FDZ-Arbeitspapier
Nr.4

Dr. Rainer Lenz
Dr. Daniel Vorgrimler

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Matching German turnover tax statistics

2005

FDZ-Arbeitspapier
Nr.4

Dr. Rainer Lenz
Dr. Daniel Vorgrimler

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Matching German turnover tax statistics

2005

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

Erscheinungsfolge: unregelmäßig
Erschienen im Mai 2005

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

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

Matching German turnover tax statistics

Rainer Lenz[1] and Daniel Vorgrimler[2]

[1]Federal Statistical Office of Germany,
Division IA, Questions of Principle,
Gustav-Stresemann-Ring 11,
65180 Wiesbaden, Germany
rainer.lenz@destatis.de

[2]Federal Statistical Office of Germany,
Division VID, Tax,
Gustav-Stresemann-Ring 11,
65180 Wiesbaden, Germany
daniel.vorgrimler@destatis.de

Abstract. In order to reveal confidential information, a data intruder has several alternatives. One of these is the so-called database cross match, where it is tried to assign as many records as possible - belonging to the same entity - of an external database (the intruder's additional knowledge) to the target data. In[1] a heuristic has been proposed for these purposes to be implemented in the next update of the software package μ-Argus. In the present paper we study the algorithm's output after application to the German turnover tax statistics as target data. Moreover, we give a detailed overview of the results obtained by different parameter settings of the algorithm, in particular we analyze the effects of the categorical key variables contained in both external and target data. The application gives an approach to generate anonymised confidential data to be disseminated to the scientifical community.

1 Introduction

In the last decade, the national statistical offices and other data providing institutions observed an increasing demand of researchers to use business microdata. Empirical economic researchers deplore the lack of access to business microdata of German official statistics: The analysis potential of surveys conducted by the statistical offices at enterprises and local units is only partially exploited. Alternative solutions applied by the economic research institutes, e.g. in the form of own surveys, are expensive and often of a less quality. Moreover, they put an additional burden on the enterprises surveyed, which may

[1] Lenz, R.: A graph theoretical approach to record linkage.

still increase the widespread weariness concerning statistics. Besides, this practice is unsatisfactory from the viewpoint of making optimum use of available public resources.[2]

In 1987, the Law on Statistics for Federal Purposes created the so-called "privilege of science", allowing scientists and researchers access to so-called de facto anonymized microdata. A data set was defined to be de facto anonymous if the costs of re-identification exceeded the benefit of re-identification. By then, only completely anonymized microdata files could be provided to scientists. That is, the statistical offices had to make sure that data intruders had no chance to deanonymize data in order to gain information about specific organizations. Moreover, the statistical offices have ethical reasons to protect respondents and they must be fully trustworthy to be able to gather data from respondents.[3]

While in the area of households and individuals the anonymization of microdata has been practised for several years and the transmission of de facto anonymized material has proven to be a good way, 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.

The software package μ-Argus offers a variety of methods for producing safe microdata files.[4] Users of the package are offered a choice between methods, or they must select suitable parameters when applying a method to a data set. Regarding a particular anonymization method, two aspects have to be taken into account: information loss and data protection. An anonymization method may perform well with respect to information loss, but it may not protect the data sufficiently according to the user's requirements. In[5] an assignment heuristic has been proposed to be implemented in the next update of the software package μ-Argus. Though the heuristic suggested has a bad worst-case performance – indeed, the results obtained can be arbitrarily bad from a theoretical point of view – it has held its own among other approaches (par-

[2] Sturm, R., Wiesbaden (2001), 1-2

[3] see also: Wende, T., Zwick, M., Geneva (2003), 141-146

[4] see: Jaro, M. A., (1989), 415-435

[5] Lenz, R., Luxembourg (2003)

ticularly, methods of multi objective optimization) and has turned out to be appropriate regarding result and complexity analyses.[6]

2 Preliminaries

In order to re-identify a statistical unit (e.g. an enterprise), several assumptions concerning the data intruder are necessary for successful attempts:[7]

- Additional knowledge about the object (in our case in the form of an external database)
- Knowledge about the participation of the organization in the target survey (response knowledge)
- Common variables contained in both target and external data (making a unique assignment possible)

Moreover, the data intruder must be personally convinced about the correctness of the assignment, for which he seems to be asking the impossible in the case of simulating a database cross match.

In the following, let $A=\{a_1,\ldots,a_m\}$ and $B=\{b_1,\ldots,b_n\}$ be two sets of records sharing a nonempty set of variables, the so-called key variables. In our case, the key variables are the common variables of the target data (confidential business microdata) and the external data (additional knowledge of the data intruder), which in the following are denoted by $(v_1,\ldots,v_k)$. We partition the set of key variables into two classes of variables, namely categorical and numerical variables.

