FDZ-Arbeitspapier
Nr.6

Rainer Lenz

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Measuring the disclosure protection of micro aggregated business microdata

2006

FDZ-Arbeitspapier
Nr.6

Rainer Lenz

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Measuring the disclosure protection of micro aggregated business 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

Erscheinungsfolge: unregelmäßig
Erschienen im Juni 2006

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.

Measuring the disclosure protection of micro aggregated business microdata

An analysis taking the example of German Structure of Costs Survey

Rainer Lenz[1]

Abstract. To assess the effectiveness of an anonymisation method with respect to data protection, the disclosure risk associated with the protected data must be evaluated. We consider the scenario where a possible data intruder matches an external database with the entire of confidential data. In order to improve his external database he tries to assign as many correct pairs of records (that is, records referring to the same underlying statistical unit) as possible. The problem of maximisation of the number of correctly assigned pairs is translated into a multi-objective linear assignment problem (MOLP).
Regarding several variants of the micro aggregation anonymisation method applied to the German structure of costs survey, we calculate approximative solutions to the MOLP obtained by using two external databases as the data intruder's additional knowledge. Finally, a standard for so-called de facto anonymity is suggested.

1 Introduction

The deep interest in secret data has long a tradition. Towards the end of December 1855 the deputy purveyor in chief in the Crimea, David Fitzgerald, submitted to the chief medical officer a *Confidential Report on the Nursing, since its introduction to the Crimea on 23rd January.* The contents of the confidential report soon became widely known in medical and military circles. In it the purveyor criticised the nurses for lack of discipline, for drunkennness and insubordination. In several letters Florence Nightingale, conducting at this time a training establishment for nurses in Scutaris, expressed her nonconformity with the report: "Having found that Mr. Fitzgerald's *Confidential Report* was *Confidential* only from myself, & has already ceased to be so in the Crimea ..."

In the last decade, the problem of confidentiality has become increasingly severe, since the number of sources available to data intruders has risen, not least because of the rapid expansion of the Internet. Disclosure occurs when an individual or an enterprise can be re-identified in data that were collected with an assurance to protect confidentiality. Therefore, distributors of data (like statistical offices or private institutions) make sure to handle confidential data with the utmost care. Their challenge is to pursue two objectives, i.e. providing useful statistical information and ensuring protection to respondents. That is, the distributing

[1] Federal Statistical Office of Germany, Research Data Centre, rainer.lenz@destatis.de

institutions have to weigh up between the preservation of accuracy and analytical validity of data and the anonymisation of data in order to minimise the risk of re-identification of statistical units to which they relate.

In 1987, the German Law on Statistics for Federal Purposes created the privilege of science, allowing scientists and researchers access to so-called `de facto anonymised microdata`. A data set was defined to be de facto anonymous if the costs of re-identification exceeded the benefit of re-identification (Sturm 2002). By then, only completely anonymised microdata files could be provided to scientists. That is, the statistical offices had to make sure that intruders had no chance to deanonymise data in order to gain infomation about specific organisations. Moreover, the statistical offices have ethical reasons to protect respondents and they must be fully trustworthy to be able to gather data from respondents.

In Rosemann et al. (2004) the analysis potential of several variants of micro aggregation of the German structure of costs survey has been examined. In the current paper the associated re-identification risk is studied. In order to re-identify statistical units, a data intruder needs additional knowledge about the units searched for (e.g. in the form of an external database) containing variables the additional and the confidential data have in common. Moreover, he needs knowledge about the participation of the units in the target survey, the so-called response knowledge.

For an estimation of the re-identification risk (in business microdata), we consider three scenarios of attack:

A *Assignment between original data and anonymised target data.*
(Calculation of an upper bound for the re-identification risk)

B1 *Assignment between external data and formally anonymised target data (i.e. original data without direct identifiers).*
(Estimation of the natural protection in the data)

B2 *Assignment between external data and anonymised target data.*
(Realistic scenario)

The results obtained by scenario A do not represent the reality. Nevertheless, it would seem advisable to involve scenario A into the estimation of the re-identification risk associated with the tested microdata file and hence into the decision about the de facto anonymity of the file, since the available additional knowledge of a data intruder is inassessable. The results obtained by scenarios B1 and B2 line out, how far by courtesy of an assumption of the best possible additional knowledge in scenario A the real re-identification risk is overestimated.

2 Basic definitions and notations

Throughout the paper we will use the following denotations: A (finite) `graph` $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$ is a relational structure, consisting of a (finite) set $V(\mathcal{G})$, the elements of which are called `vertices` (or points), and a set $E(\mathcal{G}) \subseteq V(\mathcal{G})^2$ of unordered pairs of vertices, called `edges` (or lines) of $\mathcal{G}$. We denote these sets by V or E when there is no possibility of confusion. We consider undirected graphs, fulfilling the implication $(a, b) \in E \Longrightarrow (b, a) \in E$. That is, E determines a symmetric binary relation. The edge $(x, y) \in E$ is said to be `incident` with the vertices x and y, and x, y to be `adjacent`. A graph $\mathcal{S} = (V(\mathcal{S}), E(\mathcal{S}))$ is called `subgraph` of $\mathcal{G}$ if $V(\mathcal{S}) \subseteq V(\mathcal{G})$ and $E(\mathcal{S}) \subseteq E(\mathcal{G})$ holds.

$\mathcal{G}$ is called `bipartite graph` with bipartition (X, Y) if $V(\mathcal{G})$ is a disjoint union $V = X \dot\cup Y$, so that every edge e is incident with both an $x \in X$ and a $y \in Y$. Moreover, if every $x \in X$ is connected to every $y \in Y$, the graph $\mathcal{G}$ is said to be `complete`. A `matching` $\mathcal{M}$ of $\mathcal{G}$ is a subgraph with the property that no two edges in $\mathcal{M}$ have a common vertex. $\mathcal{M}$ is called `maximum matching` if there is no matching $\mathcal{M}'$ with $\mathcal{M} \subset \mathcal{M}'$. If v is a vertex of $\mathcal{M}$, then $\mathcal{M}$ `saturates` v. Moreover, if every $v \in V$ is saturated, $\mathcal{M}$ is called `perfect matching`. A `vector-weighted` graph $\mathcal{G}$ is a graph combined with a weight function

$$\begin{array}{rcl} w : E(\mathcal{G}) & \longrightarrow & \mathbb{R}^k, \\ e & \longmapsto & (w_1(e), \ldots, w_k(e)), \end{array}$$

which maps every edge e to a k-tuple of real numbers. In the case $k = 1$ the graph $\mathcal{G}$ is a `weighted graph`.

3 Types of variables and distances

In a `database cross match`, see Elliott and Dale (1999), the data intruder matches an external database B with the whole confidential database A. For this, he uses variables which the external data have in common with the confidential data, the so-called `key variables`.

External data		
identifiers (name, address)	key variables (e.g. turnover, number of employees)	
	key variables (e.g. turnover, number of employees)	target variables (e.g. investments)
	Target data	

Fig. 1. Empirical key variables

Obviously, the reporting quality of these variables is crucial for the success of the subsequent re-identification process. The set of key variables is partitioned into two classes of variables, namely `categorical` and `numerical variables`, which are described below.

3.1 Categorical and numerical variables

`Numerical variables` are defined to be discrete or continuous variables where the difference between values has a meaning, e.g. "height", "weight" of a person or "number of employees", "total turnover" of an enterprise. Regarding a numerical variable v_i, its values are distanced by $d_i(a,b) = (a_i - b_i)^2$. In general, the treatment of numerical variables admits less diversification than the treatment of categorical variables. When analysing categorical data, it is always possible to work with counts or percentages of objects which fall within certain categories. We differentiate between `nominal variables` (there is no ranking of the categories) and `ordinal variables` (the categories have some kind of order, where differences between categories are meaningless). Regarding a nominal variable v_i, its values are compared for equality, such that we define

$$d_i(a,b) \quad = \quad \begin{cases} 0, & \text{if } a_i = b_i, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

For a linear ordered variable v_i, let $c_1 <_i c_2 <_i \ldots <_i c_r$ be its ordered range (i.e. $(\{c_1, \ldots, c_r\}; <_i)$ is a well-ordered set). We then define

$$d_i(a,b) = \frac{|\{c_j \mid \min(a_i,b_i) \leq_i c_j <_i \max(a_i,b_i)\}|}{r}. \tag{2}$$

Since in practice there will often occur categorical variables with some non-linear partial order (that is, there are at least two categories c_1 and c_2 with $c_1 \not< c_2$ and $c_2 \not\leq c_1$), formula (2) does in general not fit. It is then possible to extend this non-linear ordering to a lattice-order, where supremum $\sup\{c_i, c_j\}$ and infimum $\inf\{c_i, c_j\}$ exist for all pairs (c_i, c_j) . Using an order preserving mapping $f : \{c_1, \ldots, c_r\} \longrightarrow \mathbf{R}$, the i^{th} component distance can be defined by

$$d_i(a,b) = |f(\sup\{a_i, b_i\}) - f(\inf\{a_i, b_i\})|. \tag{3}$$

For instance, consider the n-gram approach used for string variables (Efelky et al. 2002). The distance between two strings is defined as $d_i(a,b) = \sqrt{\sum_{\forall s} |f_a(s) - f_b(s)|}$, where $f_a(s)$ and $f_b(s)$ are the number of occurrences of the substring s of length n in the two strings a and b. Let us consider a small example, where $n = 3$ and the strings HONEY and MONEY are given. We obtain $d_i(HONEY, MONEY) = \sqrt{2}$, since there are two non common substrings of length 3: HON and MON. In the case of hierarchical variables, the following distance function is suggested:

$$d_i(a,b) = \min\{f(c_j) \,|\, c_j < a_i \text{ and } c_j < b_i\}, \tag{4}$$

where for the order preserving mapping f holds $f(c_k) = 0$ if there is no c_l with $c_l <_i c_k$. Hierarchical variables may occur if some categorical variable was partially coarsened to different degrees in the data set. For instance if in a business database the variable *Branch of economic activity* (NACE code) is for some units specified on a 3-digit-level and for other units specified on a 1- or 2-digit-level (see figure 2).

