WP.33
ENGLISH ONLY

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

ANALYTICAL VALIDITY AND CONFIDENTIALITY PROTECTION OF ANONYMISED LONGITUDINAL ENTERPRISE MICRODATA – SURVEY OF A GERMAN PROJECT

Supporting Paper

Prepared by Maurice Brandt (Federal Statistical Office Germany), Michael Konold (Statistical Office of North Rhine-Westphalia, Germany), Prof. Dr. Rainer Lenz (University of applied science, Mainz, Germany) and Dr. Martin Rosemann (Institute for applied economic research, Tübingen, Germany)

Analytical validity and confidentiality protection of anonymised longitudinal enterprise microdata – Survey of a German Project

Maurice Brandt[*], Michael Konold**,
Prof. Dr. Rainer Lenz*** and Dr. Martin Rosemann[****]

[*] Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, maurice.brandt@destatis.de
[**] Statistical Office of North Rhine-Westphalia, Postfach 101105, 40002 Düsseldorf, michael.konold@lds.nrw.de
[***] University of applied science, Holzstrasse 36, 55116 Mainz, rainer.lenz@fh-mainz.de
[****]Institute for applied economic research, Ob dem Himmelreich 1, 72074 Tübingen, martin.rosemann@iaw.edu

Abstract: The access of the scientific community to cross-section data in the field of economic statistics in Germany has considerably improved over the last few years. The purpose of the new project on "Business Panel data and de facto anonymisation" is to extend the data infrastructure in Germany for longitudinal data on local units and on enterprises, so that economic statistical data can be made available to empirical researchers. This paper gives an overview of the project, describes the data sets and outlines the work done so far to assess the analysis potential and de facto anonymity of the data. At the end of the paper, first results of the project are presented.

1. Introduction

The bases for anonymising economic microdata were developed in the project on "De facto anonymisation of business microdata" (Lenz et al. 2006, Statistisches Bundesamt 2005a). A major result of the project was that de facto anonymisation of economic statistical data can be achieved on a cross-section basis. A new project on "Business Panel data and de facto anonymisation" started at the beginning of 2006 and is intended to clearly improve both the data infrastructure in Germany regarding longitudinal data on local units and enterprises and the access of the scientific community to the panel data of economic statistics[1]. The project deals with an improvement of the data supply by longitudinal linkage of statistics which so far have been used mainly on a cross-section basis. The focus is on the cost structure survey in manufacturing, the monthly reports in manufacturing, the survey of investments, the industrial small units survey and the turnover tax statistics, which

[1] The project is carried out jointly by the Institute for Employment Research (IAB), the Institute for Applied Economic Research (IAW), the Research Data Centre (FDZ) of the statistical offices of the Länder, and the Research Data Centre of the Federal Statistical Office.

are processed as longitudinal data sets as part of the project. The local units panel of the Institute for Employment Research was selected for anonymisation by means of multiple imputation method.

As another important element of the project, it is planned to complement the data linked longitudinally by information from the official business register. The main purpose of that work is to identify by means of the business register (cf. Sturm 2006) reasons for missing data, specially demographic information about enterprises, in longitudinal terms and thus to increase the analysis potential of the data. This will permit, for example, to find out on the basis of the business register whether a reporting unit has no longer been included in the survey because it changed to another economic branch or because its number of employees decreased under the cut-off limit. In these cases its ruled out that the enterprise has been shut down or has merged with another enterprise. As regards turnover tax statistics, the turnover data have already been complemented – on the basis of the business register – by employees data for the years 2000 to 2004.

Longitudinal and panel data are demanded more and more often by scientific users because only with such data is it possible to show the dynamics, changes and processes over time. Another advantage of panel data is that unobservable heterogeneity can be controlled, which can be taken account of in the analyses. However, the positive aspects provided by longitudinal data for research evaluations might also prove to be an additional challenge to anonymisation. This is because, across several waves, a structure in the data may be detected which gives additional knowledge to a potential data intruder that might be helpful in reidentification attempts (cf. Lenz 2007).

With a view to maintaining the analysis potential of panel data it must be ensured that developments over time can adequately be analysed also by means of anonymised data and that panel-econometric methods continue to produce consistent estimates (Biewen/Ronning/Rosemann 2007).

One of the questions to be answered by the project is the extent to which the anonymisation methods developed for cross-section data must be further developed for the anonymisation of panel data and what impact such methods have on data protection and on the analysis potential of the panel data of economic statistics (Lenz 2006; Rosemann 2006).

2. The data sets of the project

For the longitudinal linkage and the subsequent anonymisation, economic data were selected for which some experience is available regarding cross-section anonymisation and which are demanded most often by researchers.[2]

2.1 Monthly reports, survey on investments and annual report of small firms in the German industry

Based on the local units as a unit of analysis, the monthly reports in manufacturing, mining and quarrying are a longitudinal linkage of the years from 1995 to 2005. They contain information on employees, wages and salaries, and turnover (Statistisches Bundesamt 2007a). The survey of investments, however, provides information different types of investments (Statistisches Bundesamt 2007c) and basically contains the same local units as the monthly reports. The monthly reports represent a complete elicitation of the local units with 20 or more employees.[3] The range of data is complemented by the annual report of small industrial firms of the years 1995 to 2002, which supplies information from local units with 19 or fewer employees.

For the longitudinal data set, the individual data supplies have been aggregated to form an annual data supply. The data contain information on employees, turnover (domestic and foreign turnover), hours worked, wages and salaries, and investments (Konold 2007). Wagner (2007) contains some examples of comments on the research potential of the monthly reports.

2.2 Cost structure survey

The data of the cost structure survey in manufacturing, mining and quarrying are designed as a longitudinal data set for the years from 1995 to 2005. The cost structure survey is suited for manifold structural analyses (Fritsch et al. 2004) and provides comprehensive information on output, the production factors used, and on the value added of enterprises with at least 20 employees (cf. Statistisches Bundesamt 2007d). The longitudinal data set contains a good 43,000 enterprises for the years 1995 to 2005. The way of processing allows to perform analyses both on a cross-section basis for the reference year and on a longitudinal basis. For the period from 1995 to 2005, there are a good 2,000 enterprises which were questioned every year. A large part of those enterprises come from areas fully covered (branches with few cases, large enterprises). For the years 1999 to 2002, there are still just under 13,300 enterprises which were questioned every year, thus providing sufficient potential for scientific analyses.

