Schmollers Jahrbuch 126 (2006), 635 – 651
Duncker & Humblot, Berlin

European Data Watch

This section will offer descriptions as well as discussions of data sources that may be of interest to social scientists engaged in empirical research or teaching courses that include empirical investigations performed by students. The purpose is to describe the information in the data source, to give examples of questions tackled with the data and to tell how to access the data for research and teaching. We will start with data from German speaking countries that allow international comparative research. While most of the data will be at the micro level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) will be included, too. Suggestions for data sources to be described in future columns (or comments on past columns) should be send to: Joachim Wagner, University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@uni-lueneburg.de⟩.

Anonymising Business Micro Data – Results of a German Project

By Rainer Lenz, Martin Rosemann, Daniel Vorgrimler, and Roland Sturm

1. Introduction

In order to make business micro data of official statistics available to scientists, such data must be in line with disclosure control requirements. The statistical offices are obliged by law to ensure the confidentiality of micro data files of persons as well as of firms. For German statistics legislation, a data set is anonymous (as far as scientific uses are concerned) if the costs of identification exceed the benefits of identification. Those data bases are called scientific use files as such data can be provided exclusively to scientists. Costs and benefits depend on how "sure" a data intruder can be to reveal useful information.

In order to maintain the analytical validity of the data, they should be anonymised carefully. This is considered to be more difficult to achieve for busi-

ness micro data than for micro data on households and persons (Sturm, 2002). Compared with data on households and individuals, an anonymisation of business micro data is notably more difficult: Business surveys are based on essentially smaller sample universes than individual-related surveys so that the cell frequencies of individual groups are often also smaller. The distributions of quantitative variables are by far more heterogeneous, and dominating cases do occur. Compared with individual-related surveys, the sampling fractions of business surveys are generally much larger and with respect to some strata they even are complete counts. Besides, the number of units differs considerably for the individual business size classes. Due to the businesses' obligation to publish data, on the one hand, and to the opportunity to retrieve information from data bases against payment, on the other, an external who intends to assign micro data to the respective carrier has at his disposal a substantially larger and much better processed additional knowledge about businesses than he has about individuals or households. And finally, the advantage gained from knowing data of enterprises and local units is rated by far more highly than that gained from obtaining information about individual or household-related surveys.

The users of micro data were largely involved in the German project (carried out 2002 – 2005) on the anonymisation of business micro data. As a partner in the project, the Institute for Applied Economic Research (IAW) examined the analytical validity of a large number of data which had been anonymised for test purposes. Generally, the basic difficulty in this context is that the purposes for which users wish to use the data are not known a priori. There are manifold potential uses of enterprise or local unit data, ranging from simple descriptive evaluations to the application of complex econometric models. In our project, we focussed on three approaches for measuring the analytical validity of micro data (see chapter 3 for details).

For assessing the effectiveness of data protection, the choice of a realistic scenario of potential "data intruders" is essential. Statistical offices therefore assume real-life conditions (such as the possibility to use external commercial databases, data mining via the internet, etc.). The Statistics Law stipulates that German statistical offices may pass on data to scientists only if re-identification of the enterprises or local units concerned would require an excessive, i.e. unreasonable effort. Within the framework of the project, we therefore developed a concept to measure protectiveness.

2. Methods of Anonymisation

A broad variety of anonymisation methods is described in literature (Höhne, 2003; Ronning et al., 2005), while practical knowledge about their effects is limited. Findings often depend on the data files used and can hardly be generalised (Sturm, 2005).

Anonymisation methods may be subdivided into two groups: methods reducing the information, and more recent methods modifying the values of numerical data. When an anonymisation concept is developed, a mix of these two approaches often seems to be the best solution. Information-reducing methods such as the suppression of variables not important for the data users or e.g. the presentation of 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.