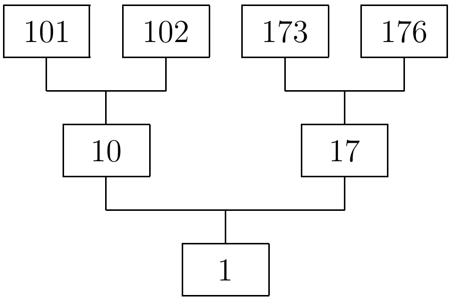

Fig. 2. Coarsening of the NACE code

Here, the distance between hierarchical variables is preferable to the simple $(0-1)$-distance in (1) being strongly separating, particularly if considerable deviations are observed between both data sets in these variables.

For the planned adaption of different types of component distances, depending on the types of variables v_r, it is necessary to standardise the distances d_r in order to avoid scaling problems, e.g. by use of the $max - min$ standardisation

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

For large data sets, it is recommended to partition the data into subsets, as described in subsection 3.2.

3.2 Blocking variables

To be a candidate for a possible assignment, it is necessary for a record pair that both records coincide in their values of some specified variables. In the following these variables are called `blocking variables` (e.g. see Jaro 1989), since they divide the whole data into disjoint blocks.

The aim of blocking data is on the one hand to reduce the complexity of the subsequent assignment procedure and the allocated main storage and on the other hand the number of mismatches. Though the number of possible mismatches grows with the number of wrongly classified records (that is, two records a and b which refer to the same individual are possibly not members of the same block), mismatches have to be expected especially in large blocks as there are many similar distances. Whether it will be possible here to find a reasonable tradeoff depends on the quality of the variable used for blocking. It is in general difficult to estimate the reporting error probability of some variable intended for blocking. In the worst case, the corresponding blocks in both data sets are disjoint, in the best case the chosen blocking variables are unique identifiers, such that truly associated records belong to the same block. If possible, the data intruder will opt for those variables, which are known to have been left out of consideration within the anonymisation process.

The treatment of specific variables like identifiers and blocking variables can also be embedded into the calculation of distances, where identifier variables are handled like nominal variables. Now let v_i be a blocking variable. In the application in section 7, the appearing blocking variables are categorical ones, where the corresponding component distances are defined by (1). Nevertheless, if a numerical variable v is intended for blocking, it is strongly recommended to aggregate the range of v into intervals, such that every value falls uniquely into some interval. That is, two values are distanced with zero if and only if their intervals (categories) coincide.

From a theoretical point of view, the setting of blocking variables is a special case of multidimensional clustering (see Schweigert 1999). Metrics often used in clustering analysis like the `general matrix metric`

$$d(x,y) = \sqrt{(x-y)^{\mathrm{T}} C (x-y)},$$

where C is an arbitrary symmetric, positive definite matrix, are in most cases not of practical relevance, since they involve an additional unacceptable computational amount already for data files of medium size. Particularly, if C determines the inverse S^{-1} of the empirical covariance matrix.

Note that there is a number of alternative methods to realise a preselection of candidate pairs. A thorough analysis of those methods can be found in Elfeki, Verykios and Elmagarmid (2002).

A formalisation of the concept of allocating individual weights to all key variables is given in the subsequent section.

4 Preference functions and matchings

Among others, the success of a record linkage algorithm depends on the choice of distance measures and on the reliability of each variable. Therefore, the decision maker rates the key variables and prefers some of them to the others. This is done by use of so-called preference functions.

Definition 1 *Let* $\Lambda = (\lambda_1, \ldots, \lambda_k) \in (\mathbb{R}^+)^k$ *be a k-tuple of positive real numbers. For a record* $r = (r^{(1)}, \ldots, r^{(s)})$*, where s is the number of all variables, let w.l.o.g. the entries* $r^{(1)}, \ldots, r^{(k)}$ *be the values of the key variables. We define a k-ary* `linear preference` `function` $f_\Lambda : \mathrm{I\!R}^k \longrightarrow \mathrm{I\!R}$ *by*

$$f_\Lambda(x_1, \ldots, x_k) = \sum_{i=1}^{k} \lambda_i x_i.$$

Setting $\sum_{i=1}^{k} \lambda_i = 1$ and hereby $\lambda_k = 1 - \sum_{i=1}^{k-1} \lambda_i$, we may reduce the set of para- meters to $\{\lambda_1, \ldots, \lambda_{k-1}\}$. The permutation τ, defined in such a way that

$$\lambda_{\tau(1)} > \lambda_{\tau(2)} > \ldots > \lambda_{\tau(k)},$$

can be understood as an individual ranking of variables by the decision maker. In the theory of multicriteria optimisation, linear preference functions are used to turn multiobjective optimisation problems into single objective ones (Schweigert 1995).
Let $\mathcal{M}$ be a matching and $M \subseteq A \times B$ its set of edges. We define componentwise

$$d_i(\mathcal{M}) := \sum_{(a,b)\in M} d_i(a,b), \; i = 1, \ldots, k$$

and further

$$\Delta(\mathcal{M}) := (d_1(\mathcal{M}), \ldots, d_k(\mathcal{M})).$$

Definition 2 *A maximum matching $\mathcal{M}$ is called a* `preference matching` *if there is a preference function f_Λ such that $f_\Lambda(\Delta(\mathcal{M})) \leq f_\Lambda(\Delta(\mathcal{M}'))$ holds for every maxi- mum matching $\mathcal{M}'$.*

A single edge (a, b) can be regarded as a (non maximum) matching. Preference functions involve for every $(a, b) \in A \times B$ the distance

$$d(a,b) := f_\Lambda(\Delta(a,b)) = \sum_{i=1}^{k} \lambda_i \cdot d_i(a,b).$$

This expression can be regarded as a weighted sum of all component distances. Now we are able to calculate the distances $d(a, b)$ for $a \in A$ and $b \in B$, split into component distances associated with categorical or numerical variables.

$$\begin{aligned} d(a,b) &= \sum_{i=1}^{k} \lambda_i \cdot d_i(a,b) \\ &= \tau \cdot \sum_{i\in CV} \tilde{\lambda}_i \cdot d_i(a,b) + \sum_{i\in NV} \lambda_i \cdot d_i(a,b), \end{aligned}$$

where CV is the set of indices of the categorical variables and NV its complement set of indices of numerical variables. The parameter $\tau = \lambda_i/\tilde{\lambda}_i$ is an adaptive control parameter, needed to balance the influence of categorical and numerical variables in order to achieve a reasonable adaption. Key variables – besides blocking variables – which are involved in distance calculations are tentatively called `matching variables`. Note that the concept of blocking data can also be embedded into the calculation of distances as mentioned in subsection 3.2. The distances $d(a,b)$ can be split into their component distances associated with blocking variables (BV) and matching variables (MV),

$$d(a,b) = \sum_{i \in BV} d_i(a,b) + \sum_{i \in MV} \lambda_i \cdot d_i(a,b),$$

where identifier variables may be contained, for the sake of easy implementation, in the set of matching variables, weighted with $\lambda_i = 0$. The weights of blocking variables are allocated by $\lambda_i = 1$. If the records a and b coincide in their blocking variables, the first sum is zero. Replacing λ_i by $\lambda_i / \sum_{j \in MV} \lambda_j$ for each matching variable v_i one obtains a convex combination of the component distances, such that the second sum is a value less than or equal one. In other words, two records a and b are distanced by $d(a,b) \leq 1$ if and only if they are classified to the same block. An alternative realisation of blocking data is to presort the whole data by blocking variables and to read the data blockwise. The experience gathered by the author has shown, however, that reading the data in and out block by block, a process usually not accounted for in complexity analyses, is extremely time-consuming. Moreover, for large data sets the summed-up distances $d(a,b)$ should be compared with some appropriate threshold value $c < 1$ – determined a priori – to decide whether the overall distances are small enough to classify the pairs (a,b) as true matches.

5 Linear assignment problem

In a non-technical way, the concept of matching may be introduced as bringing together pairwise information from two records, taken from different data sources, that are believed to refer to the same individual. The records a and b are then said to be `matched`. Since in general there is the possibility that the matching could be wrong, it is tried to minimise the number of mismatches.

