WP. 23
ENGLISH ONLY

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

**EUROPEAN COMMISSION**
**STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**
(Geneva, Switzerland, 9-11 November 2005)

Topic (iii) Confidentiality aspects of statistical information taking into account register-based data

# A RANKING APPROACH TO CONFIDENTIALITY IN SURVEY DATA

**Supporting Paper**

Submitted by Statistics Norway[1]

[1] Prepared by Johan Heldal (johan.heldal@ssb.no).

# A ranking approach to confidentiality in survey data

Johan Heldal[1]

[1] **Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway**

**Abstract.** This paper suggests a method for confidentiality protection of datasets when (continuous) numeric identifying variables have been linked to the dataset by exact matching with registers. Such variables may be highly identifying, in particular if the register is publicly available. The method, in this paper called *rank matching*, takes advantage of the registers themselves to mask the original data and eliminate the confidentiality hazard.

**Keywords.** *Superpopulation distribution, information loss, density estimation, rank matching, disclosure, re-identification probability.*

## 1. Introduction

In some countries, the national statistical office includes some variables in surveys not by asking the respondents, but by exact matching of information from registers files comprising the entire population. Such data, for instance income from tax registers, are often of high quality and may be of high value both for researchers and for an intruder trying to disclose the identity of a statistical unit, in particular if the register itself is available to him or her. Data from registers has raised concern in the context of providing anonymous datasets to researchers under EU Regulation 831/2002. The EU-SILC Anonymisation Task Force Report (Museux 2005) writes:

> *"The TF has pointed the specificities of so called register countries. For these countries, some of the income variables available in the EU-SILC may come directly from registers (DK, NO, SE, FI, LT, LI, CZ, SI, IS). If this register information together with direct identifiers is available to external users, the risk of disclosure is greatly increased. This specific issue should be carefully studied. A specific section of this report is dedicated to it. "*

This is however also a situation that opens an opportunity to apply disclosure control methods which are not otherwise available.

This paper is basically a representation of an idea called *rank matching* (rm) earlier presented in Fosen and Heldal (2001) and Heldal (2001), an idea that Statistics Norway now wishes to follow up in the context of EU-SILC. Carlson and Salabasis (2002) have (independently) worked on the same idea and in greater detail using theory of order statistics. Because of space constraints I refer to these papers for study of the statistical properties of the method from a user viewpoint. This paper will concentrate on some intruder scenarios associated with the method.

Section 2 outlines the ideas behind rank matching. In section 3 simulations and small examples are used to discuss some intruder scenarios. More work is needed to study the scenarios in more realistic settings.

## 2. Basic ideas

Consider a finite population $\mathcal{U}$ consisting of units $u_1,\ldots,u_N$ indexed by a variable $j$. To each unit a vector $\boldsymbol{X}_j^T = (X_{j1},\ldots,X_{jK})$ of absolutely continuous numeric variables is attached that for the moment can be considered as generated by a (cumulative) superpopulation distribution $F(\boldsymbol{x})$. The $N\times K$ matrix $\mathbb{X} = (\boldsymbol{X}_j^T, j\in\mathcal{U})$ is termed a *register*. Examples of numeric register variables are income from the tax assessment and age of individuals. Turnover or other economic variables stored in business registers are other examples.

A sample $\boldsymbol{s}$ of size $n$ is drawn from the finite population $\mathcal{U}$ with some sampling design $p(\boldsymbol{s})$. $\boldsymbol{s}$ gives rise to a dataset matrix $\mathbf{X} = (\boldsymbol{X}_j^T, j\in\boldsymbol{s})$. The joint non-singular density of $\boldsymbol{X}_j$ is called $f(\boldsymbol{x})$. To keep concepts as simple as possible, assume that $p(\boldsymbol{s})$ is simple random so that the $\boldsymbol{X}_j$'s are identically distributed also in the sample.