The most important methods for modifying data tested in the project were mico aggregation methods, additive and multiplicative noise, Latin hypercube sampling, resampling, post randomisation, data swapping and SAFE.[1]

This paper describes as illustration only the category of deterministic micro aggregation methods (Mateo-Sanz / Domingo-Ferrer, 1998). The basic idea of these methods is to form groups of similar objects and substitute the original values by the arithmetic mean of a group's objects. If deterministic micro aggregation is performed jointly for all numerical variables – consequently, the same groups are formed for different variables when determining the averages (MA_COM) –, it may be differentiated again by whether all variables are included into the measure of distance used to form groups or whether groups are only formed on the basis of one or several guiding variables. In contrast, an isolated micro aggregation of each numerical variable may be performed, too (MA_IND). 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 (MA_GR).

3. Measuring Analytical Validity

In general, anonymisation procedures must be judged by whether they can guarantee the anonymity of business and operational data without limiting their usefulness.

3.1 How to Assess the Analytical Validity of Anonymised Data

The analysis potential is limited on the one hand by the fact that certain analyses are excluded from the start on account of anonymisation procedures

[1] An overview of the methods is provided by Brand (2000), Ronning et al. (2005) and Höhne (2003). On the Latin hypercube sampling see Dandekar et al. (2001), on the resampling Gottschalk (2005), on the post randomisation Willenborg and de Waal (2001) and on the SAFE method Evers and Höhne (1999).

because either the issue in question cannot be analysed anymore or the method to be used and equivalent methods cannot be applied anymore. On the other, such limits result from anonymised data producing other analysis results than the original data. When anonymisation procedures are assessed which modify the data, the focus is on the second aspect. The suggestions made on this subject may be classified into three approaches: the statistic-oriented, the application-oriented and the theory-oriented approach (Ronning et al., 2005).

In the project, the effects of data-modifying anonymisation methods were first derived theoretically, where possible, both for descriptive statistical analyses and for linear and non-linear econometric models. Then the theoretical results were checked by means of simulation studies and exemplary analyses using project data. Based on these results, the data-modifying anonymisation methods studied were assessed fundamentally as to whether they were suitable for preparing scientific use files. The methods regarded as suitable were then used to prepare scientific use files of the project data. Additionally, maximum deviation thresholds were specified for major distribution measures (Ronning et al., 2005).

In the following, major findings are summarised concerning the effects which the group of deterministic micro aggregation procedures has on the analytical validity.[2]

3.2 Effects of Deterministic Micro Aggregation on the Analytical Validity

3.2.1 Deterministic Micro Aggregation and Descriptive Distribution Measures

Arithmetic means are generally preserved by micro aggregation procedures. However, the arithmetic means of sub-populations are preserved by micro aggregation only if the procedure is applied separately to each of the sub-populations studied.

For explaining the effects of deterministic micro aggregation procedures on the variance, it is helpful that the latter may generally be broken down into an external variance between the mean values of individual sub-groups and the internal variance within the sub-groups. As the individual values are substituted by the group average in micro aggregation, the internal variance of each group becomes zero, while the external variance between the group averages remains unchanged. Thus the variance is generally reduced by applying micro aggregation procedures.

[2] For deterministic micro aggregation procedures, groups are formed according to the similarity of the values. For information on stochastic micro aggregation procedures with group formation at random please refer also to Ronning et al. (2005).

It is obvious that the reduction of the variance is the smaller the more similar the values are within the groups. Therefore, separate deterministic micro aggregation (MA_IND) leads to the smallest deviations of variances and standard deviations in that group of methods. The size of the deviations in joint deterministic micro aggregation depends on the degree of correlation among the variables micro aggregated together. For high correlations, the results of joint micro aggregation approach those of separate micro aggregation, the same holds for grouped micro aggregation where only part of the variables is micro aggregated together. It can be shown that the variance distortion of deterministic micro aggregation asymptotically approaches zero (Schmid et al., 2005).

It remains to be mentioned that the value of Spearman's rank correlation coefficient is as a rule preserved in separate deterministic micro aggregation. Slight modifications may occur only due to a higher number of ties. In contrast, no general statement can be made about the effects micro aggregation procedures have on Bravais' / Pearson's correlation coefficient.