The categorical variables are divided into nominal variables (there is no ranking of the categories) and ordinal variables (the categories have a natural ordering where differences between categories are meaningless). Numerical variables are defined as being discrete or continuous variables where the difference between values has meaning, e.g. "height" and "weight" of a person or in our case "number of employees" and "turnover" of an enterprise. Note that every numerical variable can be transformed into a categorical variable by agglutination of the variable-range into intervals. For instance, turnover or employee size classes.

Let the record pair (a,b) be a candidate for a possible assignment. Then it is required that both records share some specified variables. In the following these variables are called blocking variables, since they divide the data as a whole into disjoint blocks. For a reasonable blocking of data, it is strongly recommended that those categorical variables are used whose values are as good as possible in characterizing the records (enterprises) under consideration and whose expected reporting error is as small as possible.

In addition, the potential data intruder will decide in favour of variables which have been left out of consideration during the application of the anonymization procedure.

[6] Lenz, R., London (2003)

[7] see also: Brand, R.; Bender, S.; Kohaut, S., Thessaloniki (1999), 57-74

3 Matching algorithm

The aim of the data intruder is to decide whether or not a pair *(a,b)* in *A×B* of records belongs to the same underlying individual. In a non-technical way, the concept of matching may be introduced as a way of bringing together pieces of information in pairs from two records taken from different data sources. For this, a reasonable concept of similarity is necessary. Roughly spoken, the greatest possible similarity between two records turns into identity if the considered records correspond with regard to all key variables. In the case of small deviations of the key variables, two objects are felt to be strongly related, so that the matching result essentially depends on the concept of similarity. The following types of distances were chosen for implementation:

Let *k* be the number of key variables. For a numerical key variable v_i and a record pair *(a,b)* the *i*-th component distance is defined as $d_i(a,b) = (a^{(i)} - b^{(i)})^2$, where $a=(a^{(1)},\ldots a^{(k)})$ and $b=(b^{(1)},\ldots b^{(k)})$. Further, we set

$$d_i(a,b) := \begin{cases} 0, & \text{if } a^{(i)} = b^{(i)} \\ 1 & \text{otherwise} \end{cases}$$

in the case of a nominal variable v_i and

$$d_i(a,b) := \frac{\left| \{c_j \mid \min(a^{(i)}, b^{(i)}) \le_i c_j <_i \max(a^{(i)}, b^{(i)})\} \right|}{r}$$

in the case of an ordinal variable v_i , besides blocking variables, where $c_1 <_i c_2 <_i \ldots\ldots <_i c_r$ is the ordered range of the variable v_i. In order to avoid scaling problems, the calculated component distances are standardized using the well-known *max-min*-standardization

$$\tilde{d}_r(a,b) := \frac{d_r(a,b) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}{\max_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}.$$

Let *NV* be the index set of the numerical and *CV* the index set of the categorical variables, respectively. The overall distance is then defined as a weighted sum

$$d(a,b) := \sum_{i\in NV} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in CV} \lambda_i \tilde{d}_i(a,b)$$

$$= \sum_{i\in NV} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in Cord} \lambda_i \tilde{d}_i(a,b) + \sum_{i\in Cnom} \lambda_i \tilde{d}_i(a,b),$$

where the distances associated with categorical variables are split by the contributions of ordinal (Cord) and nominal (Cnom) variables. Note that it is also possible to embed the blocking process into the distance calculation.[8] We are now able to formulate the essential steps of the matching algorithm as sketched below.

1.) INPUT: $T=(\tau_1,\ldots,\tau_k)$ vector of variable types,

$\Lambda=(\lambda_1,\ldots,\lambda_k)$ vector of variable weights,

$A=\{a_1,\ldots,a_m\}$ and $B=\{b_1,\ldots,b_n\}$ as described in section 2.

T is a k-tuple of integers with the following properties:

$\tau_i=0$ *if and only if* v_i is the identifier variable,

$\tau_i=1$ *iff* v_i is a blocking variable,

$\tau_i=2$ *iff* v_i is an ordinal variable and

$\tau_i=3$ *iff* v_i is a nominal variable.

Otherwise, v_i is assumed to be a numerical variable.

Λ is a k-tuple of real numbers compatible with T. That is, $\tau_i=0$ implies $\lambda_i=0$ and $\tau_i=1$ implies $\lambda_i=1$. The remaining weights are standardized appropriately, so that $\Sigma^k_{i=1}\ \lambda_i=1$ holds.

2.) Calculation of the summed-up distances $d_{ij}:=d(a_i,b_j)$ for $i=1,\ldots,m$ and $j=1,\ldots,n$.

3.) Listing of the distances in an ascending order (list L).

4.) While L is nonempty do

Consider the first element d_{ij} of L and assign (a_i,b_j).

Delete all elements d_{rs}, where $r=i$ or $s=j$.

5.) OUTPUT: Relative frequency of *true* and *useful true* matches.

