A RANKING APPROACH TO CONFIDENTIALITY IN SURVEY DATA

Johan Heldal, Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway

Key words: Superpopulation distribution, information loss, density estimation, rank matching, disclosure, re-identification probability.

1. Introduction

In the statistical offices of some countries some variables are included in surveys not by asking the respondents, but by exact matching of information from registers files comprising the entire population. Such data are often of high quality and may be extremely identifying. This is however also a situation that opens an opportunity to apply methods for disclosure control that are not otherwise accessible. In countries where such matching is not legally or technically feasible, samples for research are taken from censuses or administrative files like income and taxation registers.

Section 2 outlines the basic ideas behind a method that here has been termed *rank matching*. This method takes advantage of the situation where registers are available. This paper will only consider the situation when all relevant variables are continuous. The idea can be extended to discrete variables, but because of space constraints, such extensions will not be discussed here.

Section 3 discusses the effect of rank matching on information loss by analysis of a simulation experiment.

In section 4 the simulation and an example are used to discuss the confidentiality protection offered by rank matching.

Section 5 discusses problems and extensions.

2. Basic ideas

Consider a simple random sample $\boldsymbol{s}$ of size n drawn from the finite population $\mathcal{P}$. $\boldsymbol{s}$ gives rise to a dataset $\mathbf{X}$ with records $\boldsymbol{X}_j^T = (X_{j1},\ldots,X_{jK})$, $j = 1,\ldots,n$ that will here be considered as generated by a (cumulative) nonsingular superpopulation distribution $F(\boldsymbol{x})$ with density $f(\boldsymbol{x})$, which has also generated $\mathcal{P}$.

Add noise to the observed $\boldsymbol{X}_j$ vectors by transforming them randomly to vectors $\boldsymbol{Y}_j$ according to a conditional density $g(\boldsymbol{y} \mid \boldsymbol{x})$, producing a new set of observations with density

$$h(\boldsymbol{y}) = \int g(\boldsymbol{y} \mid \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} . \qquad (2.1)$$

If $g(\boldsymbol{y} \mid \boldsymbol{x})$ represents a transition density for a stochastic process with $f(\boldsymbol{x})$ as a stationary density, then $h(\boldsymbol{y}) = f(\boldsymbol{y})$. In this case the noise addition generates a new sample $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ with the same parent-distribution *F*, being statistically equivalent to the original sample. If *f* is known, a transition density *g* having this property can be constructed by orthogonal expansions. *f* must be an eigenfunction for *g* corresponding to the eigenvalue 1. Gouweleeuw et al. (1998) point at similar ideas in the context of disclosure control for categorical and discrete variables (the PRAM method), but without reference to a superpopulation.

But *f* is rarely known. Therefore, an alternative approach, based on a kind of resampling, is proposed.

Let R_{kj} be the rank of X_{jk} in a sample of size *n* and $\boldsymbol{R}_j = (R_{j1},\ldots,R_{jK})^T$. The continuity assumption guarantees uniqueness of the ranks. To denote rank orders of the observations we write $X_{(1)k} < \cdots < X_{(n)k}$ and $\boldsymbol{X}_j = (X_{j1},\ldots,X_{jK})^T = \boldsymbol{X}_{(\boldsymbol{R}_j)} = (X_{(R_{j1})1},\ldots,X_{(R_{jK})K})^T$. The marginal distribution for variable *k* is denoted by F_k. Let $\{1:n\} = \{1,\ldots,n\}$ and $\{1:n\}^K$ be the *k*-dimensional Cartesian lattice generated by $\{1:n\}$. Furthermore, let $h(\boldsymbol{r}) = P(\boldsymbol{R}_j = \boldsymbol{r})$. All marginal distributions h_k of *h* are uniform on $\{1:n\}$, but the joint distribution *h* will depend on *F*. We can write

$$F(\boldsymbol{x}) = \sum_{\boldsymbol{r} \in \{1:n\}^K} F(\boldsymbol{x} \mid \boldsymbol{r}) h(\boldsymbol{r}) \qquad (2.2)$$

For the method that will here be called *rank matching* (rm) *with registers*, there are now two versions.

I. Draw a new sample $\boldsymbol{s}_2$ according to the same design with the same sample size as $\boldsymbol{s}$. $\boldsymbol{s}_2$ gives rise to a new dataset $\mathbf{X}^{(2)}$ with records $\boldsymbol{X}^{(2)}_{j^{(2)}} = (X^{(2)}_{j^{(2)}1},\ldots,X^{(2)}_{j^{(2)}K})^T$, $j = 1,\ldots,n$, with the same variables as before and generated by the same distribution *F*. For variables attached to the samples from registers, or when samples have been drawn from censuses, this is practical thing to do without doubling the survey. Replace $X_{jk} (= X_{(R_{jk})k})$ in the original sample with the value $X^{(2)}_{(R_{jk})k}$ having the same rank on the same variable in $\mathbf{X}^{(2)}$. This produces a synthetic dataset $\mathbf{X}^*$ with rows $\boldsymbol{X}^{*T}_j = \boldsymbol{X}^{(2)T}_{(\boldsymbol{R}_j)k} = (X^{(2)}_{(R_{j1})1},\ldots,X^{(2)}_{(R_{jK})K})$. This version will be called *joint* rm (with registers). The distribution of $\boldsymbol{X}^*_j$ will depend on the original $\boldsymbol{X}_j$ only through its rank vector $\boldsymbol{R}_j$.