In the following, the distortions of distribution measures caused by different variants of deterministic micro aggregation are shown, using as an example the 1999 structure of costs survey (SCS) in manufacturing. In addition to the two variants MA_COM and MA_IND described above, two variants of grouped micro aggregation, MA_GR8 and MA_GR10, are generated. Here 8 and 11 groups of variables are formed, respectively, while it has been made sure that pairs of highly correlated variables are found in a common group.

The German Manufacturing Industry Structure of Costs Survey of the year 1999 is a projectable sample and includes some 18,000 enterprises with 20 or more employees. All enterprises with 500 or more employees or those in economic sectors with a low frequency are included. That means, a potential data intruder has knowledge about the participation of large enterprises in the survey. We look at the survey of the year 1999, which covers 33 numerical variables (among them total turnover, research and development and the number of employees) and two categorical variables, namely the branch of economic activity (abbreviated: NACE) and the regional key.

Table 1 shows how the essential univariate and multivariate distribution characteristics of the Structure of Costs Survey change through anonymisation. We will look at the mean relative error of the arithmetic means and variances, and the average absolute deviations of the correlations and rank correlations.

As was to be expected, the micro aggregation method MA_COM led to the highest faults in the variances. This is due to the fact that variation is eliminated within the groups because of averaging. The faults in the variance covariance matrix are moderate with these methods. The correlation coefficients are reduced insignificantly. The micro aggregation method MA_IND generates the lowest deviations for the variance covariance matrices and correlation

coefficients. Here, the deviation is also low for the variances and arithmetic averages. Finally, the method MA_IND shows low deviations for the correlation coefficients, arithmetic means and variances. A somewhat different picture emerges for the rank correlations. After the method MA_IND, the method MA_GR10 shows the second smallest change, followed by MA_GR8 and MA_COM.[3]

Table 1

Change of distribution characteristics due to anonymisation of the SCS

	Average (relative) deviations of		Average absolute deviations of	
	the arithmetic mean	the variances	the correlation	the rank correlation
	in %	in %	(x 100)	(x 100)
MA_COM	3.5	21.3	5.8	9.0
MA_GR8	1.8	16.7	3.5	2.4
MA_GR10	3.9	29.5	2.5	1.5
MA_IND	0.0	5.9	2.4	0.0

Source: Land Statistical Office of Berlin and IAW.

3.2.2 Deterministic Micro Aggregation and Econometric Models

Based on the theoretical considerations of the project, the Monte Carlo simulation experiments carried out and the empirical model calculations, micro aggregation procedures are assessed as follows from the viewpoint of econometrics:[4]

a) Joint micro aggregation (MA_COM)

- In linear econometric models, the estimators are unbiased, provided that the dependent variable is not considered in calculating the measure of distance for group formation in micro aggregation (Lechner / Pohlmeier, 2003).[5] If all variables are considered in calculating the measure of distance, the distortion is low due to the smaller influence of the dependent variable.

[3] Deviations of the arithmetic means are due to the storage variables, which were shown in the data record merely as turnover shares. When these variables were micro aggregated, however, this was done separately for numerator and denominator. Then the quotients were re-calculated.

[4] On the results in this section compare especially Rosemann (2005).

[5] On the subject of micro aggregation based on the dependent variable compare Schmid et al. (2005).

- If only the explanatory variables are micro aggregated (irrespective of the dependent variables), in linear models the estimators are always unbiased.
- The test statistics are biased but may be corrected if the level of aggregation is known. The bias is the lower the more similar the values are within a group.
- Estimators in non-linear models or in non-linear transformations within linear models are as a rule biased.

b) Partly joint (grouped) micro aggregation (MA_GR)

- In linear econometric models, the estimators are not unbiased or consistent. In model calculations there are much greater deviations than for separate micro aggregation.
- Estimators in non-linear econometric models or in non-linear transformations within linear models are not consistent or unbiased.
- The test statistics are distorted and cannot be corrected.

c) Separate micro aggregation (MA_IND)

- Estimators in linear econometric models are not unbiased, they are however consistent (Schmid, 2005).
- Estimators in non-linear econometric models or in non-linear transformations within linear models are not consistent or unbiased, but simulation studies and empirical research show that the estimated coefficients change only slightly.
- The test statistics are asymptotically unbiased in linear models. For the model calculations made in the project, there were only slight deviations from the original also in non-linear models.

