FDZ-Arbeitspapier
Nr.7

Rainer Lenz
Daniel Vorgrimler
Michael Scheffler

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

A standard for the release of microdata

2006

FDZ-Arbeitspapier
Nr.7

Rainer Lenz
Daniel Vorgrimler
Michael Scheffler

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

A standard for the release of microdata

2006

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Bundesamt

Fachliche Informationen
zu dieser Veröffentlichung:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 06 11 / 75 42 20
Fax: 06 11 / 72 40 00
forschungsdatenzentrum@destatis.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum
Tel.: 06 11 / 75 42 20
Fax: 06 11 / 72 40 00
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 / 9449 41 47
Fax: 0211 / 9449 40 77
forschungsdatenzentrum@lds.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Juni 2006

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

A standard for the release of microdata

Rainer Lenz, Daniel Vorgrimler, Michael Scheffler[1]

Abstract

Statistical Offices in Germany may provide microdata to the scientific community, if these data are sufficiently anonymized. We present a standard for evaluating the degree of protection of a confidential data file. In a first step distance based record linkage is used to re-identify statistical units of the confidential target data. Besides re-identification of the unit it is also important to look at the benefit to a potential data intruder in case he reveals information. The more the information in the data disseminated is disturbed the lower is the benefit a data intruder derives from re-identification. For this reason, in a second step the re-identified units are analyzed if they contribute benefit to potential data intruders.

The paper shows how the standard mentioned can be applied to real world examples, taking the German Turnover Tax Statistics (almost full survey, about 3 million units), the German Structure of Costs Survey (a sample containing about 18 000 units) and the German Retail Trade Statistics (a sample containing about 23500 units) as a basis. Recently, so called Scientific-Use-Files of these surveys have been made available for the scientific community.

1 Introduction

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

[1] Dr. Rainer Lenz, Dr. Daniel Vorgrimler, Michael Scheffler , Federal Statistical Office Germany

scientists. Costs and benefits depend on how "sure" a data intruder can be to reveal useful information. In practice, a data intruder faces several problems:

- divergence between additional knowledge and anonymized data set,
- lack of knowledge as to whether the target individual is covered by the data,
- uncertainty as to whether an assignment is correct,
- uncertainty about the quality of the data revealed.

While in the area of households and individuals the anonymization of microdata has been practised for several years, an anonymization of business microdata 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 to individual-related surveys, the sampling fractions of business surveys are generally much larger while with respect to some strata, they are even equal to complete counts. Besides, the number of units differs largely between 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 microdata 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 on enterprises and local units is rated by far more highly than that achieved from obtaining information about individual- or household-related surveys. Surveys of local units also include items which may be of interest to competing enterprises, such as information on investments. A rational data intruder will therefore accept higher expenses for deanonymization provided they are offset by the advantage gained from the information obtained.

2 Simulation of a data attack

In this chapter we discuss the concepts of additional knowledge and the most important scenarios of data attack.

2.1 Additional knowledge

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 organization 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, for which he seems to be asking the impossible in the case of simulating a database cross match described in subsection 2.2.1.

2.2 Scenarios of data attack

In Elliott and Dale (1999) several scenarios of data attack are mentioned, two of them are the so called database cross match and the match for a single individual (see also Vorgrimler and Lenz 2003).

2.2.1 Database cross match

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 a first phase, we generate a distance measure covering all common key variables of the records in the two databases. As in a real 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 of the second phase is 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 (2003a) and developed further in Lenz (2004).

2.2.2 Match for a single individual

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 order to get further information about it.

2.2.3 Combination of scenarios

In order to adequately evaluate the protection effect of an anonymization 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 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 stands above the single individual match in areas of data with high density, since in general there are many units with similar parameter values.

If by a data attack a set of additional knowledge was successfully assigned to an anonymized 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 a successful identification. An information revealed is only useful if the values revealed correspond to the "true values" or at least if the values revealed are similar to the true values to a certain extent. Some anonymization methods modify the values of the data so that the values of the data disseminated differ from the corresponding original ("true") values. Above a certain deviation (between the value revealed and the "true" value) a data intruder will not obtain a 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 fulfill 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 to 0.1 (that is, a value is considered to contribute useful information to 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*.