In the following let $n = |A| = |B| = m$. Otherwise consider w.l.o.g. the case $m < n$. Dually, the case $n < m$ can be treated. We then define new objects $b_{m+1}, \ldots, b_n$ which induce new pairs (a_i, b_j) for $i = 1, \ldots, n$ and $j = m+1, \ldots, n$, at a distance

$$d(a_i, b_j) := (\max_{(a,b) \in A \times B} d_1(a,b), \max_{(a,b) \in A \times B} d_2(a,b), \ldots, \max_{(a,b) \in A \times B} d_k(a,b)).$$

We obtain the `multi-objective linear program` described below:

$$\text{Minimise} \begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{n} d_1(a_i, b_j) x_{ij} \\ \sum_{i=1}^{n} \sum_{j=1}^{n} d_2(a_i, b_j) x_{ij} \\ \vdots \\ \sum_{i=1}^{n} \sum_{j=1}^{n} d_k(a_i, b_j) x_{ij} \end{cases} \qquad \textbf{(MOLP)}$$

$$\begin{aligned} \text{subject to} \quad & x_{ij} \in \{0,1\} \quad \text{for} \quad i,j = 1,\ldots,n, \\ & \sum_{j=1}^{n} x_{ij} = 1 \quad \text{for} \quad i = 1,\ldots,n \quad \text{and} \\ & \sum_{i=1}^{n} x_{ij} = 1 \quad \text{for} \quad j = 1,\ldots,n. \end{aligned}$$

The constraints ensure that every a_i is connected with exactly one b_j and vice versa. That is, $x_{ij} = 1$ if and only if a_i is matched with b_j.

As described in section 4, defining $d(a,b) := \sum_{i=1}^{k} \lambda_i d_i(a,b)$ and abbreviated $d_{ij} := d(a_i, b_j)$, the problem of finding an optimum matching is turned into a single objective assignment problem (AP) using linear preference functions. The main idea is to combine all objectives into one single value, as it is typically done in a linear program formulation. Note that linear approaches in general lead to a considerable loss of useful information. When the summed-up distances d_{ij} are calculated, the question of choosing weights is often glossed over, but in fact it is extremely critical. In Schweigert (1995) it is shown that under certain assumptions it suffices for the decision maker to define a range for the weights. However, there arises the following single objective assignment problem:

$$\text{Minimise} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij}, \qquad \textbf{(AP)}$$

$$\begin{aligned} \text{subject to} \quad & x_{ij} \in \{0,1\} \quad \text{for} \quad i,j = 1,\ldots,n, \\ & \sum_{j=1}^{n} x_{ij} = 1 \quad \text{for} \quad i = 1,\ldots,n \quad \text{and} \\ & \sum_{i=1}^{n} x_{ij} = 1 \quad \text{for} \quad j = 1,\ldots,n. \end{aligned}$$

This assignment problem can be formulated graph theoretically as follows: Find a preference matching on a vector-weighted bipartite graph. In other words, we have to look for a permutation π of $\{1,\ldots,n\}$ which minimises the sum $\sum_{i=1}^{n} d_{i,\pi(i)}$. That is, in order to solve (AP), we might produce all $n!$ matchings of $\mathcal{G}$ and select one of minimum weight. Unfortunately, this algorithm will certainly not be efficient and thus does not justify the transition from problem (MOLP) to (AP). Though there are classical procedures like the well-known simplex-method (e.g. see Papadimitriou and Steiglitz 1998), which – despite non-polynomial worst case run-time – turned out to be effective in practice, problems appeared already while studying data with moderate block sizes. Considering the coefficients connected with the system of linear equations of the restrictions in (AP) the following matrix is generated

$$A = \begin{pmatrix} J_1 & J_2 & \cdots & J_n \\ I_n & I_n & \cdots & I_n \end{pmatrix},$$

where J_i defines a matrix of dimension $n \times n$, whose i^{th} row is the vector $(1\,1\,\cdots\,1)$ and whose remaining entries are zero. $I_n = diag(1, 1, \ldots, 1)$ defines the identity matrix of dimension $n \times n$. The resulting coefficient matrix A, possessing $2n$ rows and n^2 columns, was in several instances responsible for exceeding the working memory.

A way out is to use the `Hungarian Method`. We give a short description of this method, originally proposed for maximum weight perfect matchings (Kuhn 1955 and Munkres 1957), in order to find a minimum weight perfect matching or a preference matching, respectively. Let us consider a complete, weighted bipartite graph $\mathcal{G} = (V, E)$. A `feasible vertex labeling` l is a mapping from the set V into the real numbers, where

$$l(a) + l(b) \leq d(a, b).$$

The number $l(v)$ is then called `label` of v. The `equality subgraph` $\mathcal{G}_l$ is a subgraph of $\mathcal{G}$ which includes all vertices of $\mathcal{G}$ but only those edges (a, b) fulfilling

$$l(a) + l(b) = d(a, b).$$

A connection between equality subgraphs and matchings of minimum weight is provided by the following theorem.

Theorem *Let l be a feasible vertex labeling of $\mathcal{G}$. If the equality subgraph $\mathcal{G}_l$ possesses a perfect matching $\mathcal{M}$, then $\mathcal{M}$ is a minimum weight perfect matching of $\mathcal{G}$.*

Proof: Let $\mathcal{M}$ be a perfect matching of $\mathcal{G}_l$ and $\mathcal{M}'$ be any perfect matching of $\mathcal{G}$. Then it holds that

$$\begin{aligned} d(\mathcal{M}') := \sum_{(a,b)\in\mathcal{M}'} d(a,b) &\geq \sum_{v \in V(\mathcal{G})} l(v) \quad \text{(since } \mathcal{M}' \text{ saturates all vertices)} \\ &= \sum_{(a,b)\in\mathcal{M}} d(a,b) \quad \text{(by definition of } \mathcal{M}) \\ &=: d(\mathcal{M}). \end{aligned}$$

Hence, $\mathcal{M}$ is a minimum weight perfect matching of $\mathcal{G}$. ◇

When applying the algorithm, we use two vectors of labels, $(l(a_1), \ldots, l(a_n))$ and $(l(b_1), \ldots, l(b_n))$, to select admissible edges. Initially, we set

$$\begin{aligned} & l(a_i) = 0 && \text{for } i = 1, \ldots, n \\ \text{and} \quad & l(b_j) = \min_{1\leq i \leq n} d(a_i, b_j) && \text{for } j = 1, \ldots, n. \end{aligned}$$

Using the concept of so-called augmenting paths, we find a matching $\mathcal{M}$ of $\mathcal{G}_l$ which saturates as many vertices as possible. If $\mathcal{M}$ is perfect, according to the above theorem $\mathcal{M}$ is a minimum weight matching of $\mathcal{G}$ and the algorithm stops. $\mathcal{M}$ is then uniquely determined up to equivalence. Else, if $\mathcal{M}$ does not determine a perfect matching, we relax the values for some $l(a)$ and $l(b)$ so that new edges will be admissible.

A competing algorithm is the auction algorithm, introduced for the assignment problem in Bertsekas (1979) and later extended to general transportation problems in Bertsekas and Castanon (1989).

6 Matching algorithm

In this section we suggest heuristic approaches to the single objective linear assignment problem. Though the greedy heuristics introduced below do not guarantee optimality, those approaches are also discussed, since their undoubted advantage is that they work in reasonable time, more precisely in square time complexity according to the number of local units. On the other hand, even when the global solution is not reached, the reached suboptimal solution is in our case a very good solution.

6.1 Greedy heuristics

Often greedy algorithms are preferred on account of easy implementation and quick run-time (e.g. see T. H. Cormen et al. 1990). In fact, the complexity of the procedures below is of order $O(nm)$, where n and m are the numbers of records in A and B, respectively, whereas the Hungarian method sketched in section 5 has complexity of order $O(\max\{n, m\}^3)$ being not practicable for data sets of large size. At this stage the distances belonging to $(a, b) \in A \times B$ can be taken for granted.

Procedure I:

```
begin {PROC I}
$\mathcal{M} := \emptyset$
  $i := 1$
While ($i \leq n$ and $B \neq \emptyset$) do
    $b' := arg\,min_{b \in B} d(a_i, b)$
    $\mathcal{M} := \mathcal{M} \cup \{(a_i, b')\}$
    $B := B \setminus \{b'\}$
    $i := i + 1$
end {PROC I}
```

The procedure's output is an assignment $\mathcal{M}$ of A to B. Obviously, the output depends on the enumeration of $a_1, \ldots, a_n$ and might be far from optimum. Let w.l.o.g. $a_1, \ldots, a_r$ be assigned to $b_{\pi(1)}, \ldots, b_{\pi(r)}$. In step $r + 1$ the target object a_{r+1} is associated with a record b of minimum distance to a_{r+1}. Record b is one of the remaining $m - r$ records in B, which have not been assigned at this stage. Note that deletion of the seventh row of the above procedure would make it possible that some a_i could be assigned to several $b \in B$ (so called 'one to many' assignment), so that the resulting assignment would not be unique and a decision maker would have to make a further selection (Lenz et al. 2004). An essential improvement of Procedure I is achieved by Procedure II below, where the enumeration has a smaller impact and can theoretically be disregarded if there is at least one continuous variable under consideration. The idea consists in a consecutive selection of pairs with the smallest distance as long as both, the external and target data sets, are nonempty.