For illustration purposes, the results of Monte Carlo simulations (1,000 observations, 1,000 replications) are presented. On the one hand, a linear model with three normally distributed regressors is estimated. First the model is estimated with the original data, then both the regressors and the dependent variable are anonymised with different variants of deterministic micro aggregation (the size of the group being 5) (Table 2).

If all variables are jointly micro aggregated in a distance-oriented manner on the basis of the regressor X_1, the result are unbiased parameter estimators. The *t*-values turn out too high due to micro aggregation because they are based on a biased estimation of the residual variance. The correct *t*-values may be determined by multiplying the test statistics shown here with the factor $f = \sqrt{\frac{M-K}{n-K}}$, M being the number of groups formed by micro aggregation. In this case, with 1,000 observations, $K = 4$ regressors (constant included) and $M = 200$ groups, the resulting correction factor is 0.44. After correcting the

Table 2

MC-simulations – linear model, different variants of deterministic micro aggregation, all variables anonymised, 1, 000 replications

	Original	MA_IND			MA_COM based on X_1			MA_COM based on Y		
			Difference not significant (5 % level of significance)			Difference not significant (5 % level of significance)			Difference no significant (5 % level of significance)	
			original value	theor. value		original value	theor. value		original value	theor. value
X_1	1.001	0.999	*	*	1.000	*	*	1.343		
(t-values)	(28.57)	(28.46)			(48.44)			(39.06)		
X_2	–0.999	–0.999	*	*	–1.001	*	*	–1.337		
(t-values)	(–30.91)	(–30.85)			(–31.09)			(–43.82)		
X_3	0.499	0.500	*	*	0.501	*	*	0.669		
(t-values)	(14.45)	(14.45)			(14.52)			(17.48)		
CKonst.	1.001	1.000	*	*	1.002	*		1.002	*	*
(t-values)	(31.66)	(31.55)			(71.14)			(61.63)		

t-values accordingly, the *t*-value obtained for the constant corresponds to the *t*-value in the original estimation. The other *t*-values will then be considerably below the original values.

As had been expected, deterministic micro aggregation based on the dependent variable produces biased parameter estimations. Here separate deterministic micro aggregation leads to nearly unbiased estimators.

On the other hand, a binary Probit model is estimated with three normally distributed regressors. Again, the estimation is made first with the original data, then the three regressors are anonymised by different variants of deterministic micro aggregation (Table 3).

It turns out that in the non-linear Probit model joint deterministic micro aggregation (MA_COM) now produces strongly biased estimators, while good results (almost unbiased) can be obtained also in the non-linear model with separate deterministic micro aggregation (MA_IND).

Table 3

MC simulation – Probit model, different variants of deterministic micro aggregation, part of the regressors micro aggregated, 1,000 replications

	Original	MA_IND			MA_COM based on X_1		
			Difference not significant (5 % level of significance)			Difference not significant (5 % level of significance)	
			oiginal value	theor. value		original value	theor. value
X_1	1.009	1.007	*		0.730		
(*t*-values)	(12.88)	(12.87)			(9.67)		
X_2	–1.008	–1.008	*		–0.727		
(*t*-values)	(–13.51)	(–13.49)			(–6.48)		
X_3	0.504	0.505	*		0.356		
(*t*-values)	(7.77)	(7.78)			(3.05)		
Const.	1.005	1.009	*		0.7245		
(*t*-values)	(14.91)	(14.94)			(14.41)		

Consequently, separate deterministic micro aggregation (MA_IND) may be regarded as the only variant of the deterministic micro aggregation procedures which is suited for anonymisation given the multiple uses to be assumed for scientific use files. In the "Handbuch zur Anonymisierung wirtschaftsstatis-

tischer Mikrodaten" (Ronning et al., 2005), this procedure is recommended for anonymising micro data in addition to multiplicative stochastic noise.