[2] In consequence of the project on “De facto anonymisation of microdata of economic statistics”, further economic statistics such as the structure of earnings survey, could be anonymised (cf. Hafner and Lenz 2007).

[3] An exception is 14 economic branches with 10 or more employees (cf. Statistisches Bundesamt 2007b/c).

2.3 Turnover tax statistics

The longitudinal linkage of turnover tax statistics comprises a data set of a total of some 4.3 million enterprises, about 1.8 million of which can be linked for the period from 2000 to 2004 to form a real panel data set. In a first step, the panel data set for 2000-2004 was established for special analyses at the Federal Statistical Office and for remote data purposes. For every case, the file contains a data set with a total of 156 variables for 5 reference years, with differing numbers of variables actually occurring, depending on the existence of the enterprise in the relevant year. Turnover tax statistics contains information on all taxable turnovers, turnover tax, prior tax, and duration of tax liability (cf. Statistisches Bundesamt 2005b).

2.4 IAB panel of local units

The IAB panel of local units is a representative survey among employers on local unit items influencing employment and covers a stratified sample of all local units with at least one employee subject to social insurance contributions in Germany. The panel contains information allowing to perform analyses of the development of labour demand on the labour market in Germany. Items covered include information on the employment trend, weekly hours worked, turnover, and export share, investments and innovation in the local unit, public subsidies, staff structure, vocational training and apprenticeship positions, staff recruited and staff leaving, search for new staff, wages and salaries, hours worked in the local unit, advanced training and continuing education. The local units panel has been produced every year since 1993 in western Germany and since 1996 in eastern Germany by the IAB research unit "Local units and employment". The local units panel contains information of the various waves on about 4,300 to a maximum of some 16,000 local units (cf. Bellmann 2002).

3. Anonymisation Methods for Panel Data and the Analytical Validity of anonymised Panel Data

In the last decade a broad variety of anonymisation methods is described in literature (see for example Brand (2000), Höhne (2003), Statistisches Bundesamt (2005a) and Rosemann (2006)). Anonymisation methods may be subdivided into two groups: methods reducing the information, and more recent methods modifying the values of numerical data (data perturbating methods). When an anonymisation concept for business micro data is developed a mix of these two approaches often seems to be the best solution. Information reducing methods such as the suppression of variables or presenting key variables in broader categories should be preferred, provided that the analyses of interest to the users can still be made. However, if it seems inevitable to additionally apply anonymisation measures which modify the data, a method has to be agreed upon and the parameters of that method need to be balanced appropriately (Lenz et al. 2006).

In Statistisches Bundesamt (2005a) most known anonymisation procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular micro aggregation or stochastic noise has been found convenient for continuous variables whereas ”Post Randomization” (PRAM) can be recommended with some reservations for discrete variables. Additionally, most recently multiple imputation has been suggested by Rubin (1993) for data protection.

The basic idea of (deterministic) micro aggregation is to form groups of similar objects and to substitute the original values by the arithmetic mean of this group (Mateo-Sanz and Domingo-Ferrer 1998).[4] The variants of deterministic micro aggregation principally differ with regard to the question whether the micro aggregation is performed jointly for all numerical variables or separately for each variable.[5] In the first case therefore the same groups are formed for different variables when determining the averages. In the second case (individual ranking) the groups are formed for the several variables separately.

In the case of panel data we have r variables, T periods and N observations. So we can perform the micro aggregation (a) separately for all variables and all periods (Individual Ranking), (b) separately for all variables but jointly for all periods, (c) separately for all periods but jointly for all variables and (d) jointly for all periods and all variables.

Micro aggregation preserves the expected values original but leads to a decreasing variance. Therefore Höhne (2004a) develops a variant of individual ranking that preserves the variances too. He builds up groups of size four. Then for two of these observations in group i anonymised values are given by $x_{i,1/2}^{a} = \bar{x}_{i.} - sd(x_i)$ whereas for the two other anonymised values $x_{i,3/4}^{a} = \bar{x}_{i.} + sd(x_i)$ is used where $\bar{x}_{i.}$ is the average of the variable x in group i and sd(xi) is the standard deviation of x in this group.

The alternative approach of addition or multiplication of stochastic noise is one of the most important data perturbating methods. In the additive case the noise variable usually is assumed to be normally distributed with expectation zero. To increase the data security one can use a mixture distribution of normal distributions where the expectations of the underlying component distributions are unequal to zero. In the case of anonymisation we can restrict ourselves to a mixture distribution of two normal distributed components with expectations $-\mu$ and μ (Roque 2000, Yancey et al. 2002, Höhne 2004b and Statistisches Bundesamt 2005a).

[4] For stochastic micro aggregation see Rosemann (2006).

[5] Also used are variants where the set of numerical variables is subdivided into groups first and where the variables of a group are then micro aggregated jointly (Statistisches Bundesamt 2005a).

We achieve better protection for larger firms if we use multiplicative noise (Statistisches Bundesamt 2005a). In this case the expectation of the noise variable should be one and the values of the noise variable should be limited to the positive band. Several distributions can be used, e.g. lognormal or uniform distribution. As an alternative, also in the multiplicative case a mixture distribution of two normal distributions is used, where the expectations are 1−f and 1+f. The parameter f as well as the standard deviations of the two components (which equal each other) are chosen in such a manner that the values of the noise variable remain positive.

A special variant of a mixture distribution was proposed by Höhne (2004b). The main idea of this approach is that for one observed unit all values are scaled down or scaled up. In other words, for every unit the probability to draw from a normal distribution with expectation 1−f is 0.5 and corresponds to the probability to draw from a normal distribution with expectation 1+f. If we adopt this anonymisation method on the case of panel data we can distinguish several variants for the multiplivative noise variable w_{ijt} of observation i, variable j and period t.

$$w_{ijt} = 1 + d_i f + \varepsilon_{ijt} \qquad (3\text{-}1)$$

$$w_{ijt} = 1 + d_{ij} f + \varepsilon_{ijt} \qquad (3\text{-}2)$$

$$w_{ijt} = 1 + d_{it} f + \varepsilon_{ijt} \qquad (3\text{-}3)$$

$$w_{ijt} = 1 + d_{ijt} f + \varepsilon_{ijt} \qquad (3\text{-}4)$$

In all cases we assume $\varepsilon_{ijt} \sim N(0, \sigma_\varepsilon^2)$ and the variable d takes on +1 and –1 with probability 0.5.