Procedure II:

```
begin {PROC II}
Sort the distances in an ascending list $\mathcal{L}$
While L is nonempty do
    Consider the first element $d_{ij}$ of $\mathcal{L}$ and assign $(a_i, b_j)$.
    Delete all elements $d_{rs}$ of $\mathcal{L}$, where $r = i$ or $s = j$.
end {PROC II}
```

The most harmful disadvantage of the above procedures is that the records are not linked simultaneously. Nevertheless, experience has shown that Procedure II yields results near the (optimum) solution of (AP). Moreover, for large statistics (for instance the whole of German turnover tax statistics) it is nearly impossible to compute the optimum solution in reasonable time.

6.2 Matching algorithm

Based on the previous considerations we propose the following matching algorithm:

1) Input: Sets $A = \{a_1, \ldots, a_n\}, B = \{b_1, \ldots, b_m\}$ of records and $V = \{v_1, \ldots, v_k\}$ of key variables.

2) Partition the problem into sub-problems by use of blocking variables $BV \subseteq V$.

3) Calculate the component distances in order to construct a vector weighted bipartite graph $\mathcal{G}$.

4) Turn from (MOLP) to (AP) by setting of individual weights $\Lambda = (\lambda_1, \ldots, \lambda_k)$, as described in section 4.

5) To solve (AP), apply alternatively

- the Hungarian Method or
- one of the procedures presented.

6) Output: $(1-1)$-assignment of A to B.

Note that the amount of the calculation of distances takes $O(knm)$, where k is the number of key variables. That is, using one of the heuristics suggested in section 6.1, the total amount of cpu-time of the whole matching algorithm is of order $O(knm)$. In our case, it holds in general $k << n$ and $k << m$, such that we may neglect the factor k within the analysis of complexity. In practise, the cpu-time is significantly reduced by blocking the data.

6.3 Illustrative example

To illustrate the matching algorithm previously mentioned, we consider a small example. We link original data $A = \{a_1, \ldots, a_4\}$ with masked data $B = \{b_1, \ldots, b_4\}$, where the objects are associated by five common variables $v_1, \ldots, v_5$.

obj.\var.	v_1	v_2	v_3	v_4	v_5
a_1	14,008,906	755,187	907,264	6,582,133	4,794,809
a_2	14,309,437	673,189	1,179,713	8,111,720	5,407,676
a_3	14,330,083	567,300	920,065	4,871,720	1,667,078
a_4	14,780,637	567,553	1,026,861	5,313,029	3,654,241
b_1	14,825,332	563,928	913,631	4,978,410	1,711,353
b_2	14,045,802	724,071	1,040,229	7,064,023	5,078,378
b_3	13,945,802	682,110	973,631	7,378,984	508,494
b_4	14,996,199	563,928	1,050,673	5,252,164	3,871,084

Enumeration of all 24 perfect matchings then leads to:

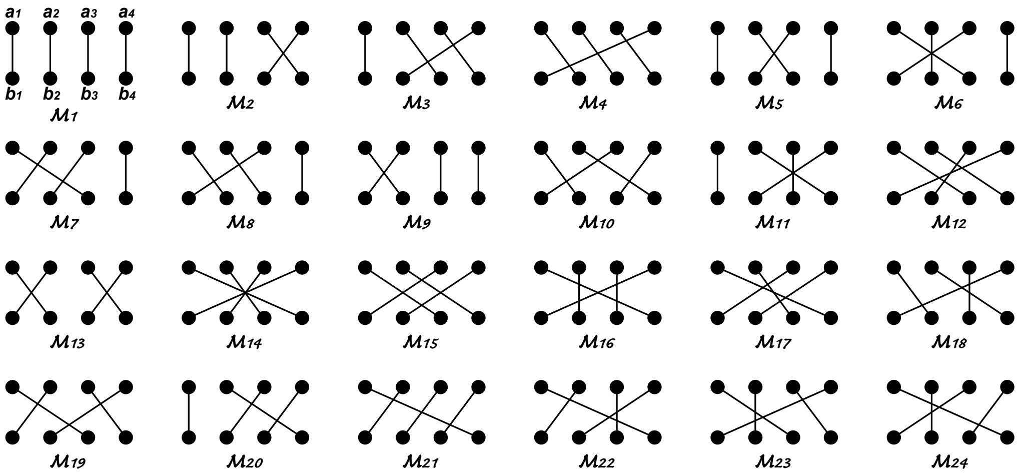

Fig. 3. Perfect matchings

We determine the standardized distances for each component and obtain the following *poset* (partially ordered set) of perfect matchings, ordered lexicographically by their vector-weights

$$\Delta(\mathcal{M}) = (d_1(\mathcal{M}), \ldots, d_5(\mathcal{M})) = \left(\sum_{(a_i,b_j)\in\mathcal{M}} d_1(a_i, b_j), \ldots, \sum_{(a_i,b_j)\in\mathcal{M}} d_5(a_i, b_j) \right).$$

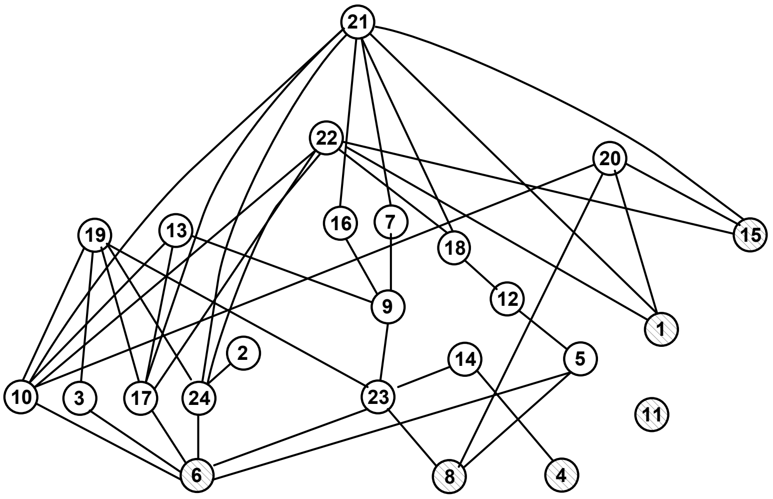

Fig. 4. Poset of vector-weighted perfect matchings

The minimum elements $\mathcal{M}_1, \mathcal{M}_4, \mathcal{M}_6, \mathcal{M}_8, \mathcal{M}_{11}$ and $\mathcal{M}_{15}$ of this poset are called efficient matchings. Taking Figure 4 as a basis, we may intuitively pick $\mathcal{M}_6$ as solution, since it is covered by all inefficient perfect matchings. This intuition is confirmed below.

In order to turn the (MOLP) into (AP), we apply the preference function f_Λ with $\Lambda = (\frac{1}{5}, \ldots, \frac{1}{5})$ to $\Delta(\mathcal{M})$, that is, no objective is prefered to the other. The following table compares the results obtained by the procedures **PROC I, PROC II** and the **H**ungarian **M**ethod. The first entry in each cell refers to the number of true assignments, the second to the total sum of distances.

PROC I	**PROC II**	**HM**
2; 2.88	4; 2.38	4; 2.38

Procedure II and the Hungarian method led to the true assignment $\mathcal{M}_6$. By modification of Λ, application of the corresponding f_Λ to the vector-weights and choice of the Hungarian Method, the decision maker is able to find another efficient matching, e.g. $\Lambda_1 = (\frac{1}{3}, 0, \frac{2}{3}, 0, 0), \Lambda_4 = (0, 1, 0, 0, 0), \Lambda_8 = (0, 0, 0, 1, 0), \Lambda_{11} = (0, 0, 1, 0, 0)$ and $\Lambda_{15} = (0, 0, \frac{2}{3}, 0, \frac{1}{3})$ yield the efficient matchings $\mathcal{M}_1, \mathcal{M}_4, \mathcal{M}_8, \mathcal{M}_{11}$ and $\mathcal{M}_{15}$.

7 Matching the German SCS

In order to advance the conflicting goals of exploiting the research potential of microdata and maintaining acceptable levels of confidentiality, there is need for appropriate anonymisation methods. In this section we study the protection effects of several variants of the micro aggregation anonymisation method (see e.g. Domingo-Ferrer and Mateo-Sanz 2002) on the German structure of costs survey, abbreviated: SCS. Up to now, these methods have proven to generate useful material according to applied econometrics. The micro aggregation method first divides the set of variables into groups. Within a group, the variables are standardised and summed up for each record, such that the records can be sorted by those called Z-scores. Afterwards, for a pregiven number k (in our case $k = 3$), the records with the greatest and smallest Z-scores are classified together with their $k - 1$ nearest neighbours (with respect to Euclidean distance) and their values are averaged. Hence, in the class of micro aggregated data from some confidential original source, the weakest degree of anonymisation is achieved where each variable forms its own group (here, the structure of data is essentially preserved), whereas putting all variables into the same group creates the strongest degree of anonymisation because there are triples of records which agree in all numerical variables and hence can only be separated using the categorical ones. That is, from the class of micro aggregated data of an original source file, the data distributing institution may extract the variant with the desired degree of anonymisation.