4. Measuring the Disclosure Risk

We measured empirical disclosure risks by exposing our data which had been anonymised for test purposes to simulated investigations in a laboratory context. To this end, we thoroughly examined the additional information on enterprises and local units a potential data intruder might have and also the quality of such information. Then we tried to make allocations based on sophisticated matching algorithms (Lenz, 2005). Unlike a real data intruder, we could determine the hit rates in our laboratory context. Since we know the real identity of the anonymised data files, we were also in a position to specify the usefulness of correctly allocated information by determining their deviation from the original data. In this way we could specify a quantitative measure for the extent of data protection.

4.1 Simulation of a Data Attack

In this section we discuss the concepts of additional knowledge and the most important scenarios of data attack (Vorgrimler and Lenz 2003).

4.1.1 Knowledge Required by Potential Data Intruders

In order to re-identify a statistical unit (e.g. a specific enterprise), several assumptions concerning the data intruder are necessary for successful attempts, see also (Brand et al., 1999):

- Additional knowledge about the object (in our case in the form of an external database and knowledge obtained by internet research)
- Knowledge about the participation of the firm in the target survey (response knowledge)
- Key variables contained in both target and external data (making a unique assignment possible)

Moreover, the data intruder must be personally convinced of the correctness of the assignment.

4.1.2 Scenarios of Data Attack

In (Elliott / Dale, 1999) several scenarios of data attack are mentioned, two of them are the database cross match, as it is called, and the match for a single individual.

Within a database cross match a data intruder matches an external database with the confidential target data. In order to enhance his external data, he tries to assign as many true pairs of records as possible.

In order to simulate a database cross match we generate a distance measure covering all common variables (so-called key variables) of the records in the two databases. As in a real data attack scenario data intruders tend to prefer a few selected variables, supposed to include less deviations from the original data to other, less reliable variables, it is left to the user to assign concrete weights w_i to variables *i,* although, for the sake of simplicity, standardised weight intervals of *[0, 1]* were laid down.

The objective is now to make assignments of records on the basis of the previously calculated distances. For that purpose, we minimize the sum of distances for all assignments to be made (total deviation). For the purpose of comparison, we use an algorithm firstly presented in (Lenz, 2003) and developed further in (Lenz, 2005).

The intention behind a single individual match is to gain information about a specific target individual. The data intruder collects information about the individual searched for, using several sources of information. For instance, he can generate additional information by commercial databases and generally accessible information (e.g. annual reports of enterprises). The collected information is then used to re-identify the target individual in the anonymised data base in order to get further information about it.

4.1.3 Combination of Scenarios

In order to adequately evaluate the protection effect of an anonymisation method, both scenarios of data attack have to be taken into account. Let $R_{SIM}(u)$ denote the estimated re-identification risk associated with a single individual match applied to some unit u and $R_{DCM}(u)$ denote the corresponding estimator for the re-identification risk associated with a database cross match. Then, the re-identification risk $R(u)$ can be estimated by the maximum of both estimators, $R(u) := \max\{R_{SIM}(u), R_{DCM}(u)\}$.

The re-identification risk for some unit strongly depends on the data blocks within the data base to which it belongs. For instance, if an enterprise is assigned to a small branch of economic activity and / or to an upper employee size class, re-identification appears much easier than in the general case. Here, the re-identification risk $R_SIM(u)$ associated with a single individual match is expected to be higher than the corresponding one $R_{DCM}(u)$ associated with a database cross match. On the other hand, the database cross match is more effective than the single individual match in areas of data with high density, since in general there are many units with similar parameter values.