3 Application to real world examples

In this chapter we describe how the above-described concepts can be applied to the German Turnover Tax Statistics 2000 (TTS), the German Structure of Costs Survey 1999 (SCS) and the German Retail Trade Statistics 1999 (RTS).

3.1 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 in the year 2000 € 16,617 and whose tax amounts to over € 511 per annum. Also excluded are enterprises with activities which are generally non-taxable or where no tax burden accrues (e.g. established medical doctors and dentists without laboratory, public authorities). Nearly all economic branches are presented in the survey. The evaluation of the year 2000 contains almost 3 million records. The survey has been conducted annually since 1996 (until then, every two years). The Federal Statistical Office of Germany published the following selected survey characteristics in tables:

- Deliveries and other performances (= taxable and non-taxable turnover)
- Branch of economic activity
- Legal form
- Bases of turnover tax (deliveries and other performances, intra-community acquisitions, input tax by tax rates, etc.)

In this section, we consider four ways to anonymize the TTS. General descriptions of these and other anonymization methods - independent from some specific survey - can be found in Höhne (2003).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)

2. The second is the use of traditional methods (like truncation and coarsening) of anonymization. Since the German turnover tax statistics determine a rather large data set, an application of traditional methods could produce reasonable results concerning confidentiality. (Traditionally anonymized)
3. The third is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 21G)
4. The fourth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.1.1 Database cross match

For our purposes, the most important variables are:

- Branch of economic activity (NACE)
- Total turnover
- Legal status
- Regional key

The above variables are the key variables of the TTS and external data (additional knowledge). The external data contains nearly 9300 enterprises with 20 or more employees, classified within NACE codes 10 - 37 (manufacturing industry). The corresponding subset of the target data contains nearly 37000 enterprises. We carried out database cross matches with different anonymizations of the categorical variables. In the original data, the NACE code has four digits. Through truncation the NACE code is reduced to zero (in this case the data intruder possesses no information on the branch of economic activity), one, two and three digits, so that we obtain four non-trivial forms of the code. Furthermore, the legal status is re-coded. In the original data, the legal status has a range of eight values, after re-coding it is coarsed to four values.

The following table contains the results obtained by blocking data using the four levels of the NACE code, with the 0-digit cases indicating that the variable was left out of

consideration. That is, the data intruder does not have additional knowledge of the branch of economic activity.

Table 1.

Matching TTS: re-identification risk (disclosure risk) distributed to employee size classes

TTS	NACE	Total	Employee size class*					
			1	2	3	4	5	6
FORMAL	4 digits	40.1 (40.1)	35.3 (35.3)	35.7 (35.7)	45.7 (45.7)	54.9 (54.9)	57.7 (57.7)	70.0 (70.0)
	3 digits	40.1 (40.1)	35.5 (35.5)	36.1 (36.1)	45.1 (45.1)	52.6 (52.6)	65.4 (65.4)	60.0 (60.0)
	2 digits	35.4 (35.4)	31.6 (31.6)	31.5 (31.5)	39.5 (39.5)	54.2 (54.2)	61.5 (61.5)	80.0 (80.0)
	1 digit	21.0 (21.0)	17.9 (17.9)	18.5 (18.5)	23.1 (23.1)	42.5 (42.5)	34.6 (34.6)	40.0 (40.0)
	0 digits	13.6 (13.6)	11.5 (11.5)	11.9 (11.9)	14.7 (14.7)	28.9 (28.9)	42.3 (42.3)	40.0 (40.0)
MA 21G	4 digits	40.1 (39.6)	35.3 (34.9)	35.7 (35.3)	45.6 (45.1)	55.2 (53.0)	57.7 (39.9)	70.0 (56.1)
	3 digits	39.9 (39.5)	36.1 (35.8)	36.1 (35.7)	44.6 (44.2)	53.3 (52.8)	57.7 (39.7)	80.0 (65.6)
	2 digits	35.4 (35.0)	31.6 (31.3)	31.5 (31.2)	39.3 (39.1)	55.2 (51.2)	61.5 (42.4)	80.0 (65.2)
	1 digit	20.8 (20.6)	17.7 (17.5)	18.3 (18.1)	22.8 (22.6)	42.2 (40.5)	34.6 (23.7)	50.0 (41.6)
	0 digits	13.7 (13.6)	11.3 (11.2)	12.0 (11.9)	14.5 (14.3)	32.5 (31.2)	30.8 (21.3)	40.0 (33.7)