[8] Lenz, R., Stockholm (2003)

Note that the computational amount in total of the above algorithm takes $O(knm)$. In most cases, it holds $k \ll n$ and $k \ll n$, so that we may neglect the factor k within the analysis of complexity. Hence, the complexity can be rated as $O(max\{m,n\}^2)$. Experience has shown that more elaborated algorithms having complexity of at least $O(max\{m,n\}^3)$ would not be tolerable for our purposes.

4 Application to real data

In this chapter the foregoing algorithm is applied to the German turnover tax statistics in order to estimate the re-identification risk associated with several variants of anonymization.

4.1 The target data used and additional knowledge

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 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, insurance agents, agricultural holdings).

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.)

Nearly all economic branches are presented in the survey. Appendix 1 contains a more detailed description of the variables available in the survey. For our purposes, the most important variables are:

- Branch of economic activity (NACE)
- Turnover
- Legal form
- Regional key

The above variables are the key variables of the target and external data. 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 the 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.

In this paper, we consider four ways to anonymize the target data:

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.
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.
3. The third is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group (see section 4.1.3).
4. The fourth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together (see also 4.1.3).

4.1.1 Deviation between the surveys

In this paper we analyze two types of data protection: On the one hand, protection by anonymization of the target database, and on the other hand, protection by the natural deviation between target database and external database ("deviations of the data"). Tables 1 and 2 demonstrate the deviation between both databases for the variables turnover and branch of economic activity. For instance, table 1 shows that nearly 30% (20%) of all enterprises in the external database differ for the variable turnover by more than 5 % (10%) from their corresponding units in the target database. Table 2 shows that only 56% of the enterprises have the same 4-digit classification in both target and external database.

Table 1. Deviations of the data I: Number of enterprises with a deviation regarding turnover in both surveys

Deviation less than...		1 %	5 %	10 %	25 %	50 %	Total
Frequency of enterprises	Absolute	3 546	6 541	7 539	8 395	8 706	9 283
	Relative	38.2	70.5	81.2	90.4	93.8	100

Table 2. Deviations of the data II: Number of enterprises with the same business classification in both surveys (with respect to different levels)

Digits of Classification		4	3	2	1	Total*
Frequency of enterprises	Absolute	5 206	5 917	7.007	7 823	9 283
	Relative	56.08	63.74	75.48	84.27	100

** without deviations regarding the regional key*

Note that around 2 % of the enterprises also have deviations regarding the regional key. For this reason, we cannot match all enterprises of the external database with their corresponding enterprises in the target database. These enterprises are protected through the natural deviation of the data. Table 3 contains the frequency of protected and unprotected enterprises with regard to the branch of economic activity. For instance, if the matchings were carried out using the NACE code on a four-digit level, 45% of the enterprises would a priori be protected by the deviations between both databases (37% using three digits and so on).

Table 3. Number of protected and unprotected enterprises by deviation concerning the NACE code between both surveys

Digits of Classification		4	3	2	1	0	Total*
Frequency of unprotected enterprises	Absolute	5 120	5 812	6 878	7 673	9 097	9 283
	Relative	55.2	62.6	74.1	82.7	98.0	100
Frequency of protected enterprises	Absolute	4 163	3 471	2 405	1 610	186	0
	Relative	44.8	37.4	25.9	17.3	2.0	0

** without deviations regarding the regional key*

4.1.2 Traditional anonymization

One of the target data anonymizations considered was obtained by traditional methods (for details of the anonymization see appendix 2). The following steps were carried out:

- Truncation of the NACE code (less than or equal to € 500 mn in turnover to two digits, above € 500 mn to one digit)
- Quorum of 3500 enterprises within one NACE group
- Re-coding of the legal status (four values instead of eight)
- Topcoding of the turnover using a lower and an upper threshold (€ 500 mn and € 1 bn)
- Rounding of turnover (applied to values less than or equal to € 500 mn)

4.1.3 Anonymization by micro aggregation

Having been derived from earlier investigations of the German structure of costs survey, some variants of the method of (multidimensional) micro aggregation have proven to generate useful and inference-valid material for applied econometricians.[9]

Micro aggregation firstly divides the set of variables into groups, where in most cases highly correlated variables are classified together. Within a group, the variables are standardized and summed up for each record, so that the records can be sorted by so-called *Z*-scores. Afterwards, for a pregiven number *k* (in our case *k=3*), the records with the greatest and the smallest *Z*-scores are put together with their *k-1* nearest neighbours w.r.t. euclidean distance (i.e. multidimensional micro aggregation) and their values are averaged.

In order to get an upper and a lower threshold for the re-identification risk connected with the method of micro aggregation, we apply the weakest and the strongest variants of micro aggregation to the target data, namely MA21G (where every numerical variable, in particular the key variable "turnover", is treated independently) and MA1G (where all numerical variables are grouped together). The latter anonymization generates triples of records, which coincide in all numerical variables and hence can only be separated by differences in the categorical variables.