We choose the variants of micro aggregation by 1, 8, 11 and 33 groups of variables, which have been extracted by the German national project on anonymisation of business microdata. Similar preservation properties regarding the analytical validity of the data were obtained using the method SAFE, introduced in Evers and Höhne (1999), which has been developed by the Land Statistical Office Berlin. The weakest of the considered variants of micro aggregation, where every numerical variable defines its proper group, is the 33-group variant MA33G. Using this variant the structure of data is widely preserved (Rosemann et al. 2004). The strongest is the multidimensional micro aggregation MA1G, where all numerical variables are grouped together. The variant MA8G is obtained by forming eight groups of a size between two elements (smallest group) and twelve elements (largest group), where highly correlated variables are put together. The variant MA11G is obtained by partitioning the set of numerical variables into three-element groups. We also consider the weakest possible form of anonymisation, formal anonymisation, consisting essentially in the deletion of direct identifiers like name, address and so on (FORMAL).

Subsection 7.1 contains a brief description of the German structure of costs survey. In subsection 7.2, we carry out the realistic scenario, where the data intruder's additional knowledge consists of an external database. For this simulation, we use as external databases both the German turnover tax statistics (subsection 7.2.1) and the commercially available `MARKUS` database (subsection 7.2.2). In subsection 7.3, the previously obtained results are contrasted with those obtained by matching records of the original German structure of costs survey with different variants of anonymisation of the survey. This may be regarded as the worst-case scenario, where the data intruder possesses the original data as the best possible external data. However, one should not presume that the data intruder possesses information about all 33 numerical variables of the survey. Realisticly, the external database of the data intruder will contain only a few key variables as in subsections 7.2.1 and 7.2.2. Regarding examinations as in subsections 7.2.1 and 7.2.2, there are in general many more difficulties to be expected for experiments with data of different sources and fewer key variables, not least because of the fact that the data intruder has – besides the reliable total distance of the assignment – no facility to evaluate his results. It is not least for that reason that the author feels it makes sense – as a concession to the data users – to run experiments also for the worst-case scenario A with variables most likely to be found in commercial enterprise databases in addition to the experiment including all variables.

7.1 The target data used

The German structure of costs survey, limited to the manufacturing industry, is a projectable sample and includes a maximum of 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 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. The complete list of variables available in the German structure of costs survey is appended to the paper.

The table below contains an excerpt of the German structure of costs survey, classified by the categorical variables mentioned above.

Table 1. Partitioning of the German structure of costs survey

nace2\bbr9	1	2	3	4	5	⋯	8	9	Sum
10	5	5	2	4	0	⋯	7	0	39
14	7	19	15	4	2	⋯	24	8	157
⋮						⋯		⋮	
20	38	54	50	15	8	⋯	57	42	504
22	356	154	57	23	91	⋯	54	18	950
24	267	174	82	32	37	⋯	66	14	901
25	97	187	90	25	16	⋯	85	41	867
26	116	108	73	49	35	⋯	100	72	965
27	120	152	44	21	18	⋯	29	16	593
30	33	28	11	2	12	⋯	13	0	153
⋮						⋯		⋮	
37	13	15	6	9	9	⋯	11	2	94
Sum	2,920	2,994	1,379	486	788	⋯	1,488	677	16,918

Totally, there are 26 economic sectors and hence 234 data blocks of a size between 0 and 670 records under consideration.

In the following sections the solutions obtained by Procedure II are presented in detail, since their structures (concerning the distributions of the number of re-identified enterprises among employee size classes) are near the structures obtained by application of Procedure I and the Hungarian method.

7.2 External versus micro aggregated confidential data

In the following we simulate the scenarios B1 and B2 mentioned in section 1 using a sample of around 9,300 records of the German turnover tax statistics (subsection 7.2.1) and a sample of around 9,400 records of the commercially available so-called MARKUS database (subsection 7.2.2) as the data intruder's additional knowledge.

7.2.1 The German turnover tax statistics

Turnover tax statistics (TTS) are based on an evaluation of monthly and quarterly advance turnover tax returns to be provided by entrepreneurs whose turnover exceeds EUR 16,617 and whose tax amounts to over EUR 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 key variables available are:

- Branch of economic activity (NACE2, blocking variable)
- Type of administrative district (BBR9, blocking variable)
- Total turnover (matching variable).

Classifying the number of true matches achieved by Procedure II into intervals relating to the number of employees, we obtain:

Table 2. Re-identifications (TTS) classified by the number of employees[*]

target data	total	1	2	3	4	5	6
MA1G	404 0.0435	103 0.0330	61 0.0259	55 0.0261	64 0.0755	47 0.0916	74 0.2151
MA8G	1,177 0.1270	366 0.1173	223 0.0949	246 0.1168	137 0.1616	96 0.1871	109 0.3169
MA11G	2,551 0.2748	824 0.2641	602 0.2561	570 0.2705	238 0.2807	180 0.3509	137 0.3983
MA33G	2,695 0.2903	894 0.2865	639 0.2718	580 0.2753	246 0.2901	189 0.3684	147 0.4273
FORMAL	2,677 0.2884	890 0.2853	635 0.2701	574 0.2724	247 0.2913	189 0.3684	142 0.4128

[*]1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

The table contains in each cell the absolute (first row) and relative (second row) frequency of successful attempts using Procedure II. The second row contains the relative frequency of correctly matched pairs concerning the number of enterprises contained in the size classes regarding the external data. It can be observed that the distribution of the shares rapidly approaches the corresponding distribution of scenario B1 (last row in table 2) as the degree of anonymisation goes down. The smallest ratios of correct assignment are obtained for enterprises with 50 to 249 employees. We should like to point out here that some caution needs to be exercised in interpreting the results as the distribution may change considerably when the employee size classes are formed differently.

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 for enterprises with at least 1,000 employees. Even in the case of the MA1G variant, about 21 per cent of the large enterprises could be re-identified.

A similar structure of this distribution is attained by an application of Procedure I [between 251 (MA1G) and 1,680 (MA33G) re-identifications overall] and the Hungarian method [between 417 (MA1G) and 2,786 (FORMAL) re-identifications overall].

As expected, the number of re-identifications rose considerably as we passed over from variant MA8G to variant MA11G. That is due to the fact that for the MA8G variant the numerical variable *Total turnover* was micro aggregated in a group containing 12 elements (including variables 8, 9 and 32, see appendix) and thus modified strongly. In variant MA11G *Total turnover* is found in a group of three elements (together with variables 9 and 15, see appendix), in which every two variables correlate with at least 0.92.

Data incompatibilities are a major reason for incorrect matchings. 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 structure of costs survey have been assigned to another branch of economic activity than their respective records of turnover tax statistics. With regard to the variable *Number of employees* there also are significant differences in both surveys. *Total turnover* figures match relatively well. Only some 18.8 % of the enterprises show deviations of more than 10 % in the surveys.

7.2.2 The MARKUS database

The MARKUS database (in German, Marketinguntersuchungen) covers selected enterprises reported on by "Creditreform". It is readily available as a CD-ROM from shops and is published quarterly, with only about 4 % of all enterprises replaced per edition. Generally, the MARKUS database contains enterprises recently examined and not having blocking notes due to insolvency. Therefore, it is not a representative sample of the population. The key variables available are (one additional variable with respect to subsection 7.2.1):

- Branch of economic activity (NACE2, blocking variable)
- Type of administrative district (BBR9, blocking variable)
- Total turnover (matching variable)
- Number of employees (matching variable).

For variant MA8G, the two numerical key variables used were mico aggregated in a common group. This means that smaller differences between the variables were lost. For variant MA11G, the variable *Number of employees* was micro aggregated in a 3-element group (together with variables 5 and 23) like the variable *Total turnover*, so that the values of these two variables were modified to a lesser degree.

In these experiments, the numerical key variables have been weighted with the same value $\lambda_1 = \lambda_2 = 0.5$. Note that a data intruder might prefer the variable *Total turnover* to *Number of employees* if he had knowledge on the data incompatibilities discussed in subsection 7.2.1.

In line with table 2 we get

Table 3. Re-identifications (`MARKUS`*) classified by the number of employees**

target data	total	1	2	3	4	5	6
MA1G	353 0.0376	59 0.0219	35 0.0150	71 0.0309	60 0.0581	53 0.0897	75 0.1667
MA8G	1,845 0.1964	343 0.1274	347 0.1490	503 0.2187	279 0.2703	210 0.3553	163 0.3622
MA11G	2,273 0.2420	419 0.1556	448 0.1924	609 0.2648	355 0.3440	244 0.4129	198 0.4400
MA33G	2,289 0.2437	420 0.1560	443 0.1902	609 0.2648	370 0.3585	246 0.4162	201 0.4467
FORMAL	2,294 0.2442	420 0.1560	442 0.1898	610 0.2652	373 0.3614	247 0.4179	202 0.4489

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

As in subsection 7.2.1 the distribution of the number of successful attempts among the employee size classes does not change significantly using Procedure I and the Hungarian method. In total, by application of Procedure I one obtains between 211 (MA1G) and 1,403 (FORMAL) re-identifications, by application of the Hungarian method between 364 (MA1G) and 2,357 (FORMAL).