Another auspicious method to anonymise panel data is multiple imputation (Rubin 1993, Raghunathan et al. 2003). In 1993 Rubin suggested to generate fully synthetic data sets to guarantee confidentiality. His idea was to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public.

However, the quality of this method strongly depends on the accuracy of the model used to impute the “missing“ values. If the model doesn’t include all the relationships between the variables that are of interest to the analyst or if the joint distribution of the variables is mis-specified, results from the synthetic data set can be biased. Furthermore, specifying a model that considers all the skip pattern and constraints between the variables can be cumbersome if not impossible

To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that are publicly available in other databases or for variables that contain especially sensitive information leaving most of the data unchanged. This approach has been adopted for some data sets in the US. In our project both approaches are tested in time with data of the IAB establishment panel (first results can be found in Drechsler et al. (2007) and Reiter and Drechsler (2007)).

The methods described above should ensure confidentiality of panel data at the same time the usefulness of data should be gained. The analytic potential is limited on the one hand by the fact that certain analyses are excluded from the start by the anonymisation procedures it selves because either the issue in question cannot be analysed anymore or the method to be used and equivalent methods cannot be applied anymore. This could be the main problem in the case of using methods reducing the information. On the other, such limits result form anonymised data producing results which differ from those based on the original data. When anonymisation procedures are assessed which modify the data, the focus is on the second aspect.

When we use data perturbating methods we have to ensure that distributional properties of the data do not change too much. However, in the project “Business Panel data and de facto anonymisation” the impacts of data perturbating methods on analysis using special qualities of panel data are in focus. On the one hand the project analyses the impacts of the described data perturbating methods on descriptive distribution measures where cross-sectional measures are supplemented by special aspects of panel data, for instance measures relating to the rates of change. On the other hand we focus on the effects of these methods on the estimation of econometric panel models, particularly if we use the within-estimator to control for individual unobservable heterogeneity. These analyses include theoretical derivations as well as simulation experiments and examples with data of official statistics. First results of this work are available.

Biewen et al. (2007) show that the within estimator is consistent in the case of anonymisation by individual ranking. These results correspond to the results of Schmid (2006) for the OLS-estimator. Biewen (2007) derives a consistent within-estimator in the case of anonymisation by multiplicative stochastic noise. The paper focuses on the case of no autocorrelation as yet. Ronning (2007) deals with the effects of stochastic noise using a mixture distribution, for instance the method proposed by Höhne (2004b). In the case of panel data he focuses on the variant described in formula (3-1). However such a distribution will imply correlation of measurement errors. This is of special concern if linear (or nonlinear) models are estimated from data anonymised in such a way. This case so far had not received much attention since usually measurement errors are assumed to be independent across variables. It can be shown that the measurement error of the dependent variable in this case no longer can be considered as harmless to estimation. A consistent fixed effects estimator using the method of Höhne can be found in Ronning (2007) as well as in Biewen (2007).

4. Approaches to assessing de facto anonymity

In order to evaluate the degree of anonymity of previously anonymised micro data, it was necessary to develop a technique for simulating data intrusion scenarios a potentially attacking data intruder might apply. One important constellation is the so-called database cross match scenario. In a database cross match scenario, an attacking data intruder tries to assign as many external database units as possible (additional knowledge) uniquely to units of an anonymised target database in order to extend the external database by target database information.

In a first phase, the database cross match scenario was mathematically modelled as a multicriteria assignment problem, which was then converted, by way of suitable parameterisation, into an assignment problem with one target function to be minimised. Then, the main concern was to choose the best-fitting coefficients of this target function. Whereas in the past a distance measure, generated across all matching variables of the two data sources (key variables and overlaps), proved to be well suited for the examination of cross-sectional data (see Lenz 2006), it turned out that the examination of panel data requires the use of additional, more elaborated measures. As the information on variables, which in the case of panel data is available to a potential data intruder, has been collected in several waves, it seems obvious that this more complex structure should be reflected in the coefficients of the linear program as well. With that goal in mind we have implemented and tested several promising approaches. A more detailed description of these approaches can be found in Lenz (2007).

4.1 Conventional distance based approach

For every numerical key variable vi and every pair of records *(a,b)* in the two data sources, the standardised square deviation is calculated. Afterwards, these component deviations are summed up. It may be advisable in some cases to assign additional weights to the various deviations on variable level. However, a weakness of that measure becomes apparent in cases where the definition of some key variable slightly differs between the two data sources, for example, if a variable such as “number of employees" relates to the number of all employees in absolute terms in one data set, whereas that number is converted into full-time workers in the other data set.

4.2 Correlation-based approach

Let $v^e_1,\dots , v^e_k$ and $v^t_1,\dots, v^t_k$ be ordinal key variables of the external and target data, respectively. We define $\boldsymbol{v^e}$ and $\boldsymbol{v^t}$ as variables from which k realisations have been drawn and calculate the empirical correlation $corr(\boldsymbol{v^e}; \boldsymbol{v^t})$ using Spearman's coefficient. The less this coefficient deviates from 1 the more likely the record pair (a,b) belongs to the same individual. Note that this coefficient can be applied either in case of numerical (and hence also ordinal) variables or in case of categorical variables, whose range forms a well-ordered set.

4.3 Distribution based approach

In a panel data situation we can take it for granted that an attacking data intruder will have information over several years for every key variable, for example, total turnover of an enterprise from 1999 to 2002. In general, we can assume the existence of a bias between the two sources of data in these variables. In order to counteract this problem, we consider the annual changes of a key variable and treat them like a frequency distribution of a discrete variable. Hence, we can apply statistical methods in order to measure the “similarity" of the frequency distributions on either side, external and target data.