Regarding the MA21G, there are very small modifications of the original values. More than 99.9 % of the modified values deviate from their original values by less than 5 %. This holds also for nearly 90 % of all enterprises with more than 500 employees, which are in general known to be very much at risk. Therefore, the risk connected with the disclosure of useless information is very small.

Regarding the MA1G, the modifications of the original values are much stronger, so that the probability that a disclosed value deviates by more than 10 % from its corresponding original value is greater than 60 %. The same holds for a restriction to enterprises with more than 500 employees.

4.2 Database cross match

We consider two types of hit rates. On the one hand, successful attempts in ratio to all enterprise and on the other hand, the adjusted hit rate, defined as the frequency of successful attempts in ratio to the unprotected enterprises as a whole (see table 3). The candidates available for blocking variables are the NACE (with its four different levels), the regional key and the legal status (with its two different variations).

4.2.1 Results depending on the level of NACE

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.

[9] Rosemann, M., Wiesbaden (2003), 154-183

Table 4. Re-identification: frequency of successful attempts in absolute terms (percentage)

Target Data	NACE	Total	Employee size class*					
			1	2	3	4	5	6
Formally anonymized	4 digits	3 726 *(40.1)*	188 *(35.3)*	1 764 *(35.7)*	1 583 *(45.7)*	169 *(54.9)*	15 *(57.7)*	7 *(70)*
	3 digits	3 720 *(40.1)*	189 *(35.5)*	1 781 *(36.1)*	1 565 *(45.1)*	162 *(52.6)*	17 *(65.4)*	6 *(60)*
	2 digits	3 287 *(35.4)*	168 *(31.6)*	1 557 *(31.5)*	1 371 *(39.5)*	167 *(54.2)*	16 *(61.5)*	8 *(80)*
	1 digit	1 951 *(21.0)*	95 *(17.86)*	912 *(18.47)*	800 *(23.1)*	131 *(42.5)*	9 *(34.6)*	4 *(40)*
	0 digits	1 259 *(13.6)*	61 *(11.5)*	586 *(11.9)*	508 *(14.7)*	89 *(28.9)*	11 *(42.3)*	4 *(40)*
MA 21G	4 digits	3 723 *(40.1)*	188 *(35.3)*	1 763 *(35.7)*	1 580 *(45.6)*	170 *(55.19)*	15 *(57.7)*	7 *(70)*
	3 digits	3 709 *(39.9)*	192 *(36.1)*	1 785 *(36.1)*	1 545 *(44.6)*	164 *(53.3)*	15 *(57.7)*	8 *(80)*
	2 digits	3 282 *(35.4)*	168 *(31.6)*	1 558 *(31.5)*	1 362 *(39.3)*	170 *(55.2)*	16 *(61.5)*	8 *(80)*
	1 digit	1 934 *(20.8)*	94 *(17.7)*	906 *(18.3)*	790 *(22.8)*	130 *(42.2)*	9 *(34.6)*	5 *(50)*
	0 digits	1 270 *(13.7)*	60 *(11.3)*	593 *(12.0)*	505 *(14.5)*	100 *(32.5)*	8 *(30.8)*	4 *(40)*
MA 1G	4 digits	2 593 *(27.9)*	114 *(21.43)*	1 076 *(21.79)*	1 214 *(35.0)*	166 *(53.9)*	17 *(65.4)*	6 *(60)*
	3 digits	2 169 *(23.4)*	84 *(15.8)*	853 *(17.3)*	1 043 *(30.1)*	162 *(52.6)*	19 *(73.1)*	8 *(80)*
	2 digits	1 332 *(14.4)*	37 *(6.9)*	471 *(9.5)*	656 *(18.9)*	144 *(46.8)*	18 *(69.2)*	6 *(60)*
	1 digit	497 *(5.4)*	11 *(2.1)*	165 *(3.3)*	244 *(7.0)*	65 *(21.2)*	9 *(34.6)*	3 *(30.0)*
	0 digits	241 *(2.6)*	5 *(0.9)*	73 *(1.5)*	117 *(3.4)*	36 *(11.7)*	7 *(26.9)*	3 *(30)*
Traditionally anonymized		2 792 *(30)*	142 *(26.7)*	1 335 *(27.0)*	1 181 *(34.0)*	127 *(41.2)*	5 *(19.2)*	2 *(20)*

** 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 no protection at all. The great deviations between the surveys are more decisive for this phenomenon than the slight (almost negligible) modifications to the target data. Regarding the 0-digit level, the anonymization even shows a disclosing effect (1270 correct assignments obtained by matching the micro aggregated data against 1259 hits by matching the formally anonymized data). As had to be expected in the authors' opinion, the variant MA 1G produces nearly safe microdata. On the other hand, this variant is connected with an unbearable abatement of statistical properties.[10] 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 surveys 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.