Here the difference between variant MA8G to MA11G is not as pronounced as in the preceding experiment. This also holds for the transition from the enterprise size class of 50 – 999 employees to the class containing enterprises with more than 999 employees. The weaker anonymisation variants MA11G, MA33G and FORMAL produce lower hit rates for smaller and medium-sized enterprises (20 to 249 employees) than in the previous experiment. It is somewhat surprising that the hit rate for variant MA8G increased against the previous experiment as there are more pronounced deviations here in both surveys. While the deviation amounting to about 24 % for all enterprises in the classification of economic activities is in line with the preceding experiment 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 surveys.

7.3 Original versus micro aggregated data

In order to get an upper bound for the disclosure risk, the results of the foregoing section are contrasted with the results obtained assuming the worst-case scenario, in which the external database equals the original data without direct identifiers. In the following, we choose several subsets of the numerical variables as matching variables. At first, the whole of 33 numerical variables is used as matching variables (worst-case scenario). We also carry out matching experiments using one matching variable, namely *Total turnover* (in order to contrast the result with the one contained in the realistic scenario in subsection 7.2.1), two matching variables, namely *Total turnover* and *Number of employees* (in order to contrast the result with the one contained in the realistic scenario in subsection 7.2.2), and three matching variables, namely *Total turnover, Number of employees* and *Total intramural R&D expenditure.* The latter variable can be in some cases obtained e.g. via internet searches. As in section 7.2, in all experiments the categorical variables BBR9 and NACE2 were used for blocking the data.

The following table shows the relative frequency of true matches obtained by Procedure II. The first and second entries in each cell refer to the relative and the absolute frequency of true matches obtained using 1, 2, 3 and 33 matching variables.

Table 4. Re-identifications classified by the number of matching variables

micro aggreg. data	33 variables	3 variables	2 variables	1 variable
MA1G	8,941 0.5285	2,156 0.1274	2,076 0.1227	1,096 0.0648
MA8G	16,792 0.9926	12,820 0.7578	11,127 0.6577	3,621 0.2140
MA11G	16,853 0.9962	16,732 0.9890	16,765 0.9910	12,066 0.7132
MA33G	16,918 1.0000	16,918 1.0000	16,912 0.9996	16,757 0.9905
FORMAL	16,918 1.0000	16,918 1.0000	16,918 1.0000	16,918 1.0000

The protection increases notably if the data intruder has only one matching variable available (*Total turnover*) instead of two matching variables (*Total turnover* and *Number of employees*). For the transition from two matching variables to three matching variables only slight differences are observed, in the case of MA11G the hit rate actually decreases. Anyway, the weak protection effect of MA11G, as already observed in subsection 7.2, is confirmed.

As in tables 2 and 3, we consider the distribution of the frequency of re-identifications among the employee size classes, starting with the one matching variable experiment:

Table 5. Re-identifications using one matching variable classified by the number of employees*

target data	total	1	2	3	4	5	6
MA1G	1,096	243	161	164	151	145	232
	0.0648	0.0459	0.0391	0.0420	0.0859	0.1336	0.3069
MA8G	3,621	1,043	681	765	417	354	361
	0.2140	0.1970	0.1653	0.1959	0.2372	0.3263	0.4775
MA11G	12,066	3,841	2,852	2,706	1,252	800	615
	0.7132	0.7255	0.6924	0.6928	0.7122	0.7373	0.8135
MA33G	16,757	5,236	4,084	3,873	1,741	1,078	745
	0.9905	0.9890	0.9915	0.9916	0.9903	0.9935	0.9854
FORMAL	16,918	5,294	4,119	3,906	1,758	1,085	756
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

The increase in the hit rate is quite marked for the transition from MA8G to MA11G. The results of Table 5 may be related to the results of the realistic scenario in subsection 7.2.1 (Table 2) as the same common variables were available in the additional knowledge (Turnover tax statistics) used there. For Procedure I the number of re-identifications is between 711 (MA1G) and 15,707 (MA33G) overall. An application of the Hungarian method yields between 1129 (MA1G) and 16,834 (MA33G) re-identifications overall.

Dually to Table 5 we obtain the distribution of re-identifications among the employee size classes concerning the experiment with two matching variables:

Table 6. Re-identifications using two matching variables classified by the number of employees*

target data	total	1	2	3	4	5	6
MA1G	2,076	394	344	420	311	275	332
	0.1227	0.0744	0.0835	0.1020	0.0796	0.1564	0.3060
MA8G	11,127	3,344	2,610	2,578	1,206	769	620
	0.6577	0.6317	0.6336	0.6600	0.6860	0.7088	0.8201
MA11G	16,765	5,237	4,076	3,879	1,746	1,079	748
	0.9910	0.9892	0.9896	0.9931	0.9932	0.9945	0.9894
MA33G	16,912	5,294	4,117	3,906	1,756	1,085	754
	0.9996	1.0000	0.9995	1.0000	0.9989	1.0000	0.9974
FORMAL	16,918	5,294	4,119	3,906	1,758	1,085	756
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

For Procedure I the number of re-identifications is between 1,674 (MA1G) and 15,721 (MA33G) overall. An application of the Hungarian method yields in total between 2,136 (MA1G) and 16,912 (MA33G) re-identifications.

Regarding MA11G and MA33G, it is observed that the percentage of true matches in the class of 1,000 or more employees is actually recessive. As in the experiment with the `MARKUS` database (see Table 3), there is a pronounced increase in the hit rate between MA1G and MA8G, while the difference between MA11G and MA33G almost seems to be negligible.

8 A standard for de facto anonymity

Two approaches seem reasonable for evaluating the protection of confidential data. There is on the one hand the calculation of the relative frequency of true matchings at the level of the units observed and on the other hand the computation of useful, successfully matched micro data at variable level (it is possible that a data intruder finds a value he can use although units were matched which do not correspond). In order to obtain a standard for de facto anonymity making more concessions to the target audience of business microdata (mainly researchers of economic and statistical sciences), we suggest a combination of both approaches. More precisely, we give an estimate for the disclosure risk of some confidential data taking into account both the percentage of re-identified records and the associated percentage of useful information gained by the data intruder. If this estimated value is below a previously specified threshold the data is then defined to be de facto anonymous. In chapter 7, the percentage of re-identifications was calculated for several simulations. We are going to continue with that example.

8.1 Benefit from re-identification

Even the successful assignment of a unit may be a fruitless disclosure attempt, i.e. when the interesting individual value (or the interesting individual information) deviates by a pre-determined threshold value γ from the actual original value.

We introduce thresholds γ_i for each variable v_i. Let $r^{(i)}$ be the value of variable v_i of some considered record r of the anonymised data and $o^{(i)}$ its corresponding original value. $r^{(i)}$ is said to provide useful information to the data intruder if

$$\tilde{r}^{(i)} := \frac{|o^{(i)} - r^{(i)}|}{|o^{(i)}|} < \gamma_i \tag{5}$$

holds. That is, the relative deviation from the original value is below some thres- hold $\gamma_i > 0$. In the following, a common threshold γ is used for all numerical variables to make the concept presented more easy to handle.

8.2 Disclosure risk

Let R be the set of re-identified units, $r = (r^{(1)}, \ldots, r^{(n)})$ a record of the target data and $o = (o^{(1)}, \ldots, o^{(n)})$ the corresponding unit of the original data. A value $o^{(i)}$ is said to be disclosed if both, $\tilde{r}^{(i)} < \gamma$ and $r \in R$, hold for some predefined γ. That is, we have to estimate the probability $P_\gamma(o^{(i)}\,\text{disclosed}) := P(\tilde{r}^{(i)} < \gamma \text{ and } r \in R)$.

As an estimate for the probability $P(r \in R)$ we set the percentage $\hat{P}(r \in R)$ of true matches (records which have been re-identified by the matching algorithm). The re-identified records of the target data are now tested to contribute useful information in the following way. For each true match and all numerical variables (including matching variables), the percentage of values fulfilling the inequality (5) for some pregiven threshold $\gamma > 0$ is an estimate for the conditional probability $P(\tilde{r}^{(i)} < \gamma \,|\, r \in R)$ and denoted by $\hat{P}(\tilde{r}^{(i)} < \gamma \,|\, r \in R)$. Thus, as an estimate for the disclosure risk associated with the anonymised microdata file on trial we set

$$\hat{P}_\gamma(o^{(i)}\,\text{disclosed}) := \hat{P}(r \in R) \cdot \hat{P}(\tilde{r}^{(i)} < \gamma \,|\, r \in R). \tag{6}$$

In table 8 the concept is applied to the experiments in subsection 7.3 with two matching variables. The first entry in each cell corresponds to the value $\hat{P}_\gamma(o^{(i)}\,\text{disclosed})$, the second to the value $\hat{P}(r \in R)$ and the third to $\hat{P}(\tilde{r}^{(i)} < \gamma \,|\, r \in R)$. If no threshold γ has been defined, the estimator of the risk of disclosing useful values is reduced to the share of correctly matched units (see table 8, column $\gamma = \infty$).