Let $R_{jk}$ be the rank of the observed value $X_{jk}$ in the $j$-th column of $\mathbb{X}$ where $R_{jk}\in\{1, \ldots, N\}$. Further, let $\boldsymbol{R}_j^T = (R_{j1},\ldots,R_{jK})$ and $\mathbb{R} = (\boldsymbol{R}_j^T, j\in\mathcal{U})$ the $N\times K$ rank matrix corresponding to the register $\mathbb{X}$. Let $i$ index the sample units and $j_i$ (stochastic) be the population index (label) of sample unit $i$ and let $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ik}) = \boldsymbol{X}_{j_i}$. Let $r_{jk}\in\{1, \ldots, n\}$ be the rank of $x_{ik}$ in $\mathbf{X}$, $\boldsymbol{r}_i = (r_{i1},\ldots,r_{iK})^T$ and let $\mathbf{R} = (\boldsymbol{r}_1^T,\ldots,\boldsymbol{r}_n^T)$ be the sample rank matrix. The latter should be distinguished from $\mathbb{R}_{\boldsymbol{s}} = (\boldsymbol{R}_j^T, j\in\boldsymbol{s})$ which contains the population ranks for the sample units. The continuity assumption guarantees uniqueness of the ranks. For the mapping from the ranks to the labels the (somewhat simplified) notation $i = (r_{ik})$ and $j = (R_{jk})$ is being used, $k=1, \ldots, K$.

*Rank matching* (rm) now goes as follows: Draw a new sample $\boldsymbol{s}_2$ independently of $\boldsymbol{s}$, according to the same design and sample size as $\boldsymbol{s}$. $\boldsymbol{s}_2$ gives rise to a new sample dataset $\mathbf{X}^{(2)} = (\boldsymbol{X}_j^T, j\in\boldsymbol{s}_2) = (x_1^{(2)T},\ldots,x_n^{(2)T})$ with the same variables as before and generated by the same superpopulation distribution $F$. Replace $x_{ik}(= x_{(R_{ik})k})$ in the original sample with the value $x_{(R_{ik})k}^{(2)}$ having the same rank on the same variable in $\mathbf{X}^{(2)}$. This produces a synthetic dataset $\mathbf{X}^*$ with rows $x_i^{*T} = \boldsymbol{x}_{(\boldsymbol{R}_i)k}^{(2)T} = (x_{(R_{i1})1}^{(2)},\ldots,x_{(R_{iK})K}^{(2)})$. This version will be called *joint* rm. The distribution of $\boldsymbol{x}_i^*$ will depend on the original $\boldsymbol{x}_i$ only through its rank vector $\boldsymbol{r}_i$. The method can be applied for an entire sample or within strata or domain. It is also possible to draw one sample for each variable. This is computationally more intensive, but analytically somewhat simpler to deal with. This will be called *independent* rank matching. It will however not matter much from the point of view of the intruder (see section 3).

Generally there is information loss associated with rm. In the original dataset, the marginal distribution of $x_{ik}$ given $\boldsymbol{r}_i$ can depend on components of $\boldsymbol{r}_i$ other than $r_{ik}$. In other words, for independent rm,

$$F_k^*(x_k \mid \boldsymbol{r}) = F_k^*(x_k \mid r_k) = F_k(x_k \mid r_k) \neq F_k(x_k \mid \boldsymbol{r})$$

For joint rm the first equality will not be exact.

An option to rank matching with register is rank *swapping* (rs) and related methods (Moore 1996). Contrary to rm, rs preserves exactly the observed marginal distributions of all variables as in the original dataset, but not the exact multivariate rank order structure. Simulations and theoretical considerations in Carlson and Salabasis (2002) based on correlations between normally distributed variables indicate that attenuation due to independent rm is slightly larger than due to joint rm. The attenuation is larger for smaller sample sizes than for big ones. A proper comparison to various versions of rs remains, but Heldal (2001) indicates that simple half sample rs is inferior.

A similar approach can be attempted on discrete ordinal variables. Then an artificial ordering must be introduced between units having the same value on the discrete variable. Care must be taken to avoid illegal edits. Some variables, and income components in particular, take both discrete and continuous values. For many income components, there are typically many zero values and otherwise positive values. Improper ordering of the original zeroes in **X** can easily introduce positive values on units that according to the values of other variables cannot be positive, like giving a 20 year old man retirement pension. Detailed discussions about how to handle discrete values will not be given here.

In most cases, the variables available from registers only make up some of the variables in a survey dataset. Non-register variables, usually collected in the survey are not affected by the register rank matching, but may be target variables for a disclosure attempt.