4.4 Collinearity approach

A data intruder might have information on two key variables over a period of n years in both sources of data, e.g., “total turnover" $(u_1, \dots , u_n)$ and “number of employees" $(b_1, \dots , b_n)$ of an enterprise. If we interpret the pairs of values (u_i,b_i) as realisations of two random variables, those units that belong together in the different data sources can be expected to reveal empirical correlation cofficients that are ’similar’. It should, however, be considered that what is measured by correlation is just the linear interrelation of two variables. In special cases the two estimated correlation coefficients can diverge from each other very clearly, even if the variables are linked by a direct functional relationship.

4.5 Combination of approaches

Because of the mentioned weaknesses of the various measures described above they are combined in a suitable way. Here we distinguish between two types of combination, *hybrid* and *composite* matching, see Lenz (2007). Once the coeffcients $d(a_i,b_j)$ are calculated, one can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2003).

5. Outlook

Already within the scope of the project on “Business Panel data and de facto anonymisation”, some longitudinal data sets were supplied. They are already used in some scientific research projects. The cost structure survey for the years 1995 to 2005, the monthly reports from 1995 to 2005, the survey of investments from 1995 to 2005 and the survey of small units for the years 1995 to 2002 in manufacturing as well as the data of the turnover tax statistics for 2000 to 2004 are available through remote data access and by using safe scientific workstations at the statistical offices. The data access is also possible for foreign researchers who are interested in the data. The scientists can use the data by remote data access or they can work at the scientific workstations in Germany.

First Scientific Use Files for data utilisation on one’s own workstation will presumably be made available at the beginning of 2009. The project should permit to automate the processing and anonymisation of other economic statistics over time. Also, the experience thus acquired will be used for further projects such as the integration of economic data from various surveys and years.

References

Bellmann, L. (2002). *Das IAB-Betriebspanel. Konzeption und Anwendungsbereiche. In Allgemeines Statistisches Archiv, Bd. 86, H. 2, 177-188.*

Biewen, E. (2007). *Within-Schätzung bei anonymisierten Paneldaten, IAW-Diskussionspapier, to appear.*

Biewen, E., Ronning, G. und Rosemann, M. (2007). *Estimation of Linear Panel Models with Anonymised Business Data. In: IAW-Report 1/2007, 87-114.*

Brand, R. (2000). *Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, Beiträge zur Arbeitsmarkt- und Berufsforschung, 237.*

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2007). *A new approach for disclosure control in the IAB establishment panel.Multiple imputation for a better data access.Tech. rep., IAB Discussion Paper, No.11/2007.*

Fritsch, M., Görzig, B., Hennchen, O. und Stephan, A. (2004). *Cost Structure Surveys in Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 124, 557-566.*

Hafner, H.-P., Lenz, R. (2007). *Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, FDZ-Arbeitspapier Nr. 18.*

Höhne, J. (2003). *Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in:* Gnoss, R./Ronning, G. (Eds.): *Anonymisierung wirtschaftsstatistischer Einzeldaten,* Wiesbaden, *pp. 69-94.*

Höhne, J. (2004a). *Weiterentwicklung von Mikroaggregationsverfahren, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Höhne, J. (2004b). *Varianten von Zufallsüberlagerungen, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Konold, M. (2007). *New possibilities for economic research through integration of establishment-level panel data of German official statistics, Schmollers Jahrbuch / Journal of Applied Social Science Studies 127.*

Lenz, R. (2003). *Disclosure of confidential information by means of multi-objective optimisa- tion. Proceedings of the Comparative Analysis of Enterprise Data Conference (CAED),* London. *(CD-ROM publication, see http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp)*

Lenz, R. (2006). *Measuring the disclosure protection of micro aggregated business microdata - An analysis taking the example of German Structure of Costs Survey. Journal of Official Statistics 22 (4),* Sweden*, 681-710.*

Lenz, R. (2007). *Risk Assessment Methodology for Longitudinal Business Micro Data. To appear.*

Lenz, R., Rosemann, M., Vorgrimler, D. und Sturm, R. (2006). *Anonymising business micro data - results of a German project. Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126 (4), 635-651.*

Little, R. (1993). *Statistical Analysis of Masked Data, in: Journal of Official Statistic, Vol. 9; pp. 407-426.*

Mateo-Sanz, J., Domingo-Ferrer, J. (1998). *A Method for Data-Oriented Multivariate Microaggregation, in: Statistical Data Protection, Proceedings of the conference Eurostat 1999.*

Rosemann, M. (2006). *Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW-Forschungsbericht, Nr. 66,* Tübingen.

Raghunathan, T., Reiter, J., Rubin, D. (2003). *Multiple Imputation für Statistical Disclosure Limitation, Journal of Official Statistics, 19, pp. 1.16.*

Rat für Sozial- und Wirtschaftsdaten (2006). *Eine moderne Dateninfrastruktur für exzellente Forschung und Politikberatung. Bericht über die Arbeit des Rates für Sozial- und Wirtschaftsdaten (RatSWD) während der ersten Berufsperiode vom 31.12.2006,* Frontenhausen

Reiter, J and Drechsler, J. (2007). *Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality. IAB Discussion Paper, No.20/2007.*

Ronning, G. (2007). *Stochastische Überlagerung mit Hilfe der Mischungsverteilung, IAW-Diskussionspapier Nr. 30*

Roque, G. (2000). *Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. thesis, University of California, Riverside.*

Rubin, D. (1993). *Discussion: Statistical Disclosure Limitation, in: Journal of Official Statistics, 9(2), pp. 461-468.*

Statistisches Bundesamt (2007a). *Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007b). *Produktionserhebungen, Qualitätsbericht.*

Statistisches Bundesamt (2007c). *Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007d). *Kostenstrukturerhebung im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2005a). *Statistik und Wissenschaft, Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Band 4.*

Statistisches Bundesamt (2005b). *Umsatzsteuerstatistik, Qualitätsbericht.*

Sturm, R. und Tümmler, T. (2006). *Das statistische Unternehmensregister - Entwicklungsstand und Perspektiven. In: Wirtschaft und Statistik 10/2006, 1021-1036.*

Yancey, W., Winkler, W. and Creezy, R. (2002). *Disclosure Risk Assessment in Perturbative Micro Data Protection, in*: Domingo-Ferrer, J. (Ed.): *Inference Control in Statistical Databases – From Theory to Practice,* Berlin, *pp. 135-152.*