The following table contains the adjusted hit rates. These rates show the actual protection effect of the anonymization. It also turns out that MA 21G does not develop any protection effect. On the other hand the table shows the anonymization effect through truncation of the NACE code. The first line demonstrates the effectiveness of the applied matching algorithm. If there were no deviations of the blocking variables and no additional anonymization, nearly 3 of 4 enterprises would be re-identified.

[10] Rosemann, M., Wiesbaden (2003), 40-58

Table 5. Re-identification: percentage of successful attempts without deviations concerning NACE code and regional key

Target data	NACE	Total	Employee size class*					
			1	2	3	4	5	6
Formally anonymized	4 digits	72.8	70.7	70.0	75.1	84.9	83.3	100
	3 digits	64.0	62.0	61.0	66.7	76.4	81.0	75.0
	2 digits	47.8	46.3	44.4	50.2	68.4	72.7	100
	1 digit	25.4	22.8	23.0	26.7	49.1	39.1	50.0
	0 digits	13.8	11.7	12.1	15.0	29.7	42.3	44.4
MA 21G	4 digits	72.7	70.7	69.9	75.0	85.4	83.3	100
	3 digits	63.8	62.9	61.2	65.8	77.4	71.4	100
	2 digits	47.7	46.3	44.4	49.9	69.7	72.7	100
	1 digit	25.2	22.5	22.9	26.4	48.7	36.1	62.5
	0 digits	14.0	11.5	12.2	14.9	33.3	30.8	44.4
MA 1G	4 digits	50.6	42.9	42.7	57.6	83.4	94.4	85.7
	3 digits	37.3	27.5	29.2	44.4	76.4	90.5	100
	2 digits	19.4	10.2	13.4	24.0	59.0	81.8	75.0
	1 digit	6.5	2.6	4.2	8.2	24.4	39.1	37.5
	0 digits	2.7	1.0	1.5	3.5	12.0	26.9	33.3
Traditionally anonymized		40.6	39.1	38.0	43.2	52.1	22.7	25.0

** 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.*

According to table 6 it can be said that the higher the frequency of enterprises in an economic sector, the lower the probability of true re-identification.

Table 6. Correlation between the percentage of successful attempts and the frequency of enterprises in a branch of economic activity

Target data	NACE		
	4 digits	3 digits	2 digits
Formally anonymized	-0.504	-0.683	-0.777
MA 21G	-0.499	-0.665	-0.779
MA 1 G	-0.474	-0.644	-0661
Traditionally anonymized	-0.413		

The following table contains descriptive statistics applied to the sets of re-identified enterprises concerning the total number of employees. The first line contains the descriptive statistics of the confidential data as a whole.

Table 7. Descriptive statistics of re-identified enterprises (total employees)

Target data	NACE	Re-identification of the biggest enterprise*	Mean	Minimum	Standard deviation	Frequency Of enterprises
		-	295.8	20	2 888.7	9 283
Formally anonymized	4 digits	1	431.3	20	4 326.3	3 726
	3 digits	1	421.8	20	4 279.4	3 720
	2 digits	1	488.6	20	4 673.4	3 287
	1 digit	0	445.5	20	5 469.0	1 951
	0 digits	0	626.0	20	6 897.7	1 259
MA 21G	4 digits	1	433.2	20	4 328.9	3 723
	3 digits	1	443.3	20	4 396.3	3 709
	2 digits	1	491.5	20	4 677.9	3 282
	1 digit	1	532.4	20	8 924.5	1 934
	0 digits	1	644.5	20	5 883.1	1 270
MA 1G	4 digits	1	561.3	20	5 117.5	2 593
	3 digits	1	688.4	20	5 743.1	2 169
	2 digits	1	936.6	20	7 126.3	1 332
	1 digit	1	1200.6	21	8 924.5	497
	0 digits	1	1946.6	20	12 750.1	241
Traditionally anonymized		0	305.6	20	2 279.2	2 792

** 1 = yes; 0 = no*

Apparently, a stronger anonymization of the data (among others caused by coarsening the NACE code) induces a larger number of employees of the correctly matched enterprises. However, the standard deviation is in step with the size of the re-identified enterprises. For this reason, it is stated again that on the one hand the anonymization methods considered have stronger effects on the smaller enterprises, but on the other hand not all small enterprises could be protected by these methods.

4.2.2 Effects of coarsening NACE classification and legal form

In the following, we present the results obtained by coarsening both the NACE code and the legal status. The latter is coarsened in the same way as described in section 4.1 for the method of traditional anonymization. Thus, we have four different values for this attribute.