## 3. Inference about population units

While information loss should be considered at superpopulation level, probability of disclosure is definitely a finite population matter. The samples $\boldsymbol{s}$ and $\boldsymbol{s}_2$ are (simple random) samples from a labelled set of units $\mathcal{U}$ to which realisations from the population distribution $F$ have been associated. Identity disclosure is inference about the label in this finite population. Such inference is possible only when someone with access to the dataset **X** has information about some of these variable values associated to given labels.

It is clear that an intruder having accurate information about the value of at least one absolutely continuous numeric variable for some unit drawn to the sample will be able to identify that unit. If intruder's information or the measured values of the variables in the sample is not quite accurate, inference about a label can never the less very often be done with high degree of confidence.

**Question 1:** Which information on labels associated with the units in $\boldsymbol{s}$ is still present in the rank matched dataset $\mathbf{X}^*$?
**Question 2: H**ow can an intruder make use of this information to make a disclosure?

The answer to these questions will depend on the intruder scenario. Two worst-case scenarios will be discussed:

a. The intruder knows that some members in her Identification File (IF) are in $\boldsymbol{s}$ and their true values on some $X_{ij}$.
b. The intruder has access to the entire population register, but does not know which units were drawn to $\boldsymbol{s}$.

Case a will be studied in a simulation experiment presented in section 3.1. This shows that with an increasing number of variables available for disclosure the probability of doing correct identification using distance techniques increases rapidly. Case b will be illustrated with an example in section 3.2. This is an extreme case, but is interesting. Someone having access to the entire register can extract its rank structure (population ranks) and from that identify all possible samples whose sample rank matrix $\mathbf{R}$ equal the rank matrix of $\mathbf{X}$. On such a basis the probability that an individual with a given sample rank vector corresponds to a given population unit can be computed exactly for every records in the sample.

### 3.1. Situation a, a simulated intrusion

With what confidence can an intruder identify the original record number associated with the synthetic record $\boldsymbol{x}^*$? Assume that the intruder in her identification file has access to an original record $\boldsymbol{x}$ from $\mathbf{X}$ and knows that the owner of $\boldsymbol{x}$ is in $\mathbf{X}$. To disclose the corresponding record in $\mathbf{X}^*$ (and $\mathbf{X}^+$), she uses discriminant analysis and decides for the following decision rule: Choose the record $\boldsymbol{x}_i^*$ in $\mathbf{X}^*$ that minimizes a distance

$$\left\| \boldsymbol{x} - \boldsymbol{x}_i^* \right\|_{\mathbf{W}}^2 = (\boldsymbol{x} - \boldsymbol{x}_i^*)' \mathbf{W} (\boldsymbol{x} - \boldsymbol{x}_i^*) . \qquad (3.1)$$

A thorough discussion of the use of discriminant analysis in the context of disclosure control is given in Paaß and Wauschkuhn (1985). In order to test the capacity of this decision rule, $\mathbf{W}$ was taken as the inverse of the diagonal of $\hat{\boldsymbol{\Sigma}}^*$ and $\hat{\boldsymbol{\Sigma}}^+$, the obvious

estimates of the covariance matrices based on $\mathbf{X}^*$ and $\mathbf{X}^+$. All 63 possible combinations of one to six variables were tested and the number of correct hits recorded. The results are summarized in table 1.

| The number of variables used | Number of correct hits | | The number of variables used | The number of variables used | |
|---|---|---|---|---|---|
| One (of 6 variables) | rm | 6-41 | Four (15 combs.) | rm | 845-989 |
| | rs | 0 | | rs | 722-945 |
| Two (of 15 pairs) | rm | 137-545 | Five (6 combs.) | rm | 983-996 |
| | rs | 93-321 | | rs | 924-981 |
| Three (20 triples) | rm | 472-933 | Six (1 comb.) | rm | 996 |
| | rs | 244-720 | | rs | 987 |

**Table 1.** Minimum and maximum numbers of correct identifications of records in $\mathbf{X}^*$ (rm rows) and $\mathbf{X}^+$ (rs rows) with various numbers of identification variables.