Zühlke S., Zwick, M., Scharnhorst S., Wende, T. (2004). *The research data centres of the Federal Statistics Office and the statistical offices of the Länder, Schmollers Jahrbuch 124, 567-578.*

Analytical validity and confidentiality protection of anonymised longitudinal enterprise microdata – Survey of a German Project

Maurice Brandt[*], Michael Konold**,
Prof. Dr. Rainer Lenz*** and Dr. Martin Rosemann[****]

[*] Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden,
maurice.brandt@destatis.de
[**] Statistical Office of North Rhine-Westphalia, Postfach 101105, 40002 Düsseldorf,
michael.konold@lds.nrw.de
[***] University of applied science, Holzstrasse 36, 55116 Mainz, rainer.lenz@fh-mainz.de
[****] Institute for applied economic research, Ob dem Himmelreich 1, 72074 Tübingen,
martin.rosemann@iaw.edu

Abstract: The access of the scientific community to cross-section data in the field of economic statistics in Germany has considerably improved over the last few years. The purpose of the new project on "Business Panel data and de facto anonymisation" is to extend the data infrastructure in Germany for longitudinal data on local units and on enterprises, so that economic statistical data can be made available to empirical researchers. This paper gives an overview of the project, describes the data sets and outlines the work done so far to assess the analysis potential and de facto anonymity of the data. At the end of the paper, first results of the project are presented.

1. Introduction

The bases for anonymising economic microdata were developed in the project on "De facto anonymisation of business microdata" (Lenz et al. 2006, Statistisches Bundesamt 2005a). A major result of the project was that de facto anonymisation of economic statistical data can be achieved on a cross-section basis. A new project on "Business Panel data and de facto anonymisation" started at the beginning of 2006 and is intended to clearly improve both the data infrastructure in Germany regarding longitudinal data on local units and enterprises and the access of the scientific community to the panel data of economic statistics[1]. The project deals with an improvement of the data supply by longitudinal linkage of statistics which so far have been used mainly on a cross-section basis. The focus is on the cost structure survey in manufacturing, the monthly reports in manufacturing, the survey of investments, the industrial small units survey and the turnover tax statistics, which

[1] The project is carried out jointly by the Institute for Employment Research (IAB), the Institute for Applied Economic Research (IAW), the Research Data Centre (FDZ) of the statistical offices of the Länder, and the Research Data Centre of the Federal Statistical Office.

are processed as longitudinal data sets as part of the project. The local units panel of the Institute for Employment Research was selected for anonymisation by means of multiple imputation method.

As another important element of the project, it is planned to complement the data linked longitudinally by information from the official business register. The main purpose of that work is to identify by means of the business register (cf. Sturm 2006) reasons for missing data, specially demographic information about enterprises, in longitudinal terms and thus to increase the analysis potential of the data. This will permit, for example, to find out on the basis of the business register whether a reporting unit has no longer been included in the survey because it changed to another economic branch or because its number of employees decreased under the cut-off limit. In these cases its ruled out that the enterprise has been shut down or has merged with another enterprise. As regards turnover tax statistics, the turnover data have already been complemented – on the basis of the business register – by employees data for the years 2000 to 2004.

Longitudinal and panel data are demanded more and more often by scientific users because only with such data is it possible to show the dynamics, changes and processes over time. Another advantage of panel data is that unobservable heterogeneity can be controlled, which can be taken account of in the analyses. However, the positive aspects provided by longitudinal data for research evaluations might also prove to be an additional challenge to anonymisation. This is because, across several waves, a structure in the data may be detected which gives additional knowledge to a potential data intruder that might be helpful in reidentification attempts (cf. Lenz 2007).

With a view to maintaining the analysis potential of panel data it must be ensured that developments over time can adequately be analysed also by means of anonymised data and that panel-econometric methods continue to produce consistent estimates (Biewen/Ronning/Rosemann 2007).

One of the questions to be answered by the project is the extent to which the anonymisation methods developed for cross-section data must be further developed for the anonymisation of panel data and what impact such methods have on data protection and on the analysis potential of the panel data of economic statistics (Lenz 2006; Rosemann 2006).

2. The data sets of the project

For the longitudinal linkage and the subsequent anonymisation, economic data were selected for which some experience is available regarding cross-section anonymisation and which are demanded most often by researchers.[2]

2.1 Monthly reports, survey on investments and annual report of small firms in the German industry

Based on the local units as a unit of analysis, the monthly reports in manufacturing, mining and quarrying are a longitudinal linkage of the years from 1995 to 2005. They contain information on employees, wages and salaries, and turnover (Statistisches Bundesamt 2007a). The survey of investments, however, provides information different types of investments (Statistisches Bundesamt 2007c) and basically contains the same local units as the monthly reports. The monthly reports represent a complete elicitation of the local units with 20 or more employees.[3] The range of data is complemented by the annual report of small industrial firms of the years 1995 to 2002, which supplies information from local units with 19 or fewer employees.

For the longitudinal data set, the individual data supplies have been aggregated to form an annual data supply. The data contain information on employees, turnover (domestic and foreign turnover), hours worked, wages and salaries, and investments (Konold 2007). Wagner (2007) contains some examples of comments on the research potential of the monthly reports.

2.2 Cost structure survey

The data of the cost structure survey in manufacturing, mining and quarrying are designed as a longitudinal data set for the years from 1995 to 2005. The cost structure survey is suited for manifold structural analyses (Fritsch et al. 2004) and provides comprehensive information on output, the production factors used, and on the value added of enterprises with at least 20 employees (cf. Statistisches Bundesamt 2007d). The longitudinal data set contains a good 43,000 enterprises for the years 1995 to 2005. The way of processing allows to perform analyses both on a cross-section basis for the reference year and on a longitudinal basis. For the period from 1995 to 2005, there are a good 2,000 enterprises which were questioned every year. A large part of those enterprises come from areas fully covered (branches with few cases, large enterprises). For the years 1999 to 2002, there are still just under 13,300 enterprises which were questioned every year, thus providing sufficient potential for scientific analyses.

[2] In consequence of the project on “De facto anonymisation of microdata of economic statistics”, further economic statistics such as the structure of earnings survey, could be anonymised (cf. Hafner and Lenz 2007).