When by a data attack a set of additional knowledge was successfully assigned to an anonymised data set, all target variables which are contained in this data set were revealed. The benefit of a successful assignment hence arises from the "useful" information which a data intruder can reveal by successful identification. A piece of information revealed is only useful if the values revealed correspond to the „true values" or if the values revealed at least bear some similarity to the true values. If the deviation (between the value revealed and the true value) exceeds a certain level, a data intruder will not benefit from the information revealed. In our case, deviation is defined as the relative difference between the disseminated value and the true value of a variable.

This means that individual data will fulfil the criterion of being "anonymous" if the correctly assigned data set provides mainly useless information (the value revealed is outside a "deviation threshold" of the "true value"). It is a task of the statistical office to specify this deviation threshold. In the following examples, the deviation threshold has been set at 0.1 (that is, a value is considered to provide useful information for a data intruder if its relative difference from the true value is less than 10 percent) and the risk of revealing useful information is called *disclosure risk.*

4.2 Application to Real World Examples

In this chapter we describe how the above concepts can be applied to the German Turnover Tax Statistics 2000 (TTS) and the German Structure of Costs Survey 1999 (SCS). A more detailed description of the results can be found in (Lenz et al., 2005b).

4.2.1 German Turnover Tax Statistics

Turnover tax statistics are based on an evaluation of monthly and quarterly advance turnover tax returns to be provided by entrepreneurs whose turnover exceeds € 16,617 in the year 2000 and whose tax amounts to more than € 511 per annum. Nearly all economic branches are presented in the survey. The evaluation of the year 2000 contains almost 3 million records.

We consider three ways to anonymise the TTS, namely the formal anonymisation (FORMAL) and the previously described MA_IND and MA_COM.

Database cross match

The key variables of the TTS and the external data (additional knowledge) are branch of economic activity (NACE), total turnover, legal status and the regional key.

For laboratory data attack experiments we built a data base of external data containing nearly 9,300 enterprises with 20 or more employees, classified within NACE codes 10 – 37 (manufacturing industry). The corresponding subset of the target data contains nearly 37,000 enterprises. The following table 4 contains the results obtained by blocking data using the two-digit NACE code.

Table 4

Matching TTS: disclosure risk distributed among employee size classes

TTS	Total	Employee size class*					
		1	2	3	4	5	6
FORMAL	35.4	31.6	31.5	39.5	54.2	61.5	80.0
MA_IND	35.0	31.3	31.2	39.1	51.2	42.4	65.2
MA_COM	2.9	1.2	1.8	4.2	7.2	7.6	4.7

* 1 = less than 25; 2 = 25 – 100; 3 = 100 – 1 000; 4 = 1 000 – 5 000; 5 = 5 000 – 15 000; 6 = more than 15 000.

Obviously, variant MA_IND provides less protection than the other. The great deviations between the two data sources are more decisive for this phenomenon than the slight (almost negligible) modifications to the TTS. While only about 1 % of the enterprises have been classified differently with regard to regional information, nearly 25 % of the enterprises covered by the German turnover tax statistics have been assigned to another branch of economic activity than their respective records of external data. *Total turnover* figures match relatively well. Only some 18.8% of the enterprises show deviations of more than 10% between both data sources.

As had to be expected in the authors' opinion, the variant MA_COM produces safe micro data. On the other hand, this variant is connected with an unbearable abatement of statistical analysis. All in all, the disclosure grow (obtained by involving the concept of useful information) fall as the enterprise size increases.

Match for a single individual

We repeated the single individual match for 15 enterprises with the target data set being only formally anonymised. The key variables were the regional key, branch of economic activity, legal status and total turnover of the years 1999 and 2000. Using these key variables, only 6 out of 15 enterprises could be re-identified.