Table 2 shows that the identifying capacity of combinations of variables increases rapidly with the number of variables available for disclosure for both methods. This is no surprise. The number of correct identifications with the same number of variables shows large variations. The tendency is, as expected, that among the combinations with the same number of variables, those showing higher correlations produce the smallest number of correct hits and vice versa. The results in table 2 may seem discouraging. But this was for an intruder knowing that the target is there. An intruder not knowing that the target unit is in the dataset will need to verify that. For some discussion of that case, see Heldal (2001)

### 3.2. Case b.

Consider an intruder with access to a population register $\mathbb{X}$ described in the beginning of section 2. This intruder can extract the population rank matrix $\mathbb{R} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)^T = \rho(\mathbb{X})$ from $\mathbb{X}$. Without loss of generality we can take the ranks in the first column of $\mathbb{R}$ and $\mathbf{R}$ as population and sample labels, setting $R_{j1} = j$ and $r_{i1} = i$. Let $j_i$ be the stochastic variable that maps sample label $i$ to a population label. The intruder observes $\mathbf{X}^*$ and $\mathbb{X}$ and wishes to calculate $P(j_i = j \mid \mathbf{X}^*, \mathbb{X})$ for all $i$ and all $j \in \mathcal{U}$. With a little algebra we prove that $\mathbf{R}$ and $\mathbb{R}$ are sufficient for the intruders inference.

$$P(j_i = j \mid \mathbf{X}^*, \mathbb{X}) = P(j_i = j \mid \mathbf{R}, \mathbb{R})$$

The sample version of $\mathbb{R}$, $\mathbb{R}_{\boldsymbol{s}} = (\boldsymbol{R}_j^T,\ j \in \boldsymbol{s})$, will not be directly observable in the sample. Never the less, there is a 1-1 correspondence between the sample space $\boldsymbol{\mathcal{S}}$ and $\{\mathbb{R}_{\boldsymbol{s}} : \boldsymbol{s} \in \boldsymbol{\mathcal{S}}\}$. $\mathbb{R}_{\boldsymbol{s}}$ uniquely determines $\mathbf{R} = \rho(\mathbb{R}_s) = \rho(\mathbf{X}) = \rho(\mathbf{X}^*)$, and the structure of $\mathbb{R}$ determines the probability structure of $\mathbf{R}$. There are $(n!)^{K-1}$ possible (unordered) sample rank matrices $\mathbf{R}$. They define a partition of $\boldsymbol{\mathcal{S}}$ into disjoint subsets $\boldsymbol{\mathcal{S}}_{\mathbf{R}}$, some of which may be empty by the configuration of $\mathbb{R}$. If for an observed matrix $\mathbf{R}$, $\boldsymbol{\mathcal{S}}_{\mathbf{R}}$ is identified, then the probability $P(j_i = j \mid \boldsymbol{\mathcal{S}}_{\mathbf{R}})$ that a given sample unit $i$ corresponds to a given population unit $j$ can be calculated exactly. However, it does not seem to be feasible to do this by formula except when $K = 1$. For large $N$ and $n$ efficient algorithms will be necessary to identify $\boldsymbol{\mathcal{S}}_{\mathbf{R}}$.

***Example*:** Assume $N = 7$, $K = 1$ and $n = 3$. Then $\mathbb{R} = [1, 2, 3, 4, 5, 6, 7]^T$ and $\mathbf{R} = [1, 2, 3]^T$. Then

$$P(j_i = j \mid \boldsymbol{\mathcal{S}}_{\mathbf{R}}) = \binom{j-1}{i-1}\binom{N-j}{n-i} \Big/ \binom{N}{n} = \binom{j-1}{i-1}\binom{7-j}{3-i} \Big/ 35$$

| $i\backslash j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 15/35 | 10/35 | 6/35 | 3/35 | 1/35 | 0 | 0 |
| 2 | 0 | 5/35 | 8/35 | 9/35 | 8/35 | 5/35 | 0 |
| 3 | 0 | 0 | 1/35 | 3/35 | 6/35 | 10/35 | 15/35 |

Table 2: Tabulation of the distribution of $P(j_i=j \mid \mathbf{R}, \mathbb{R})$.