[3] An exception is 14 economic branches with 10 or more employees (cf. Statistisches Bundesamt 2007b/c).

2.3 Turnover tax statistics

The longitudinal linkage of turnover tax statistics comprises a data set of a total of some 4.3 million enterprises, about 1.8 million of which can be linked for the period from 2000 to 2004 to form a real panel data set. In a first step, the panel data set for 2000-2004 was established for special analyses at the Federal Statistical Office and for remote data purposes. For every case, the file contains a data set with a total of 156 variables for 5 reference years, with differing numbers of variables actually occurring, depending on the existence of the enterprise in the relevant year. Turnover tax statistics contains information on all taxable turnovers, turnover tax, prior tax, and duration of tax liability (cf. Statistisches Bundesamt 2005b).

2.4 IAB panel of local units

The IAB panel of local units is a representative survey among employers on local unit items influencing employment and covers a stratified sample of all local units with at least one employee subject to social insurance contributions in Germany. The panel contains information allowing to perform analyses of the development of labour demand on the labour market in Germany. Items covered include information on the employment trend, weekly hours worked, turnover, and export share, investments and innovation in the local unit, public subsidies, staff structure, vocational training and apprenticeship positions, staff recruited and staff leaving, search for new staff, wages and salaries, hours worked in the local unit, advanced training and continuing education. The local units panel has been produced every year since 1993 in western Germany and since 1996 in eastern Germany by the IAB research unit “Local units and employment”. The local units panel contains information of the various waves on about 4,300 to a maximum of some 16,000 local units (cf. Bellmann 2002).

3. Anonymisation Methods for Panel Data and the Analytical Validity of anonymised Panel Data

In the last decade a broad variety of anonymisation methods is described in literature (see for example Brand (2000), Höhne (2003), Statistisches Bundesamt (2005a) and Rosemann (2006)). Anonymisation methods may be subdivided into two groups: methods reducing the information, and more recent methods modifying the values of numerical data (data perturbating methods). When an anonymisation concept for business micro data is developed a mix of these two approaches often seems to be the best solution. Information reducing methods such as the suppression of variables or presenting key variables in broader categories should be preferred, provided that the analyses of interest to the users can still be made. However, if it seems inevitable to additionally apply anonymisation measures which modify the data, a method has to be agreed upon and the parameters of that method need to be balanced appropriately (Lenz et al. 2006).

In Statistisches Bundesamt (2005a) most known anonymisation procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular micro aggregation or stochastic noise has been found convenient for continuous variables whereas ”Post Randomization” (PRAM) can be recommended with some reservations for discrete variables. Additionally, most recently multiple imputation has been suggested by Rubin (1993) for data protection.

The basic idea of (deterministic) micro aggregation is to form groups of similar objects and to substitute the original values by the arithmetic mean of this group (Mateo-Sanz and Domingo-Ferrer 1998).[4] The variants of deterministic micro aggregation principally differ with regard to the question whether the micro aggregation is performed jointly for all numerical variables or separately for each variable.[5] In the first case therefore the same groups are formed for different variables when determining the averages. In the second case (individual ranking) the groups are formed for the several variables separately.

In the case of panel data we have r variables, T periods and N observations. So we can perform the micro aggregation (a) separately for all variables and all periods (Individual Ranking), (b) separately for all variables but jointly for all periods, (c) separately for all periods but jointly for all variables and (d) jointly for all periods and all variables.

Micro aggregation preserves the expected values original but leads to a decreasing variance. Therefore Höhne (2004a) develops a variant of individual ranking that preserves the variances too. He builds up groups of size four. Then for two of these observations in group i anonymised values are given by $x_{i,1/2}^{a} = \bar{x}_{i.} - sd(x_i)$ whereas for the two other anonymised values $x_{i,3/4}^{a} = \bar{x}_{i.} + sd(x_i)$ is used where $\bar{x}_{i.}$ is the average of the variable x in group i and sd(xi) is the standard deviation of x in this group.

The alternative approach of addition or multiplication of stochastic noise is one of the most important data perturbating methods. In the additive case the noise variable usually is assumed to be normally distributed with expectation zero. To increase the data security one can use a mixture distribution of normal distributions where the expectations of the underlying component distributions are unequal to zero. In the case of anonymisation we can restrict ourselves to a mixture distribution of two normal distributed components with expectations $-\mu$ and μ (Roque 2000, Yancey et al. 2002, Höhne 2004b and Statistisches Bundesamt 2005a).

[4] For stochastic micro aggregation see Rosemann (2006).

[5] Also used are variants where the set of numerical variables is subdivided into groups first and where the variables of a group are then micro aggregated jointly (Statistisches Bundesamt 2005a).

We achieve better protection for larger firms if we use multiplicative noise (Statistisches Bundesamt 2005a). In this case the expectation of the noise variable should be one and the values of the noise variable should be limited to the positive band. Several distributions can be used, e.g. lognormal or uniform distribution. As an alternative, also in the multiplicative case a mixture distribution of two normal distributions is used, where the expectations are 1−f and 1+f. The parameter f as well as the standard deviations of the two components (which equal each other) are chosen in such a manner that the values of the noise variable remain positive.

A special variant of a mixture distribution was proposed by Höhne (2004b). The main idea of this approach is that for one observed unit all values are scaled down or scaled up. In other words, for every unit the probability to draw from a normal distribution with expectation 1−f is 0.5 and corresponds to the probability to draw from a normal distribution with expectation 1+f. If we adopt this anonymisation method on the case of panel data we can distinguish several variants for the multiplivative noise variable w_{ijt} of observation i, variable j and period t.

$$w_{ijt} = 1 + d_i f + \varepsilon_{ijt} \qquad (3\text{-}1)$$

$$w_{ijt} = 1 + d_{ij} f + \varepsilon_{ijt} \qquad (3\text{-}2)$$

$$w_{ijt} = 1 + d_{it} f + \varepsilon_{ijt} \qquad (3\text{-}3)$$

$$w_{ijt} = 1 + d_{ijt} f + \varepsilon_{ijt} \qquad (3\text{-}4)$$

In all cases we assume $\varepsilon_{ijt} \sim N(0, \sigma_\varepsilon^2)$ and the variable d takes on +1 and –1 with probability 0.5.