II. An alternative that is slightly easier to deal with analytically (but not computationally) is to draw one sample for each of the *K* variables in $\mathbf{X}$. Instead of the vectors $\boldsymbol{X}^{(2)}_{j^{(2)}} = (X^{(2)}_{j^{(2)}1},\ldots, X^{(2)}_{j^{(2)}K})^T$, we observe $X^{(2)}_{j^{(2)}_1 1},\ldots,X^{(2)}_{j^{(2)}_K K}$, $j^{(2)}_1,\ldots,j^{(2)}_K = 1,\ldots,n$. This version will be called *independent* rank matching. The synthetic data vectors $\boldsymbol{X}^*_j$ are otherwise formed in the same way as in I. Mixtures of joint and independent rank matching are possible, resampling some variables jointly and others independently.

Generally there is information loss associated with rank matching. The theoretical transition probability $g(\boldsymbol{x}^* \mid \boldsymbol{x})$ associated with the method is not stationary. In the original dataset, the marginal distribution of X_{jk} given the rank vector $\boldsymbol{R}_j$ can depend on components of $\boldsymbol{R}_j$ other than R_{jk}. In other words, for independent rm,

$$\begin{aligned} &F^*_k(x_k \mid \boldsymbol{r}) = F^*_k(x_k \mid r_k) \\ &= F_k(x_k \mid r_k) \neq F_k(x_k \mid \boldsymbol{r}) \end{aligned} \qquad (2.3)$$

For joint rm the first equality will not be exact.

For some multivariate distributions the last inequality in (2.3) may be equality. If that is the case, the transition function is stationary and (in expectation) no information will be lost in the rank matching procedure. Already rank matched data are of this type. Repeated rank matching of an already rank matched sample will (in expectation) not lead to further loss of information.

Lemma 1: As $n \to \infty$, $\boldsymbol{X}_j - \boldsymbol{X}^*_j \to 0$ in probability. Hence, we also have $\boldsymbol{X}^*_j \to \boldsymbol{X}_j$ in law.

Proof of the lemma is skipped. An option to rank matching with register is half-sample rank *swapping* (FCSM 1994, Moore 1996). Split the sample in two random half samples. This can eventually be done separately for different sets of variables in the dataset. For each half sample, replace the values with the values having the same rank on the same variable in the other half sample. Each half sample will then have the properties of a fully rank matched sample with registers, half the size of the original one. The two half samples can then be stacked to form one synthetic dataset $\mathbf{X}^+$. Contrary to rm, rank swapping, preserves the observed marginal distributions of all variables as in the original dataset exactly. This can be desirable in some contexts, but may also make the sample more vulnerable to re-identification. Rank swapping does not preserve the rank structure of the original dataset. For that reason and unlike rank matching, repeated rank swapping is not stationary and leads to further loss of information on the joint distribution. Simulations presented in section 4 indicates that this mixing of the rank structure leads to a somewhat larger loss of information on the structures of the joint distribution than rank matching with registers.

3. A simulation experiment

A user of a confidentiality-protected dataset will be interested how the statistical properties of a dataset have been affected by application of a given confidentiality protection method. If rank matching or rank swapping has protected the dataset, the information loss will mostly be seen as a dilution of the multivariate structures in $\mathbf{X}^*$, such as correlations and regressions, compared to $\mathbf{X}$. This dilution will be studied for rank matching and rank swapping in a small simulation experiment. More extensive simulations will be performed later.

A simulation study was performed to investigate the joint statistical properties of a rank matched and rank swapped dataset with 6 correlated variables

and n =1000 observations $\boldsymbol{X}_1,\ldots\ldots,\boldsymbol{X}_n \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. All continuous variables can be transformed to a normal scale marginally. The effects of rank matching and swapping on the estimated correlations in $\mathbf{X}^*$ and $\mathbf{X}^+$ are shown in table 1.