Table 8. Disclosure risk on a γ-level using two matching variables (scenario A)

target data\γ	∞	0.001	0.005	0.01	0.05	0.1	0.2
MA1G	0.1227	0.0246 0.1227 0.2002	0.0254 0.1227 0.2072	0.0266 0.1227 0.2168	**0.0352** 0.1227 0.2865	0.0446 0.1227 0.3636	0.0603 0.1227 0.4913
MA8G	0.6577	0.1858 0.6577 0.2825	0.2088 0.6577 0.3175	0.2326 0.6577 0.3536	**0.3790** 0.6577 0.5762	0.4851 0.6577 0.7376	0.5683 0.6577 0.8642
MA11G	0.9910	0.3226 0.9910 0.3255	0.4546 0.9910 0.4587	0.5565 0.9910 0.5616	**0.8094** 0.9910 0.8167	0.8815 0.9910 0.8895	0.9291 0.9910 0.9376
MA33G	0.9996	0.8908 0.9996 0.8911	0.9811 0.9996 0.9815	0.9904 0.9996 0.9908	**0.9976** 0.9996 0.9980	0.9985 0.9996 0.9989	0.9990 0.9996 0.9994

As an example, the disclosure risks related with the anonymisation variants at the $\gamma = 0.05$ level were printed in bold letters. Especially in the stronger variants of micro aggregation, the marked change in the original values is accounted for by the benefit it provides. For instance, a hit rate of 65.8 % contrasts with a disclosure risk of 37.9 % for variant MA8G.

As variant MA33G modifies the original data only slightly, the suggested concept will be of little effect here.

Note that from the point of view of a risk-adverse data intruder, the probability complementary to (6) is decisive:

$$\begin{aligned} P_\gamma(o^{(i)} \text{ not disclosed}) &= 1 - P_\gamma(o^{(i)} \text{ disclosed}) \\ &= P(r \notin R) + P(\tilde{r}^{(i)} \geq \gamma \,|\, r \in R). \end{aligned}$$

For instance, in variant MA8G with the threshold of $\gamma = 0.05$, he would be put off by knowledge of the estimated probability of 62.1 % of finding a useless value.

8.3 De facto anonymity

To be able to determine a threshold for de facto anonymity, scenarios B1 and B2, in the following taken together as B, should also be considered since the actual risk of disclosing useful information was overestimated considerably on the sole basis of worst-case scenario A. A possible approach would be to estimate the probability (6) also for scenario B (where formally anonymised data are analysed, scenario B2, otherwise B1) and to define a convex combination

$$\hat{P}_\gamma(o^{(i)}\text{disclosed}) := \lambda \cdot \hat{P}_{A/\gamma}(o^{(i)}\text{disclosed}) + (1-\lambda) \cdot \hat{P}_{B/\gamma}(o^{(i)}\text{disclosed}) \quad (7)$$

where the position parameter $\lambda \in [0, 1]$ has to be set individually, depending on the structure of the confidential data and the quality of the data provider's additional knowledge. If $\lambda = 1$, almost completely anonymised data material would be produced provided that λ was determined with caution. If $\lambda = 0$, the data provider would have unlimited confidence in the realistic simulations performed and be certain that a potential data intruder could not have additional knowledge of a better quality.

With the aim of obtaining a reasonable estimator $\hat{P}_{B/\gamma}(o^{(i)}\text{disclosed})$, the realistic scenario should be repeated quite often in practice (for additional knowledge from different sources). We analogously carry out the calculations in table 8 for the two realistic scenarios in 7.2 and contrast the results with the worst-case scenario for $\gamma = 0.05$ in table 9 and figure 5.

Table 9. Disclosure risk on a 0.05 *level depending on three experiments*

target data \ external data	Original	TTS	MARKUS
MA1G	0.035	0.017	0.013
MA8G	0.379	0.072	0.108
MA11G	0.809	0.225	0.194
MA33G	0.997	0.290	0.243
FORMAL	1.000	0.288	0.244

As a general rule, there are areas at risk within the data material which should receive special protection. Therefore we strongly recommend that the concept is applied also to parts. After breaking down the results by size classes of employees, we obtain the disclosure risks listed in table 10 and illustrated in figure 6 for a threshold of usefulness $\gamma = 0.05$ and

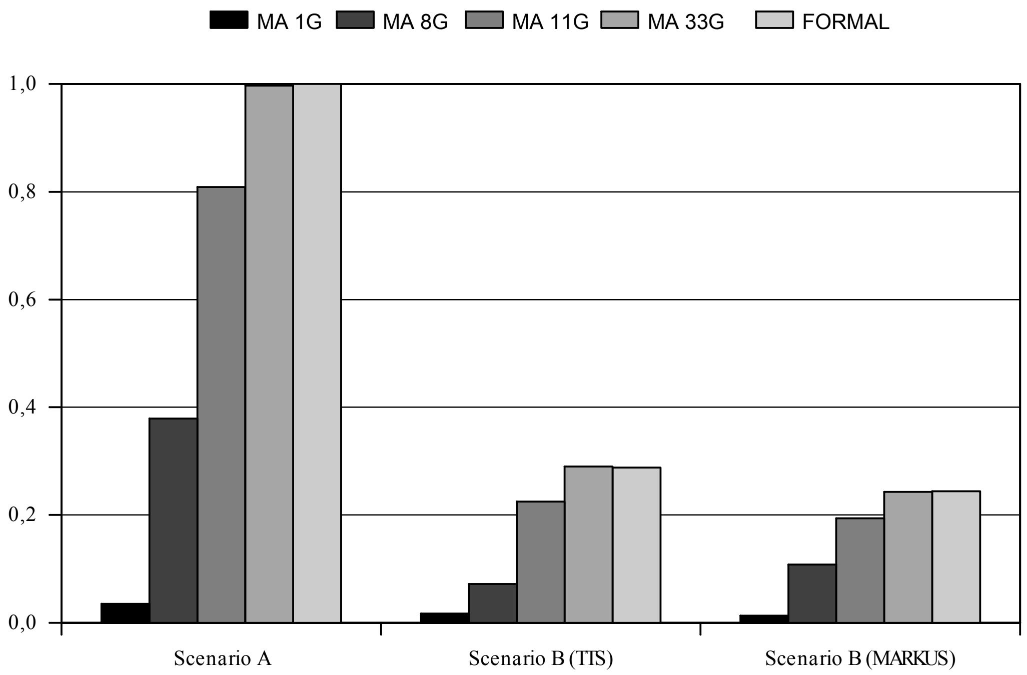

Fig. 5. Disclosure risk on a 0.05 *level depending on three experiments*

a position parameter $\lambda = 0.2$ using formula (7). As an estimator for $P_{B/\gamma}(o^{(i)}$ disclosed) in table 10, the arithmetic mean of the respective estimator was calculated via scenarios using the `MARKUS` data and the turnover tax statistics as external data.

Table 10. Aggregated disclosure risks at the 0.05 *level for size classes of employees**

target data	total	1	2	3	4	5	6
MA1G	0.0191	0.0094	0.0064	0.0080	0.0174	0.0214	0.0366
MA8G	0.1278	0.1156	0.1112	0.1282	0.1426	0.1350	0.1464
MA11G	0.3474	0.3178	0.3194	0.3364	0.3426	0.3704	0.3528
MA33G	0.4130	0.3756	0.3840	0.4150	0.4611	0.5146	0.5447
FORMAL	0.4127	0.3744	0.3840	0.4136	0.4592	0.5104	0.5464

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

When, for instance, the MA8G anonymisation on trial at the $\gamma = 0.05$ level is examined, the value of 0.3790 printed in bold letters in table 8 – a first risk-averse approximation to the disclosure risk of useful information – can thus be undercut considerably with 0.1278.

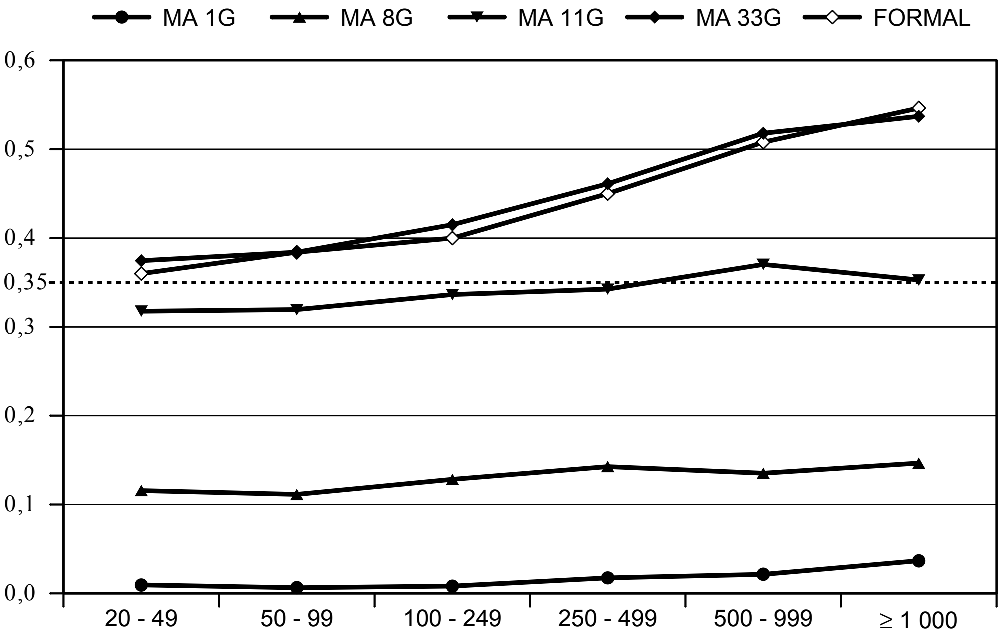

Fig. 6. Aggregated disclosure risks at the 0.05 *level for size classes of employees*

As the number of employees rises, the share of correctly assigned enterprises increases on the one hand, while on the other the share of useful information in the records of these enterprises falls, so that via the product set up in (6) the observed rise in the hit rate for large enterprises (see tables 2, 3, 5 and 6) is slowed down. This effect is particularly obvious with variant MA11G. Here the disclosure risk remains rather constant across the size classes of employees, and for enterprises with at least 1000 employees it is even below that of enterprises with 500 to 999 employees, which is due also to the particularly strong effect of the micro aggregation procedures in intervals containing only small numbers of values for the variables to be anonymised. If the estimated disclosure risk is below some balanced second threshold τ, that is,

$$P_\gamma(o^{(i)}\text{disclosed}) < \tau,$$

the tested microdata file is defined to be de facto anonymous. At this stage, the data distributing institution has to select the anonymisation variant with the best analytical validity among the candidates being de facto anonymous. It is now the task of the data distributing institutions to discuss suitable thresholds γ and τ. For a risk threshold of 35 per cent ($\tau = 0.35$), in our example the MA1G and MA8G anonymisations on trial at the $\gamma = 0.05$ level could be regarded as de facto anonymous, while the units in the upper size classes of employees would still have to be modified slightly for variant MA11G.