Another auspicious method to anonymise panel data is multiple imputation (Rubin 1993, Raghunathan et al. 2003). In 1993 Rubin suggested to generate fully synthetic data sets to guarantee confidentiality. His idea was to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public.

However, the quality of this method strongly depends on the accuracy of the model used to impute the “missing“ values. If the model doesn’t include all the relationships between the variables that are of interest to the analyst or if the joint distribution of the variables is mis-specified, results from the synthetic data set can be biased. Furthermore, specifying a model that considers all the skip pattern and constraints between the variables can be cumbersome if not impossible

To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that are publicly available in other databases or for variables that contain especially sensitive information leaving most of the data unchanged. This approach has been adopted for some data sets in the US. In our project both approaches are tested in time with data of the IAB establishment panel (first results can be found in Drechsler et al. (2007) and Reiter and Drechsler (2007)).

The methods described above should ensure confidentiality of panel data at the same time the usefulness of data should be gained. The analytic potential is limited on the one hand by the fact that certain analyses are excluded from the start by the anonymisation procedures it selves because either the issue in question cannot be analysed anymore or the method to be used and equivalent methods cannot be applied anymore. This could be the main problem in the case of using methods reducing the information. On the other, such limits result form anonymised data producing results which differ from those based on the original data. When anonymisation procedures are assessed which modify the data, the focus is on the second aspect.

When we use data perturbating methods we have to ensure that distributional properties of the data do not change too much. However, in the project "Business Panel data and de facto anonymisation" the impacts of data perturbating methods on analysis using special qualities of panel data are in focus. On the one hand the project analyses the impacts of the described data perturbating methods on descriptive distribution measures where cross-sectional measures are supplemented by special aspects of panel data, for instance measures relating to the rates of change. On the other hand we focus on the effects of these methods on the estimation of econometric panel models, particularly if we use the within-estimator to control for individual unobservable heterogeneity. These analyses include theoretical derivations as well as simulation experiments and examples with data of official statistics. First results of this work are available.

Biewen et al. (2007) show that the within estimator is consistent in the case of anonymisation by individual ranking. These results correspond to the results of Schmid (2006) for the OLS-estimator. Biewen (2007) derives a consistent within-estimator in the case of anonymisation by multiplicative stochastic noise. The paper focuses on the case of no autocorrelation as yet. Ronning (2007) deals with the effects of stochastic noise using a mixture distribution, for instance the method proposed by Höhne (2004b). In the case of panel data he focuses on the variant described in formula (3-1). However such a distribution will imply correlation of measurement errors. This is of special concern if linear (or nonlinear) models are estimated from data anonymised in such a way. This case so far had not received much attention since usually measurement errors are assumed to be independent across variables. It can be shown that the measurement error of the dependent variable in this case no longer can be considered as harmless to estimation. A consistent fixed effects estimator using the method of Höhne can be found in Ronning (2007) as well as in Biewen (2007).

4. Approaches to assessing de facto anonymity

In order to evaluate the degree of anonymity of previously anonymised micro data, it was necessary to develop a technique for simulating data intrusion scenarios a potentially attacking data intruder might apply. One important constellation is the so-called database cross match scenario. In a database cross match scenario, an attacking data intruder tries to assign as many external database units as possible (additional knowledge) uniquely to units of an anonymised target database in order to extend the external database by target database information.

In a first phase, the database cross match scenario was mathematically modelled as a multicriteria assignment problem, which was then converted, by way of suitable parameterisation, into an assignment problem with one target function to be minimised. Then, the main concern was to choose the best-fitting coefficients of this target function. Whereas in the past a distance measure, generated across all matching variables of the two data sources (key variables and overlaps), proved to be well suited for the examination of cross-sectional data (see Lenz 2006), it turned out that the examination of panel data requires the use of additional, more elaborated measures. As the information on variables, which in the case of panel data is available to a potential data intruder, has been collected in several waves, it seems obvious that this more complex structure should be reflected in the coefficients of the linear program as well. With that goal in mind we have implemented and tested several promising approaches. A more detailed description of these approaches can be found in Lenz (2007).

4.1 Conventional distance based approach

For every numerical key variable vi and every pair of records *(a,b)* in the two data sources, the standardised square deviation is calculated. Afterwards, these component deviations are summed up. It may be advisable in some cases to assign additional weights to the various deviations on variable level. However, a weakness of that measure becomes apparent in cases where the definition of some key variable slightly differs between the two data sources, for example, if a variable such as “number of employees" relates to the number of all employees in absolute terms in one data set, whereas that number is converted into full-time workers in the other data set.

4.2 Correlation-based approach

Let $v^e_1,\dots , v^e_k$ and $v^t_1,\dots, v^t_k$ be ordinal key variables of the external and target data, respectively. We define $\boldsymbol{v^e}$ and $\boldsymbol{v^t}$ as variables from which k realisations have been drawn and calculate the empirical correlation $corr(\boldsymbol{v^e}; \boldsymbol{v^t})$ using Spearman's coefficient. The less this coefficient deviates from 1 the more likely the record pair *(a,b)* belongs to the same individual. Note that this coefficient can be applied either in case of numerical (and hence also ordinal) variables or in case of categorical variables, whose range forms a well-ordered set.

4.3 Distribution based approach

In a panel data situation we can take it for granted that an attacking data intruder will have information over several years for every key variable, for example, total turnover of an enterprise from 1999 to 2002. In general, we can assume the existence of a bias between the two sources of data in these variables. In order to counteract this problem, we consider the annual changes of a key variable and treat them like a frequency distribution of a discrete variable. Hence, we can apply statistical methods in order to measure the "similarity" of the frequency distributions on either side, external and target data.

4.4 Collinearity approach

A data intruder might have information on two key variables over a period of n years in both sources of data, e.g., "total turnover" $(u_1, \dots , u_n)$ and "number of employees" $(b_1, \dots , b_n)$ of an enterprise. If we interpret the pairs of values (u_i,b_i) as realisations of two random variables, those units that belong together in the different data sources can be expected to reveal empirical correlation cofficients that are 'similar'. It should, however, be considered that what is measured by correlation is just the linear interrelation of two variables. In special cases the two estimated correlation coefficients can diverge from each other very clearly, even if the variables are linked by a direct functional relationship.