MA 1G	4 digits	27.9 (5.6)	21.4 (3.6)	21.8 (4.8)	35.0 (7.4)	53.9 (8.1)	65.4 (7.3)	60.0 (5.1)
	3 digits	23.4 (4.5)	15.8 (2.6)	17.3 (3.5)	30.1 (6.5)	52.6 (7.6)	73.1 (9.0)	80.0 (7.2)
	2 digits	14.4 (2.9)	6.9 (1.2)	9.5 (1.8)	18.9 (4.2)	46.8 (7.2)	69.2 (7.6)	60.0 (4.7)
	1 digit	5.4 (1.0)	2.1 (0.4)	3.3 (0.7)	7.0 (1.6)	21.2 (3.1)	34.6 (4.5)	30.0 (2.8)
	0 digits	2.6 (0.5)	0.9 (0.2)	1.5 (0.3)	3.4 (0.7)	11.7 (1.9)	26.9 (2.9)	30.0 (2.5)
Traditionally anonymized		30.0 (29.9)	26.7 (26.7)	27.0 (27.0)	34.0 (34.0)	41.2 (39.7)	19.2 (11.2)	20.0 (8.1)

* = 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, the weakest variant MA 21G provides lesser protection than the other variants of anonymization. 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 the 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 the 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 1G produces safe microdata. On the other hand, this variant is connected with an unbearable abatement of statistical properties. The matching results obtained by coarsening the NACE code to 3 or 4 digits are comparable. In the case of NACE 4 the increase in the number of enterprises protected due to deviations in both sources is compensated by the decrease in the re-identification risk in the case of NACE 3 due to larger blocks. An improved effect of protection is achieved by reducing the NACE code to 2 digits. Regarding the traditional method, it is observed in contrast to the other

methods that this method – roughly spoken - protects the larger insecure enterprises much better. All in all, the disclosure risks (obtained by involving the concept of useful information) are slowed down in line with an increasing growth of enterprises.

3.1.2 Match for a single individual

We repeated the single individual match for 15 enterprises with the target data set being only formally anonymized. The key variables were the regional key, the business classification, the legal status and the turnovers of the years 1999 and 2000 (note that the key variables were not available for the observations as a whole). 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 deviations in both surveys (irrespective of the method of anonymization 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.

3.2 German Structure of Costs Survey

The German structure of costs survey of the year 1999, limited to the manufacturing industry, is a projectable sample and includes a maximum of 18000 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 is, a potential data intruder has knowledge about the participation of large enterprises in the survey. We consider the survey of the year 1999, covering 33 numerical variables (among which are *Total turnover, Research and Development* and the *Number of employees*) and two categorical variables, namely the *Branch of economic activity* (abbreviated: NACE), broken down to

the 2-digit level, and the *Type of administrative district* (abbreviated: BBR9), which has 9 values depending on the degree of urbanisaton of the region considered.

In this section, we consider five ways to anonymize the SCS. A detailed description of these methods can be found in Lenz (2003b).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)
2. The second is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 30G)
3. In the third variant, the set of variables is textually divided into three-element groups. (MA 10G)
4. Grouping highly correlated variables leads to groups of size between two (smallest group) and twelve elements (largest group). (MA 8G)
5. The fifth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.2.1 Database cross match

For our purposes, the most important variables are:

- Branch of economic activity (NACE 2), reduced to two digits
- Type of administrative district (BBR9), containing 9 categories
- Total turnover
- Number of employees

The above variables are the key variables of the SCS and external data (additional knowledge). The external data contains nearly 9400 enterprises with 20 or more employees, classified within NACE codes 10 - 37 (manufacturing industry).