Hence, the results are in accordance with the database cross match, where the influence of divergencies between both surveys (irrespective of the method

of anonymisation decided for) were the main reason for unsuccessful attempts. But we can also observe that in contrast to other statistics (like the German Structure of Costs Survey SCS) the structure of the German turnover tax statistics does not offer a data intruder more key variables within a single match scenario than in the scenario of a database cross match. Therefore, the risk of re-identification of a specific enterprise with respect to a single match scenario is not higher than the risk regarding a database cross match.

4.2.2 German Structure of Costs Survey

The following results refer to the German Structure of Costs Survey, which was described in subsection 3.1.

Database cross match

The key variables of the SCS and the external data (additional knowledge) are the branch of economic activity reduced to two digits, the regional key, total turnover and the number of employees.

In our laboratory experiments, the external data contain nearly 9,400 enterprises with 20 or more employees, classified to NACE codes 10 – 37 (manufacturing industry).

We carried out database cross matches with four different versions of micro aggregation of the numerical variables *total turnover* and *number of employees*. The results obtained are shown in table 5.

As had to be expected, the ratio of correct assignments increases with the size class. Although it is normal that for larger enterprises the micro aggregation procedures cause more pronounced changes in the variables, the column on the right of table 5 shows a notably high risk of re-identification and disclosure for enterprises with at least 1,000 employees.

Match for a single individual

We carried out the single individual match for 41 enterprises with the target data set being only formally anonymised. In general, the key variables were the same as in the previous subsection. In some instances, the variables *total revenue, research and development investments* (yes or no), *trade activity* (yes or no) appeared as further key variables. With these keys, 19 of the 41 enterprises searched for could be re-identified. Only one enterprise could be re-identified among the 15 enterprises with less than 250 employees. On the other hand, among the larger enterprises a total of 18 out of 26 could be re-identified.

Table 5

Matching SCS: disclosure risk distributed among employee size classes

SCS	Total	Employee size class*					
		1	2	3	4	5	6
FORMAL	24.4	15.6	19.0	26.5	36.1	41.8	44.9
MA_IND	24.2	5.5	18.9	26.4	35.8	41.4	43.8
MA_GR10	19.8	12.8	16.9	21.5	26.1	29.7	24.1
MA_GR8	10.8	7.7	8.9	12.5	14.6	18.4	16.3
MA_COM	1.1	0.7	0.4	0.8	1.5	2.1	2.5

* 1 = 20 – 49; 2 = 50 – 99; 3 = 100 – 249; 4 = 250 – 499; 5 = 500 – 999; 6 = more than 999.

5. Conclusion

As was hoped, the results of the project will allow us to provide scientific users with wide access to business micro data. Besides, we realised that there are far reaching reservations among the scientific community about the application of data-perturbing anonymisation measures. Users seemingly prefer a clear reduction of information (e.g. suppressed regional information) to modifications in numerical values (e.g. turnover) (Sturm, 2005). Therefore, the application of data-modifying procedures should largely be restricted. If data-modifying procedures appear inevitable, a mix of information-reducing and data-modifying procedures is recommendable. Of the data-modifying procedures, the methods which may be used are in particular multiplicative stochastic noise and the micro aggregation variant treating individual variables separately.

Several stocks of anonymised data were made available in the course of project implementation. They included data on cost structures in industry and on structures in retail trade, and also turnover tax data.

Detailed descriptions of the scientific use files can be found in (Lenz et al., 2005a), (Vorgrimler et al., 2005) and (Scheffler, 2005). Recently, similar approaches were made in order to anonymise further business statistics like the German data of the Continuing Vocational Training Survey1999 and the German Structure of Earnings Survey 2001.

Now, the anonymisation of data from many other surveys has become comparatively easy as each time the most promising anonymisation strategy can be selected from our project-related anonymisation approaches. Its protective effects can be determined with the help of our measuring tools and a catalogue of meta information be prepared for evaluating the analytical validity. In the meta data, the procedures applied are made transparent to the user in the first place. The findings suggest that the anonymisation process will always have to be tailored to the specific data set (Sturm, 2005).