Table 8. Re-identification: frequency of successful attempts in absolute terms (percentage) by coarsening the legal status

Target data	NACE	Total	Employee size class*					
			1	2	3	4	5	6
Formally Anonymized	4 digits	3 527 *(37.9)*	175 *(32.9)*	1 644 *(33.3)*	1 526 *(44.0)*	163 *(52.9)*	13 *(50)*	6 *(60)*
	3 digits	3 501 *(37.7)*	175 *(32.9)*	1 656 *(33.5)*	1 492 *(43.0)*	156 *(50.7)*	15 *(70)*	7 *(70)*
	2 digits	2 945 *(31.7)*	146 *(27.4)*	1 389 *(28.1)*	1 244 *(35.9)*	151 *(49.0)*	12 *(46.1)*	3 *(30)*
	1 digit	1 587 *(17.1)*	75 *(14.4)*	718 *(14.5)*	669 *(19.3)*	115 *(37.3)*	8 *(30.8)*	2 *(20)*
	0 digits	1 081 *(11.4)*	51 *(9.6)*	488 *(9.9)*	441 *(12.7)*	92 *(12.7)*	7 *(26.9)*	2 *(20)*
MA 21G	4 digits	3 527 *(37.9)*	175 *(32.9)*	1 645 *(33.3)*	1 525 *(44.0)*	163 *(52.9)*	13 *(50.0)*	6 *(60)*
	3 digits	3 499 *(37.7)*	175 *(32.9)*	1 656 *(33.5)*	1 491 *(43.0)*	155 *(50.3)*	15 *(57.7)*	7 *(70)*
	2 digits	2 941 *(31.7)*	146 *(27.4)*	1 393 *(28.2)*	1 236 *(35.6)*	151 *(49.0)*	11 *(42.3)*	4 *(40)*
	1 digit	1 569 *(16.9)*	76 *(14.3)*	712 *(14.4)*	660 *(19.0)*	108 *(35.1)*	8 *(30.8)*	5 *(50)*
	0 digits	1 055 *(11.4)*	52 *(9.8)*	484 *(9.8)*	426 *(12.3)*	81 *(26.3)*	7 *(26.9)*	5 *(50)*
MA 1G	4 digits	2 340 *(25.2)*	100 *(18.8)*	919 *(18.6)*	1 143 *(32.9)*	158 *(51.3)*	15 *(57.7)*	5 *(50)*
	3 digits	1 881 *(20.3)*	69 *(13.0)*	690 *(14.0)*	957 *(27.6)*	141 *(45.8)*	17 *(65.4)*	7 *(70)*
	2 digits	1 037 *(11.2)*	23 *(4.3)*	355 *(15.1)*	525 *(15.1)*	116 *(37.7)*	11 *(42.3)*	7 *(70)*

	1 digit	317 *(3.4)*	7 *(1.3)*	89 *(1.8)*	160 *(4.6)*	49 *(15.9)*	6 *(23.1)*	6 *(60)*
	0 digits	199 *(2.1)*	7 *(1.3)*	72 *(1.5)*	85 *(2.5)*	28 *(9.1)*	3 *(11.5)*	4 *(40)*
Traditionally anonymized		2 792 *(30)*	142 *(26.7)*	1 335 *(27.0)*	1 181 *(34.0)*	127 *(41.2)*	5 *(19.2)*	2 *(20)*

** 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.*

The protection effect of coarsening the legal status is much weaker than had been expected by the authors. Since there were no deviations in the legal status between both surveys, the variable legal status seems to be partially redundant in the remaining set of key variables. Comparing tables 5 and 9, this observation becomes more transparent.

Table 9. Re-identification: percentage of successful attempts without deviations concerning NACE code and regional key and by coarsening the legal status

Target data	NACE	Total	Employee size class*					
			1	2	3	4	5	6
Formally anonymized	4 digits	68.9	65.8	65.2	72.4	81.9	72.2	85.7
	3 digits	60.2	57.4	56.7	63.6	73.6	71.4	87.5
	2 digits	42.8	40.2	39.6	45.6	61.9	54.6	37.5
	1 digit	20.7	17.9	18.1	22.4	43.1	34.8	25.0
	0 digits	11.9	9.8	10.1	13.0	30.7	26.9	22.2
MA 21G	4 digits	68.9	65.8	65.2	72.4	81.9	72.2	85.7
	3 digits	60.2	57.4	56.7	63.5	73.1	71.4	87.5
	2 digits	42.8	40.2	39.7	45.3	61.9	50.0	50.0
	1 digit	20.5	18.2	18.0	22.0	40.5	34.8	62.5
	0 digits	11.6	10.0	10.0	12.6	27.0	26.9	55.6
MA 1G	4 digits	45.7	37.6	36.4	54.3	79.4	83.3	71.4
	3 digits	32.4	22.6	23.6	40.8	66.5	80.9	87.5
	2 digits	15.1	6.3	10.1	19.2	47.5	50.0	87.5
	1 digit	4.1	1.7	2.2	5.4	18.4	26.1	75
	0 digits	2.2	1.3	1.5	2.5	9.3	11.5	44.4
Traditionally anonymized		40.6	39.1	38.0	43.2	52.1	22.7	25.0