We carried out database cross matches with five different degrees of perturbation of the categorical variables *Total turnover* and *Number of employees*. The results obtained are shown in table 2.

Table 2.

Matching SCS: re-identification risk (disclosure risk) distributed to employee size classes

SCS	Total	Employee size class*					
		1	2	3	4	5	6
FORMAL	24.4 (24.4)	15.6 (15.6)	19.0 (19.0)	26.5 (26.5)	36.1 (36.1)	41.8 (41.8)	44.9 (44.9)
MA 30G	24.4 (24.2)	15.6 (15.5)	19.0 (18.9)	26.5 (26.4)	35.9 (35.8)	41.6 (41.4)	44.7 (43.8)
MA 10G	24.2 (19.8)	15.6 (12.8)	19.2 (16.9)	26.5 (21.5)	34.4 (26.1)	41.3 (29.7)	44.0 (24.1)
MA 8G	19.6 (10.8)	12.7 (7.7)	14.9 (8.9)	21.9 (12.5)	27.0 (14.6)	35.5 (18.4)	36.2 (16.3)
MA 1G	3.8 (1.1)	2.2 (0.7)	1.5 (0.4)	3.1 (0.8)	5.8 (1.5)	9.0 (2.1)	16.7 (2.5)

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

As to be expected, the frequency of correct assignments grows with the number of employees. 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 2 shows a notably high risk of re-identification and disclosure for enterprises with at least 1000 employees.

While the deviation amounting to about 24% for all enterprises in the *Branch of economic activity* is in line with the preceding section as are the slight deviations in the regional data of less than 2%, there are much more marked differences regarding *Total*

turnover. About 50% of the enterprises deviate from each other by more than 10% in the two data sources.

3.2.2 Match for a single individual

We repeated the single individual match for 41 enterprises, without consideration of commercial databases. 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.

3.3 German Retail Trade Statistics

The German Retail Trade Statistics of the year 1999 is a projectable sample containing about 23500 enterprises. In each branch of economic activity, the dominant enterprises have been included into the survey. The RTS consists of 33 numerical and 3 categorical variables. The results of this annual survey yield important information to economic-political problems concerning the structure, profitability and productivity of enterprises of this sector. In this section, we consider four ways to anonymize the RTS. A detailed description of these methods can be found in Scheffler (2005).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)
2. The second is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 31G)
3. The third was obtained by groupwise application of micro aggregation to 9 groups of numerical variables. (MA 9G)
4. The fourth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.3.1 Database cross match

In order to simulate database cross matches with the RTS, we generated additional knowledge containing about 12100 enterprises, classified within NACE codes 521 - 527 (retail trade) on a three-digit level. Hence, the key variables are

- Branch of economic activity (NACE 2), reduced to three digits
- Type of administrative district (BBR9), containing 9 categories
- Total turnover

Table 3 below contains the re-identification and disclosure risks associated with the four variants of anonymization distributed to employee size classes.

Table 3.

Matching RTS: re-identification risk (disclosure risk) distributed to employee size classes

SCS	Total	Employee size class*						
		1	2	3	4	5	6	7
FORMAL	22.2 (22)	20.9 (20.9)	23.7 (23.7)	29.3 (29.3)	27.4 (27.4)	39.6 (39.6)	25.0 (25.0)	48.1 (48.1)
MA 31G	22.0 (21.8)	20.9 (20.8)	23.6 (23.6)	28.1 (28.1)	29.3 (29.3)	38.4 (38.4)	30.0 (29.9)	45.1 (42.8)
MA 9G	3.1 (2.4)	3.2 (2.7)	3.1 (2.6)	6.0 (4.9)	10.4 (8.0)	17.2 (12.9)	13.1 (9.9)	30.2 (19.9)
MA 1G	2.1 (1.3)	2.0 (1.6)	2.2 (1.4)	3.5 (2.1)	4.8 (2.1)	7.6 (4.2)	9.1 (5.2)	24.8 (11.4)

* = 1-19 ; 2 = 20-49 ; 3 = 50-99; 4 = 100-249 ; 5 = 250-499 ; 6 = 500-999 , 7 = more than 999.