Assume $K = 2$ and that the population rank matrix is

$$\mathbb{R} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 2 & 3 & 1 & 7 & 6 \end{bmatrix}^T,$$

The sample space still consists of 35 samples. Now there are 6 possible sample rank matrices $\mathbf{R}$. The 6 sample rank matrices, their associated partition sets and the probabilities $p(\ j|\ i) = P(\ j_i = j \mid \boldsymbol{\mathcal{S}}_{\mathbf{R}})$ are given in table 3. The table shows large variation of the number of samples in each partition. The cases where $p(\ j \mid i\ ) = 1$ define identity disclosure with probability one. This occurs for at least one unit in eleven samples in three partition subsets, meaning that before sampling the probability of a disclosure producing dataset is 11/35.

| $\mathbf{R}^T$ | $\{\mathbb{R}_s{}^T : s \in \mathcal{S}_\mathbf{R}\}$ | $p(j \mid i) = P(j_i = j \mid \mathbf{R})$ |
|---|---|---|
| $\begin{pmatrix}1&2&3\\1&2&3\end{pmatrix}$ | $\begin{bmatrix}1&2&6\\4&5&7\end{bmatrix},\begin{bmatrix}1&2&7\\4&5&6\end{bmatrix},\begin{bmatrix}3&4&6\\2&3&7\end{bmatrix},\begin{bmatrix}3&4&7\\2&3&6\end{bmatrix}$ | $p(1 \mid 1) = p(3 \mid 1) = 1/2$<br>$p(2 \mid 2) = p(4 \mid 2) = 1/2$<br>$p(6 \mid 3) = p(7 \mid 3) = 1/2$ |
| $\begin{pmatrix}1&\mathit{2}&\mathit{3}\\3&1&2\end{pmatrix}$* | $\begin{bmatrix}1&\mathit{3}&\mathit{4}\\4&2&3\end{bmatrix},\begin{bmatrix}2&\mathit{3}&\mathit{4}\\5&2&3\end{bmatrix}$ | $p(1 \mid 1) = p(2 \mid 1) = 1/2$<br>$p(\mathit{3} \mid \mathit{2}) = p(\mathit{4} \mid \mathit{3}) = \mathbf{1}$ |
| $\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}$ | $\begin{bmatrix}1&2&3\\4&5&2\end{bmatrix},\begin{bmatrix}1&2&4\\4&5&3\end{bmatrix},\begin{bmatrix}1&2&5\\4&5&1\end{bmatrix},\begin{bmatrix}3&4&5\\2&3&1\end{bmatrix}$ | $p(1 \mid 1) = 3/4,\ p(3 \mid 1) = 1/4$<br>$p(2 \mid 2) = 3/4,\ p(4 \mid 2) = 1/4$<br>$p(3 \mid 3) = p(4 \mid 3) = 1/4,\ p(5 \mid 3) = 1/2$ |
| $\begin{pmatrix}1&\mathit{2}&\mathit{3}\\1&3&2\end{pmatrix}$* | $\begin{bmatrix}1&\mathit{6}&\mathit{7}\\4&7&6\end{bmatrix},\begin{bmatrix}2&\mathit{6}&\mathit{7}\\5&7&6\end{bmatrix},\begin{bmatrix}3&\mathit{6}&\mathit{7}\\2&7&6\end{bmatrix},$<br>$\begin{bmatrix}4&\mathit{6}&\mathit{7}\\3&7&6\end{bmatrix},\begin{bmatrix}5&\mathit{6}&\mathit{7}\\1&7&6\end{bmatrix}$ | $p(1 \mid 1) = p(2 \mid 1) = p(3 \mid 1)$<br>$= p(4 \mid 1) = p(5 \mid 1) = 1/5$<br>$p(\mathit{6} \mid \mathit{2}) = p(\mathit{7} \mid \mathit{3}) = \mathbf{1}$ |
| $\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}$ | $\begin{bmatrix}1&3&6\\4&2&7\end{bmatrix},\begin{bmatrix}1&3&7\\4&2&6\end{bmatrix},\begin{bmatrix}1&4&6\\4&3&7\end{bmatrix},\begin{bmatrix}1&4&7\\4&3&6\end{bmatrix}$<br>$\begin{bmatrix}1&5&6\\4&1&7\end{bmatrix},\begin{bmatrix}1&5&7\\4&1&6\end{bmatrix},\begin{bmatrix}2&3&6\\5&2&7\end{bmatrix},\begin{bmatrix}2&3&7\\5&2&6\end{bmatrix}$<br>$\begin{bmatrix}2&4&6\\5&3&7\end{bmatrix},\begin{bmatrix}2&4&7\\5&3&6\end{bmatrix},\begin{bmatrix}2&5&6\\5&1&7\end{bmatrix},\begin{bmatrix}2&5&7\\5&1&6\end{bmatrix}$<br>$\begin{bmatrix}3&5&6\\2&1&7\end{bmatrix},\begin{bmatrix}3&5&7\\2&1&6\end{bmatrix},\begin{bmatrix}4&5&6\\3&1&7\end{bmatrix},\begin{bmatrix}4&5&7\\3&1&6\end{bmatrix}$ | $p(1 \mid 1) = p(2 \mid 1) = 3/8$<br>$p(3 \mid 1) = p(4 \mid 1) = 1/8$<br>$p(3 \mid 2) = p(4 \mid 2) = 1/4$<br>$p(5 \mid 2) = 1/2$<br>$p(6 \mid 3) = p(7 \mid 3) = 1/2$ |
| $\begin{pmatrix}1&2&\mathit{3}\\3&2&1\end{pmatrix}$* | $\begin{bmatrix}1&3&\mathit{5}\\4&2&1\end{bmatrix},\begin{bmatrix}1&4&\mathit{5}\\4&3&1\end{bmatrix},\begin{bmatrix}2&3&\mathit{5}\\5&2&1\end{bmatrix},\begin{bmatrix}2&4&\mathit{5}\\5&3&1\end{bmatrix}$ | $p(1 \mid 1) = p(2 \mid 1) = 1/2$<br>$p(3 \mid 2) = p(4 \mid 2) = 1/2$<br>$p(\mathit{5} \mid \mathit{3}) = \mathbf{1}$ |