** 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.*

As in section 4.2.1, the above table confirms the general conjecture that large enterprises are easier to re-identify than smaller ones, though in some variants of micro aggregation the effect of protection concerning larger enterprises was stronger than expected. Partially, the frequency of re-identified enterprises was recessive regarding the employee size classes 5 and 6. The section concludes with table 10 containing the associated descriptive statistics:

Table 10. Descriptive statistics of re-identified enterprises (total employees) by coarsening the legal status

Target data	NACE	Re-identification of the biggest enterprise*	Mean	Minimum	Standard Deviation	Frequency of enterprises
		-	295.8	20	2 888.7	9 283
Formally anonymized	4 digits	1	423.9	20	4 307.2	3 527
	3 digits	1	441.8	20	4 390.4	3 501
	2 digits	1	420.6	20	4 426.5	2 945
	1 digit	0	427.7	20	3 023.7	1 587
	0 digits	0	518.4	20	3 918.6	1 081
MA 21G	4 digits	1	424.9	20	4 307.8	3 527
	3 digits	1	442.5	20	4 392.2	3 499
	2 digits	1	432.6	20	4 495.5	2 941
	1 digit	1	640.0	20	6 223.9	1 569
	0 digits	1	835.9	20	7 709.9	1 055
MA 1G	4 digits	1	571.1	20	5 213.1	2 340
	3 digits	1	715.6	20	5 977.1	1 881
	2 digits	1	1 076.2	20	8 006.0	1 037
	1 digit	1	2 394.4	21	14 147.5	317
	0 digits	1	2 302.1	20	14 306.2	199
Traditionally anonymized		0	305.6	20	2 279.2	2 792

1 = yes; 0 = no

The general structure of the above table is similar to the one of table 7. Due to the additional coarsening of the legal status the distribution characterisitcs “mean” and “standard deviation” have grown.

4.3 Match for a single individual

The intention behind the single individual match is to gain information about a specific target individual. The data intruder collects information using several sources. The collected information is then used to re-

identify the target individual in order to get additional information about it. In the German turnover tax statistics such a data attack was simulated. 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)[11] 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.

5 Conclusion

The present paper provides an overview of the functions and forms of application of a matching algorithm which will be also available in the next version of the software package µ-Argus.

Since weak anonymization of data implies a threat of confidentiality and strong anonymization implies a lack of scientific usefulness, in praxi different degrees of anonymization should be compared in order to find a balanced file taking into account both objectives. The application in section 4 presents a modus operandi for the generation of consecutive anonymisations of some confidential data approximising a pregiven threshold of re-identification risk. Certainly, regarding this example there are various meaningful anonymisation variants uncared-for, containg further variants of micro aggregation, whose protection effects can be placed – applying the same traditional methods to the data - between the calculated risks for MA 1G and MA 21G. Also it is possible to coarsen the variable regional key. Some ways of presenting this variable in a less detailed form are shown in appendix 3.

In general, we observe that the suppression of information regarding the categorical variables, particularly the coarsening of variables, accounts vastly for the protection of the data and thus allows much weaker application of, perturbation methods to the remaining metric variables. In such a way the statistical offices of the federation and the Länder could recently generate and disseminate anonymised data for scientific purposes (so called Scientific-Use-Files) arisen from the German turnover tax statistics of the year 2000 and the German structure of costs survey of the year 1999.

This work was partially supported by the EU project IST-2000-25069, Computational Aspects of Statistical Confidentiality, and by the German national project De Facto Anonymization of Business Microdata.

[11] see: Vorgrimler, D., Wiesbaden (2003), 40-58

References

Brand, R., Bender, S., Kohaut, S.: 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 (1999), 57-74

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

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

Hundepool, A., Van de Wetering, A., Ramaswamy, R. et al.: μ-Argus Version 3.2 User's Manual. Netherlands Official Statistics, Vooburg (2003)

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

Jaro, M. A.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, Vol. 89 (1989), 415-435

Lenz, R.: A graph theoretical approach to record linkage. Appears in: Proceedings of the joint UN-ECE/Eurostat work session on statistical data confidentiality, Luxembourg (2003)

Lenz, R.: A way to combine probabilistic with deterministic record linkage. Appears in: Proceedings of the Workshop on Microdata, Stockholm (2003)

Lenz, R.: Disclosure of confidential information by means of multi objective optimisation. Comparative analysis of enterprise (micro) data conference, London (2003)

Lovasz, L., Plummer, M. D.: Matching Theory. Annals of Discrete Mathematics, vol. 29, North Holland, Amsterdam (1986)

Rosemann, M.: Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik. Forum der Bundesstatistik, Vol. 42, Wiesbaden (2003), 154-183