References

Brand, R. / *Bender,* S. / *Kohaut,* S. (1999): Possibilities for the creation of a scientific-use-file for the IAB-Establishment-Panel, Proceedings of the Joint Eurostat / UN-ECE Work Session on Statistical Data Protection, Thessaloniki, 57 – 74.

Dandekar, R. / *Cohen,* M. / *Kirkendall,* N. (2001): Applicability of Latin Hypercube Sampling to Create Multivariate Synthetic Micro Data, Proceedings of ETK-NTTS, Eurostat, Luxemburg, 839 – 847.

Elliot, M. / *Dale,* A. (1999): Scenarios of attack: the data intruder's perspective on statistical disclosure risk, Netherlands Official Statistics, 6 – 10.

Evers, K. / *Höhne,* J. (1999): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten, Spektrum der Bundesstatistik 14, Wiesbaden, 136 – 147.

Gottschalk, S. (2005): Unternehmensdaten zwischen Datenschutz und Analysepotenzial, NOMOS-Verlag.

Höhne, J. (2003): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten (German), Forum der Bundesstatistik 42, Wiesbaden, 69 – 94.

Lechner, S. / *Pohlmeier,* W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten (German), Forum der Bundesstatistik 42, Wiesbaden, 69 – 94.

Lenz, R. (2003): A graph theoretical approach to record linkage, Monographs of Official Statistics – Research in Official Statistics, Luxembourg, 324 – 334.

Lenz, R. (2005): Measuring the disclosure protection of micro aggregated business micro data – An analysis taking the example of German Structure of Costs Survey, to appear, Journal of Official Statistics 22(3), Sweden.

Lenz, R. / *Vorgrimler,* D. / *Rosemann,* M. (2005a): Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe (German), Wirtschaft und Statistik 2, 91 – 96.

Lenz, R. / *Vorgrimler,* D. / *Scheffler,* M. (2005b). A standard for the release of micro data, Monographs of Official Statistics (2006 edition), paper presented at the Joint UN / ECE work session on statistical data confidentiality, Geneva, 9 – 11, 197 – 206.

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

Ronning, G. / *Sturm,* R. / *Höhne,* J. / *Lenz,* R. / *Rosemann,* M. / *Scheffler,* M. / *Vorgrimler,* D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten (German), Statistik und Wissenschaft 4.

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

Scheffler, M. (2005): Ein Scientific-Use-File der Einzelhandelsstatistik 1999 (German), Wirtschaft und Statistik 3, 197 – 200.

Schmid, M. / *Schneeweiß,* H. / *Küchenhoff,* H. (2005): Consistent Estimation of a Simple Linear Model Under Microaggregation, SFB Discussion Paper No. 415.

Schmid, M. (2005): Estimation of a Linear Model under Microaggregation by Individual Ranking, SFB Discussion Paper No. 453.

Sturm, R. (2002): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten (German), Journal of the German Statistical Society 86, 468 – 477.

Sturm, R. (2003): Anonymization of Business Micro Data – A Glimpse of Work in Progress, Bulletin of the 54th Session of the International Statistical Institute (ISI), Berlin, Contributed Papers 2, 491 f.

Sturm, R. (2005): Anonymization of Business Statistics – Findings and Recommendations, Bulletin of the 55th Session of the International Statistical Institute (ISI), Sydney, Contributed Papers (forthcoming).

Vorgrimler, D. / *Dittrich,* S. / *Lenz,* R. / *Rosemann,* M. (2005): Wissenschaftliche Analy-sen anhand der Umsatzsteuerstatistik, Wirtschaftswissenschaftliches Studium 10, 327 – 332.

Vorgrimler, D. / *Lenz,* R. (2003): Disclosure risk of anonymized business micro data files – Illustrated with empirical key variables, Bulletin of the 54th International Statistical Institute (ISI), book 2, Berlin, 594 – 595.

Willenborg, L. / *de Waal,* T. (2001): Elements of Statistical Disclosure Control. Lecture Notes in Statistics 155.