As was to be expected, the protection effect of the weakest variant of micro aggregation, MA 31G, is similar to the effect of formally anonymized data. For enterprises with 500-

999 employees, this method even has a disclosive impact. In accordance with the previous sections, the relative frequencies of correct assignments grow with the number of employees.

3.3.2 Match for a single individual

We repeated single individual matches for a sample of 20 enterprises drawn by the size class of enterprises with more than 999 employees. In several passes, the variable *Number of branch offices* turned out to be a key variable between additional knowledge (mainly generated by internet research) and the target enterprise.

At first, the matches were carried out using only the internet as additional knowledge. In doing so, 8 of the 20 enterprises searched for could be uniquely and correctly assigned to their corresponding target individuals (re-identified). In a second step, the matches were carried out using only the external database described in 3.3.1. Here, 11 of the 20 enterprises participated in the external survey, where 6 of them could be re-identified using the external data and 4 of them using the Internet.

Finally, the matches were carried out using both, internet and external database, as additional knowledge. In this simulation, 8 of the 11 enterprises searched for could be re-identified. This means an increase from 4 (Internet) over 6 (external database) to 8 re-identifications.

3.4 Scientific-Use-Files

For each of the above-described surveys a so called Scientific-Use-File has been generated, i.e. data available for scientific purposes. Since the TTS consists of many records (about 2.9 million) and less numerical variables (most of them strongly correlated with *Total turnover*, a strong emphasis was put on anonymization of categorical variables (essentially information reducing methods). Anonymizing the SCS, consisting of less records (about 18.000) and about 30 numerical variables, a stronger emphasis was put on numerical variables (data perturbing methods) as well as in the

case of the RTS. Detailed descriptions of the Scientific-Use-Files can be found in Lenz et al. (2005), Vorgrimler et al. (2005) and Scheffler (2005).

4 Conclusion

In this paper we examined the risk a data intruder must take into account when he conducts an identification attempt. Economic rationale suggests that if the risk to fail is too high, an intruder will refrain from an identification attempt and the data sets can be regarded as protected. The concepts have been applied to three different business surveys of German official statistics. Currently, similar approaches are made in order to anonymize further business statistics like the Continuing Vocational Training Survey 1999 and the Structure of Earnings Survey 2001.

This work was partially supported by the EU project IST-2000-25069, **C**omputational **A**spects of **S**tatistical **C**onfidentiality (CASC), and by the German national project De Facto Anonymization of Business Microdata.

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.

Domingo-Ferrer, J., Torra, V. (2003): *Record linkage methods for multidatabase data mining*. *Information Fusion in Data Mining*. Springer-Verlag, Berlin, 99-130.

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

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

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

Lenz, R. (2003b). *Disclosure of confidential information by means of multi-objective optimisation*. Proceedings of the Comparative Analysis of Enterprise Micro Data Conference (CAED), London, 2003.

Lenz, R. (2004). *Measuring the disclosure protection of micro aggregated business microdata – An analysis taking the example of German Structure of Costs Survey*. Appears in: Journal of Official Statistics, 2004.

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

Rosemann, M., Vorgrimler, D., Lenz, R. (2004). *Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten (German)*. Journal of the German Statistical Society, Vol. 88, 73-99.

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

Vorgrimler, D. (2003): *Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios (German)*. Forum der Bundesstatistik, 42, Wiesbaden, 40-58.

Vorgrimler, D., Dittrich, S., Lenz, R., Rosemann, M. (2005): *Wissenschaftliche Analysen mit Hilfe der amtlichen Umsatzsteuerstatistik (German)*. Wirtschaftswissenschaftliches Studium 10, 585-590.

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