4.5 Combination of approaches

Because of the mentioned weaknesses of the various measures described above they are combined in a suitable way. Here we distinguish between two types of combination, *hybrid* and *composite* matching, see Lenz (2007). Once the coeffcients $d(a_i,b_j)$ are calculated, one can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2003).

5. Outlook

Already within the scope of the project on “Business Panel data and de facto anonymisation”, some longitudinal data sets were supplied. They are already used in some scientific research projects. The cost structure survey for the years 1995 to 2005, the monthly reports from 1995 to 2005, the survey of investments from 1995 to 2005 and the survey of small units for the years 1995 to 2002 in manufacturing as well as the data of the turnover tax statistics for 2000 to 2004 are available through remote data access and by using safe scientific workstations at the statistical offices. The data access is also possible for foreign researchers who are interested in the data. The scientists can use the data by remote data access or they can work at the scientific workstations in Germany.

First Scientific Use Files for data utilisation on one’s own workstation will presumably be made available at the beginning of 2009. The project should permit to automate the processing and anonymisation of other economic statistics over time. Also, the experience thus acquired will be used for further projects such as the integration of economic data from various surveys and years.

References

Bellmann, L. (2002). *Das IAB-Betriebspanel. Konzeption und Anwendungsbereiche. In Allgemeines Statistisches Archiv, Bd. 86, H. 2, 177-188.*

Biewen, E. (2007). *Within-Schätzung bei anonymisierten Paneldaten, IAW-Diskussionspapier, to appear.*

Biewen, E., Ronning, G. und Rosemann, M. (2007). *Estimation of Linear Panel Models with Anonymised Business Data. In: IAW-Report 1/2007, 87-114.*

Brand, R. (2000). *Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, Beiträge zur Arbeitsmarkt- und Berufsforschung, 237.*

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2007). *A new approach for disclosure control in the IAB establishment panel.Multiple imputation for a better data access.Tech. rep., IAB Discussion Paper, No.11/2007.*

Fritsch, M., Görzig, B., Hennchen, O. und Stephan, A. (2004). *Cost Structure Surveys in Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 124, 557-566.*

Hafner, H.-P., Lenz, R. (2007). *Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, FDZ-Arbeitspapier Nr. 18.*

Höhne, J. (2003). *Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in:* Gnoss, R./Ronning, G. (Eds.): *Anonymisierung wirtschaftsstatistischer Einzeldaten,* Wiesbaden, *pp. 69-94.*

Höhne, J. (2004a). *Weiterentwicklung von Mikroaggregationsverfahren, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Höhne, J. (2004b). *Varianten von Zufallsüberlagerungen, Arbeitspapier des Projekts `Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten´.*

Konold, M. (2007). *New possibilities for economic research through integration of establishment-level panel data of German official statistics, Schmollers Jahrbuch / Journal of Applied Social Science Studies 127.*

Lenz, R. (2003). *Disclosure of confidential information by means of multi-objective optimisa- tion. Proceedings of the Comparative Analysis of Enterprise Data Conference (CAED),* London. *(CD-ROM publication, see http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp)*

Lenz, R. (2006). *Measuring the disclosure protection of micro aggregated business microdata - An analysis taking the example of German Structure of Costs Survey. Journal of Official Statistics 22 (4),* Sweden*, 681-710.*

Lenz, R. (2007). *Risk Assessment Methodology for Longitudinal Business Micro Data. To appear.*

Lenz, R., Rosemann, M., Vorgrimler, D. und Sturm, R. (2006). *Anonymising business micro data - results of a German project. Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126 (4), 635-651.*

Little, R. (1993). *Statistical Analysis of Masked Data, in: Journal of Official Statistic, Vol. 9; pp. 407-426.*

Mateo-Sanz, J., Domingo-Ferrer, J. (1998). *A Method for Data-Oriented Multivariate Microaggregation, in: Statistical Data Protection, Proceedings of the conference Eurostat 1999.*

Rosemann, M. (2006). *Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW-Forschungsbericht, Nr. 66,* Tübingen.

Raghunathan, T., Reiter, J., Rubin, D. (2003). *Multiple Imputation für Statistical Disclosure Limitation, Journal of Official Statistics, 19, pp. 1.16.*

Rat für Sozial- und Wirtschaftsdaten (2006). *Eine moderne Dateninfrastruktur für exzellente Forschung und Politikberatung. Bericht über die Arbeit des Rates für Sozial- und Wirtschaftsdaten (RatSWD) während der ersten Berufsperiode vom 31.12.2006,* Frontenhausen

Reiter, J and Drechsler, J. (2007). *Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality. IAB Discussion Paper, No.20/2007.*

Ronning, G. (2007). *Stochastische Überlagerung mit Hilfe der Mischungsverteilung, IAW-Diskussionspapier Nr. 30*

Roque, G. (2000). *Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. thesis, University of California, Riverside.*

Rubin, D. (1993). *Discussion: Statistical Disclosure Limitation, in: Journal of Official Statistics, 9(2), pp. 461-468.*

Statistisches Bundesamt (2007a). *Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007b). *Produktionserhebungen, Qualitätsbericht.*

Statistisches Bundesamt (2007c). *Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2007d). *Kostenstrukturerhebung im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden, Qualitätsbericht.*

Statistisches Bundesamt (2005a). *Statistik und Wissenschaft, Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Band 4.*

Statistisches Bundesamt (2005b). *Umsatzsteuerstatistik, Qualitätsbericht.*

Sturm, R. und Tümmler, T. (2006). *Das statistische Unternehmensregister - Entwicklungsstand und Perspektiven. In: Wirtschaft und Statistik 10/2006, 1021-1036.*

Yancey, W., Winkler, W. and Creezy, R. (2002). *Disclosure Risk Assessment in Perturbative Micro Data Protection, in*: Domingo-Ferrer, J. (Ed.): *Inference Control in Statistical Databases – From Theory to Practice,* Berlin, *pp. 135-152.*

Zühlke S., Zwick, M., Scharnhorst S., Wende, T. (2004). *The research data centres of the Federal Statistics Office and the statistical offices of the Länder, Schmollers Jahrbuch 124, 567-578.*