9 Aims and scope

On the whole, the general conjecture is confirmed that larger enterprises are easier to re-identify than smaller ones. In this context, it is fortunately observed that the method of micro aggregation is more effective for very large enterprises (above a certain total number of employees). In our application, this inflecting point is lowest for the method of micro aggregation by 11 groups.

From the theoretical viewpoint, experiments drawing upon additional knowledge taken from reality always have to be regarded as exemplary. The data distributing institution can never be 100 per cent sure that a potential data intruder does not have better additional knowledge at his disposal than the one used for simulation. In order to make concessions to the data users the present paper proposes an approach accounting for both, scenarios using available databases for potential data intruders and the worst-case scenario matching the original data against the anonymised data in order to determine an upper bound for the disclosure risk associated with the anonymised data.

Regarding the anonymisations considered, the number of true matches obtained by Procedure II is – taking the previously calculated distances as a basis – very near the optimum solution. Thereby, for the near future, Procedure II, despite its obviously bad worst-case performance, seems to be a real alternative to more elaborated methods of optimisation.

To handle the matching algorithm, first of all the areas at risk have to be identified within the data material. In the present paper, size classes of employees have been examined for that purpose. Furthermore, the analyses have shown that some economic sectors (rows in table 1) are more insecure than other sectors and require particularly confidential treatment. Here it seems necessary that branches of economic activity containing only a small number of values are excluded or aggregated further. In general, the following holds: The coarsening or exclusion of categorical variables contributes considerably to anonymising and, provided that the scientist can do without the information thus lost, makes it possible to modify the numerical variables to a smaller extent. It has turned out, for example, that coarsening the BBR9 code leads to a marked reduction of the disclosure risk calculated in section 8. Here we have the case mentioned in subsection 3.2, namely that on the one hand the number of mismatches within the blocks is reduced through coarsening, while on the other very large blocks are created making it much more difficult to find true matches.

In the present paper we have studied different degrees of anonymisation of the confidential data in order to balance the two main objectives "minimisation of the disclosure risk" and "maximisation of the analytical validity". Nevertheless, the investigations have to be continued by testing other anonymisation methods with less influence on the structure of data. Currently in progress are investigations on the disclosure protection and analytical validity of data modified by multiplicative and additive noise (Ronning 2004) and different variants of resampling (Gottschalk 2004).

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 Anonymisation of Business Microdata.

Appendix: The German Structure of Costs Survey

The following variables are available in the German structure of costs survey. More information about the survey can be found in Statistisches Bundesamt (2005).

1. Branch of economic activity (NACE - Classification of Economic Activities)
2. Type of administrative district (BBR[2]9 code – so-called category 9)
3. Size class of employees
4. Working proprietors
5. Employees (salary and wage earners)
6. Part-time employees
7. Part-time employees in full-time equivalent units
8. Total of active persons
9. Turnover of the unit's own products
10. Turnover of goods for resale
11. Total turnover (does not correspond to the sum of items 9 and 10)
12. Initial stocks of work in progress and finished products manufactured by the unit measured against turnover of the unit's own products
13. Final stocks of work in progress and finished products manufactured by the unit measured against turnover of the unit's own products
14. Change in stocks of work in progress and finished products
15. Gross output/production value
16. Initial stocks of raw materials and other intermediary products purchased and consumables, measured against turnover of the unit's own products
17. Final stocks of raw materials and other intermediary products purchased and consumables, measured against turnover of the unit's own products
18. Consumption of raw materials
19. Energy consumption
20. Initial stocks of goods for resale measured against turnover of goods for resale
21. Final stocks of goods for resale measured against turnover of goods for resale
22. Input of goods for resale
23. Wages and salaries
24. Statutory social security costs
25. Other social security costs
26. Payments for agency workers
27. Costs of contract processing
28. Repair costs
29. Renting and leasing
30. Other costs
31. Interest on borrowed capital
32. Total costs
33. Value-added at factor cost

[2] Federal Agency for Construction and Regional Planning

34. Net value-added at factor cost
35. Total intramural R&D expenditure
36. Total number of R&D personnel

References

Bertsekas, D. P. (1979). A distributed algorithm for the assignment problem. Lab. for Information and Decision Systems, Working Paper, MIT Press.

Bertsekas, D. P. and Castanon, D. A. (1989). The auction algorithm for transportation problems. Annals of Operations Research 20, 67-96.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (1990). Introduction to Algorithms. Cambridge, Mass., MIT Press.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1971). Maximum Likelihood From Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society B, 39, 1-38.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 39 (1), 189-201.

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

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata. Journal of the Royal Statistical Society B, (4) 64, 855-867.

Evers, K. and Höhne, J. (1999). SAFE - A procedure to anonymise and safeguard the statistical confidentiality of microdata of economic statistics (German). Spektrum der Bundesstatistik, 14, 136-147.

Fellegi, I. P. and Sunter, A. P. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183-1210.

Efelky, M., Verykios, V. and Elmagarmid, A. (2002). TAILOR: A Record Linkage Toolbox. Proc. of the 18^{th} Int Conf. on Data Engineering, San Jose, California.

Gottschalk, S. (2004). Microdata disclosure by resampling – empirical findings for business survey data. Journal of the German Statistical Society, (3) 88, 279-302.

Hernandez, M. and Stolfo, S. (1998). Real World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Journal of Data Mining and Knowledge Discovery, 2 (1), 9-37.

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

Kadane, J. B. (2001). Some Statistical Problems in Merging Data Files. Journal of Official Statistics, 17 (3), 423-433.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. Naval Res. Logist. Quart., 2, 83-97.

Lenz, R. (2003). A graph theoretical approach to record linkage. Mono- graphs of Official Statistics – Research in Official Statistics, Eurostat, 324-334.

Lenz, R., Doherr, T. and Vorgrimler, D. (2004). Simulation of a database cross match – as applied to the German structure of costs survey. Proc. of the European Conference on Quality and Methodology in Official Statistical (Q2004), ISBN 3-8246-0733-6.

Munkres, J. (1957). Algorithms for the assignment and transportation problem. L. Soc. Indust. Appl. Math., 5, 32-38.

Papadimitriou, C. H. and Steiglitz, K. (1998). Combinatorial Optimization. Dover Publ., Mineola, New York.

Porter, E. H. and Winkler, W. E. (1999). Approximate String Comparison and its Effect on an Advanced Record Linkage System. Record Linkage Techniques – 1997. National Academy Press, Washington DC, 190-199.

Ronning, G. (2004). Mixtures Distributions and Masking of Data (Roque and the alternative approach by Yancey, Winkler and Creecy). Working paper of the project team De Facto Anonymisation of Business Microdata.

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

Rosemann, M., Vorgrimler, D. and Lenz, R. (2004). First results of de facto anonymising micro data of economic statistics (German). Journal of the German Statistical Society, (1) 88, 73-99.

Statistisches Bundesamt (2005): Kostenstruktur im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden (German).
See http://www.destatis.de/download/qualitaetsberichte/

Schweigert, D. (1995). Vector Weighted Matchings. Combinatorial Advances (eds. C.J.Colbourn, E.S.Mahmoodian), Kluwer, 267-276.

Schweigert, D. (1999). Order and Clustering. Human Centered Processes, 10^{th} Mini Euro Conference, Brest, 397-401.

Sturm, R. (2002). De facto anonymisation of micro data of economic statistics (German). Journal of the German Statistical Society, (4) 86, 468-477.

Vorgrimler, D. and Lenz, R. (2003). Disclosure risk of anonymized business microdata files – illustrated with empirical key variables. Proceedings of the 54^{th} session of the International Statistical Institute, 2, 594-595.

Yancey, W. E., Winkler, W. E. and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. Inference Control in Statistical Databases, Springer, 135-152.

Zaslavsky, A. M. and Horton, N. J. (1998). Balancing disclosure risk against the loss of nonpublication. Journal of Official Statistics, 14, 411-419.