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

Sturm, R.: Anonymisation of economic microdata. Methods/Approaches/Developments 1/2001, Information of the German Federal Statistical Office, Wiesbaden (2001), 1-2

Sturm, R.: Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten. Allgemeines Statistisches Archiv, Vol. 86 (2002), 468-477

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

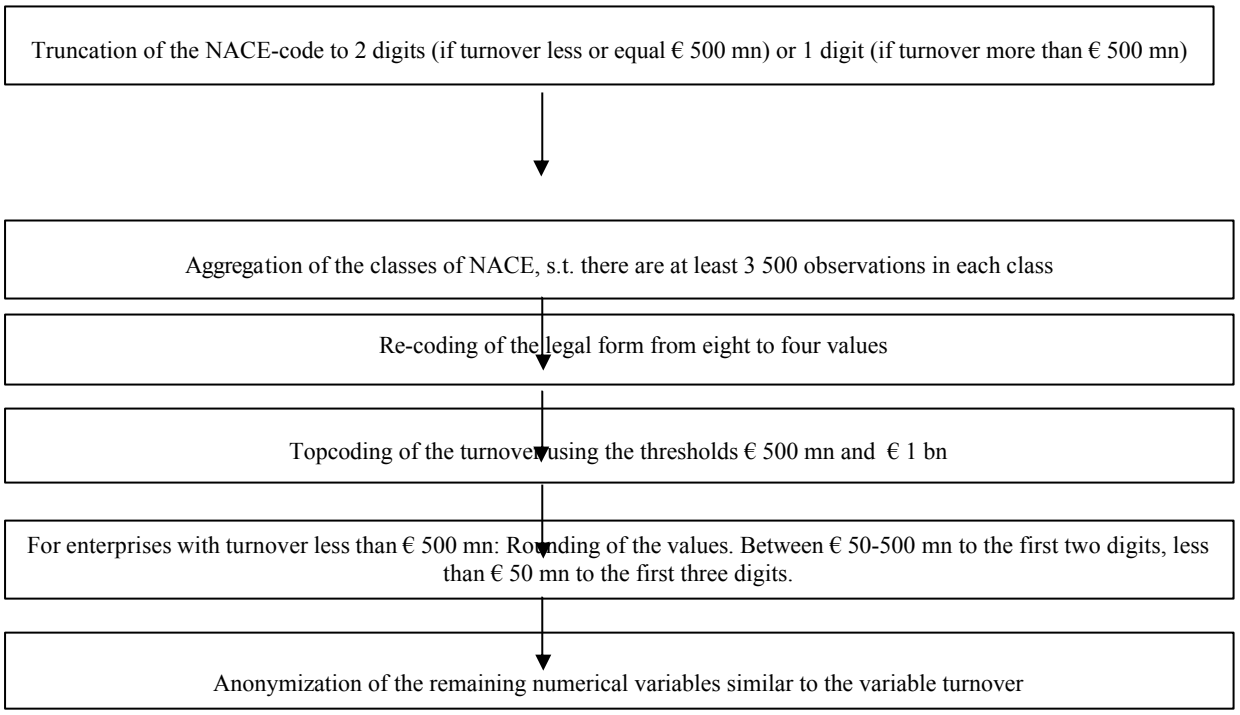
Vorgrimler, D., Rosemann, M.: Effects of anonymization on analytical validity and reidentification risk. Appears in: Proceedings of the Workshop on Microdata, Stockholm (2003)

Wende, T., Zwick, M.: Research data centres of the official statistics. Proceedings of the Seminar Session of the Conference of European Statisticians, Geneva (2003), 141-146

Appendix 1: The relevant variables in the German turnover tax statistics

- Regional reference (so called category 9)
- Branch of economic activity
- Duration of liability to pay taxes (standardized)
- Integrated group of companies pursuant to §2, Para. 2, item 2 (0= no, 1 = yes)
- Legal status
- Deliveries and other performances
  - Taxable deliveries and other performances
    - at 16 %
    - at 7 %
  - Non-taxable deliveries and other performances
    - with prior-tax deduction
      - intra-Community deliveries and other performances
    - without prior-tax deduction
- Turnover tax before deducting prior tax
  - for deliveries and other performances
  - for intra-Community acquisitions
- Deductible portion of prior tax
  - for deliveries and other performances
    - from invoices of other enterprises
    - import turnover tax
  - for intra-Community acquisitions
- Tax prepayment target
- Memo item: intra-Community acquisitions
- Previous year's values in euros
  - Deliveries and other performances
  - Tax prepayment target

Appendix 2: Traditional anonymization procedure



Appendix 3: Types of areas delimited to the basis of settlement structure

The following three variations of the regional key, broken down into three, seven and nine regional types, (*more detailed types of regions, basic types* and *types of districts*), are appropriate for further investigations.