. In the scientific-use-file of the MIP, firm specific random numbers are used and only "sales" and "number of employees" are perturbed with multiplicative error. Hence, productivity (sales per number of employees) remains constant and the scenarios in this paper are not completely transferable to the scientific-use-files of the MIP.

correlations, whereas errors occur in each statistic when applying multiplicative perturbation to the data.

Table 5: MIP - Average Relative Absolute Deviation in %

Method	Means	Variances	Correlations	Rank-Correlations	Covariances
Resample A1	0.01	0.00	0.00	3.07	0.00
Resample A2	0.02	0.01	0.00	5.48	0.01
Resample B1	0.00	0.00	0.00	1.94	0.00
Resample B2	0.00	0.00	0.00	1.91	0.00
Multiplcative Error	6.38	19.95	4.12	2.81	11.15

Table 6 present the results of paramter estimations of an exemplary linear model, explaining productivity defined as sales per employees. Regression results with perturbed data do not strikingly differ from the original values, although only a few coefficients remain significantly the same (marked with *). But significant parameter estimates with the original data are still significant unequal to zero by using the perturbed samples.

5 Resume

The Monte Carlo studies and application with real data, the Mannheim Innovation Panel, shows the effects of resampling in comparison to multiplicative data perturbation on different kinds of analyses. Univariate and multivariate distributions can be better reproduced by resampling. As expected, with extension of the bandwidth bias and confidentiality increases, but a critical or optimal point is not identified, till now. The consideration of directional information within the original data, in weighting the kernel with the covariance-matrix, does not improve the performance of the resample. This finding might be a consequence of the data generating process and cannot generally be assessed. Further work is necessary.

Linear regression results can be reproduced quite well with perturbed data as well as with resamples. Essential findings of OLS-analyses do not differ. It remains to examine effects on non-linear and semi-parametric model estimations as well as different kinds of model specifications with the MIP.

Though resamples consists of synthetic values confidentiality problems remain. Anonymized data with multiplicative perturbed values performs much better. Confidentiality measures and estimations of re-identification risks should be computed in realistic scenarios, where additional databases are used for a match with perturbed datasets.

Table 6: MIP - OLS-Regression Results

Variable	Original	Resample A1	Resample A2	Resample B1	Resample B2	Multiplicative Error
R&D-intensity	-0.253	-0.302	-0.299	-0.254*	-0.214	-0.238
(t-stat.)	(-2.51)	(-3.96)	(-3.92)	(-3.32)	(-2.80)	(-2.23)
Size	0.014	0.014*	0.014*	0.014*	0.014*	0.017
(t-stat.)	(4.28)	(5.63)	(5.62)	(5.56)	(5.90)	(5.07)
Investment per sales	-0.053	-0.069	-0.067	-0.072	-0.073	-0.047
(t-stat.)	(-2.40)	(-3.32)	(-3.33)	(-3.48)	(-3.59)	(-2.33)
Sector: Chemistry	0.117	0.110	0.110	0.106	0.103	0.133
(t-stat.)	(4.74)	(5.98)	(5.95)	(5.75)	(5.57)	(5.11)
Sector: MPO	-0.036	-0.031	-0.031	-0.032	-0.034	-0.040
(t-stat.)	(-1.37)	(-1.57)	(-1.56)	(-1.65)	(-1.71)	(-1.45)
Sector: Metal	-0.028	-0.026	-0.026	-0.025	-0.024	-0.014
(t-stat.)	(-1.54)	(-1.78)	(-1.77)	(-1.73)	(-1.66)	(-0.75)
Constant	0.208	0.215	0.215	0.215	0.215	0.187
(t-stat.)	(11.00)	(14.85)	(14.84)	(14.78)	(14.70)	(9.38)
R^2	0.04	0.04	0.04	0.03	0.03	0.04
Obs.	1537	2870	2870	2870	2870	1537

A The Mannheim Innovation Panel

The Mannheim Innovation Panel (MIP) was asigned by the German government to conduct an innovation survey representative for the German economy leading to international comparable data on the innovation behaviour of German firms. It started in 1993 as a voluntary mail survey and is constructed as a panel with yearly waves. The population of the MIP covers legally independent German firms in the sectors mining and manufacturing. In 1995 the survey on innovation activities of distributive and business related service sector firms (Mannheim Innovation Panel - Services, MIP-S) was additionally initiated. Up to 2002 the MIP and MIP-S have been running ten times in co-operation with infas Institute for Applied Social Science. The MIP is strongly based on the recommendations on innovation surveys manifested in the Oslo-Manual of the OECD and Eurostat (OECD, 1997 [11]). It gives basic information on product and process innovations, innovation activities and components of innovation expenditure related to these activities (see Janz et al., 2001 [10]). The data are available in an anonymized version (scientific-use-file) to external users for non-commercial basic research. Currently, more than 30 researchers utilize the scientific-use-files.

References

[1] Brand, R. (2000), Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, Nürnberg.

[2] Congdon, P. (2001), Bayesian Statistical Modelling, Wiley Series in Probability and Statistics, New York.

[3] Devroye, L. and L. Györfi (1985), Nonparametric Density Estimation, New York.

[4] Fienberg, S.E. (1997), Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, Technical Report No. 161, Carnegie Mellon University, Pittsburgh.

[5] Fienberg, S.E. (1994), A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality, Technical Report No. 611, Carnegie Mellon University, Pittsburgh.

[6] Fienberg, S.E., R.J. Steele und U. Makov (1996), Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models, Proceedingsof Bureau of the Census 1996 Annual Research Conference, US Bureau of the Census, Washington DC, 87-105.

[7] Gottschalk, S. (2002), Anonymisierung von Unternehmensdaten - Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels, ZEW Discussion Paper 02-23.

[8] Härdle, W., S. Klinke und M. Mller (1991), XploRe - Learning Guide, Berlin.

[9] Hwang, J.T. (1986), Multiplicative Errors-in-Variables Models with Application to Recent Data Released by U.S. Department of Energy, Journal of the American Statistical Association, 81, 395, 680-688.

[10] Janz, N., G. Ebling, S. Gottschalk und H. Niggemann (2001), The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW), Schmollers Jahrbuch 121, Journal of Applied Social Science Studies, 123-129.

[11] OECD (1997), Oslo Manual: Proposed Guidelines for Collecting and Interpreting Technological Innovation Data, Paris.

[12] Rober, C.P. and G. Casella (2002), Monte Carlo Statistical Methods, New York.

[13] Rubin, D. (1993), Discussion - Statistical Disclosure Limitation, Journal of Official Statistics, 9, 461-468.

[14] Rubin, D. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, USA.

[15] Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability 26, London.

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