**Table 3** The partition of $\mathcal{S}$ generated by the sample rank matrices **R** and the induced identification probabilities and disclosure probabilities given **R**. * marked **R**s generate some certain disclosures, indicated by italic sample and population labels.

## 4. Future work

In 2006 the ideas presented in this paper will be attempted on the Norwegian SILC survey will start up in Statistics Norway. An application of this kind will require further work on the method and may require use of other methods as well. Questions

related to sample design and rm-domains, balance of information loss versus disclosure risk and data integrity must be addressed. What about discrete or mixed mode variables?

We know that the methods suggested in this paper can be relevant for other countries as well and the so-called 'register countries' in particular. We wish to do our work in an international context and we hereby invite workers who may be interested in this kind of problems for collaboration.

## References


Carlsson, M. and Salabasis, M. (2002): *A data-swapping technique using ranks – A method for disclosure control.* Research in official Statistics, Vol. 4 no. 2 pp 35- 67 (with comment by S.E. Fienberg).

Duncan, G. T. and Lambert, D. (1989): *Risk of Disclosure for Microdata.* J. of Business & Economic Statistics, Vol 7., no. 2 pp 207-217

Fuller, W. A. (1993): *Masking Procedures for Microdata Disclosure Limitation.* Journal of Official Statistics, vol 9 no. 2 pp 383-406.

Hurkens, C.A.J. and Tiourine, S.R. (1998*): Models and Methods for the Microdata Protection Problem.* Journal of Official Statistics, 14, pp 437-447.

Little, R.J.A. (1993): *Statistical Analysis of Masked Data*. Journal of Official Statistics, vol 9 no. 2 pp 407-426.

Moore, R. (1995): *Controlled Data Swapping Techniques For Masking Public Use Data Sets,* U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).

Museux, J-M. (2005): *EU-SILC anonymisation: Results of the Eurostat Task Force.* UNECE/Eurostat work session in Confidentiality 2005, WP no. 20.

Paaß, G. and Waushkuhn, U. (1985): Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von Anonymisierten Individualdaten. München: Oldenburg Verlag

Paaß, G. (1988): *Disclosure Risk and Disclosure Avoidance for Microdata*. J. of Business & Economic Statistics, Vol. 6., no. 4 pp 487-500.

Reiss, R.-D. (1989): *Approximate Distributions of Order Statistics.* With applications to Nonparametric Statistics. Springer Verlag.

Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). *Disclosure Control for Census Microdata.* Journal of Official Statistics, 10, pp 31-51.

Strudler, M., Oh, H. L. and Scheuren, F. (1986): *Protection of Taxpayers Confidentiality With Respect to the Tax Model*. Proceedings of the Section on Survey Research Methods, American Statistical Assoc. pp 375-381