Var	Variable numbers				
no.	1	2	3	4	5
2	0.170 0.165 0.167	1.000			
3	0.483 0.481 0.477	0.123 0.124 0.121	1.000		
4	0.707 0.704 0.700	0.102 0.095 0.099	0.317 0.317 0.312	1.000	
5	0.893 0.889 0.888	0.137 0.136 0.135	0.441 0.438 0.433	0.644 0.640 0.637	1.000
6	0.982 0.981 0.979	0.171 0.168 0.168	0.469 0.469 0.463	0.694 0.692 0.685	0.875 0.871 0.869

Table 1. Correlations between variables in a simulated dataset. The upper entries are the correlations $\hat{\boldsymbol{P}}$ in the original dataset $\mathbf{X}$. The middle entries are the correlations $\hat{\boldsymbol{P}}^*$ in $\mathbf{X}^*$, and the lower entries show the correlations $\hat{\boldsymbol{P}}^+$ in $\mathbf{X}^+$.

The simulation result in table 1 is a rather typical one. Among 15 estimated pairs of correlation in $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}^*$, using four significant digits, 13 were smaller in $\hat{\boldsymbol{P}}^*$ than in $\hat{\boldsymbol{P}}$. This indicates that some minor dilution of the multivariate structure has taken place. This is not surprising since it is well known (Kruskal 1958) that for multivariate normal variables the rank correlation itself (Spearman's ρ_S) is directly related to the original correlation by

$$\rho \geq \rho_S = (6/\pi)\arcsin(\rho/2)$$
$$\geq (3/\pi)\rho \approx 0.955\rho$$

where minimum is taken in the vicinity of $\rho = 0$.

Next, 13 out of 15 correlations of the rank swapped data in $\boldsymbol{P}^+$ are slightly less than those in $\boldsymbol{P}^*$. This indicates that rank swapping causes a somewhat larger dilution and information loss than rank matching. This was expected and can be attributed to the mixing of the rank-structure taking place. Obviously, the effects of rank matching and rank swapping on the estimates of correlations are insignificant compared to the random variation in these estimates.

Direct estimation of correlations on rank matched and rank swapped data from variables that are highly non-normal can produce adverse results as well as it is adverse to calculate correlations from highly non-normal data at all.

4. 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 also (simple random) samples from the labeled set of units $\mathcal{P}$ to which realizations 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 anonymous dataset $\mathbf{X}$ has information about some of the variable values associated to given labels.

It is clear that an individual having exact information about the value of at least one continuous variable for some unit drawn to the sample will be able to identify that unit in the sample. 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.

The only information on labels associated with the units in $\boldsymbol{s}$ still present in the rank matched dataset $\mathbf{X}^*$ is the rank-structure $\mathbf{R}$. How an intruder can make use of this information to make a disclosure will depend on what kind of extra information on the individuals she has available in her identification file. Two cases will be considered:

a) She has access to one or more variable in $\mathbf{X}$ (but not other variables in the survey) for (at least) one unit and knows that this unit is in the sample. The unit can be considered as a randomly chosen unit from the sample.

b) She has access to the entire population register, but does not know who were drawn to $\boldsymbol{s}$.

Case a will be studied in a simulation experiment presented in section 4.1.

Situation b is an extreme case, but is interesting. This case will be considered in section 4.2 by a small example.

4.1 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 $\boldsymbol{x}_j^*$ in $\mathbf{X}^*$ that minimizes a distance

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

A thorough discussion of the use of discriminant analysis in the context of disclosure control is given in Paaß and Wauschkuhn (1985) and Paas (1988). In order to test the capacity of this decision rule, $\mathbf{W}$ was taken as the inverse of the diagonal of $\hat{\Sigma}^*$ and $\hat{\Sigma}^+$, 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. Table 2 shows that the identifying capacity of combinations of variables increases rapidly with the number of variables available for disclosure for both methods. The number of correct identifications with the same number of variables shows large variations. The simulations indicate, 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. Nevertheless, table 2 gives a rough indication of how the probability of correct disclosure *D*, given that the target is in the sample, $P(D \mid i \in \boldsymbol{s})$, increases with increasing number of variables.

Table 2 may seem discouraging, but for an intruder not knowing that the target person is in the dataset $P(D \mid i \in \boldsymbol{s})$ must be multiplied by $P(i \in \boldsymbol{s})$ which is usually small. $P(D \mid i \notin \boldsymbol{s}) = 0$.

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

Table 2. Minimum and maximum numbers of correct identifications of records in $\mathbf{X}^*$ and $\mathbf{X}^+$ with various numbers of identification variables.

4.2 Case b.

Consider an intruder with access to the population register $\mathbb{X} = (\xi_1, \ldots, \xi_N)^T$ where $\xi_i = (\xi_{i1}, \ldots, \xi_{iK})^T$ are the values of the *K* continuous variables for finite population unit *i*. It is then possible to extract the population rank matrix $\mathbb{R} = (\rho_1, \ldots, \rho_N)^T$ where *N* is the finite population size. Actually, all information about labels contained in $\mathbf{X}^*$ and $\mathbb{X}$ is contained in $\mathbf{R}$ and $\mathbb{R}$. The rank vector ρ_{i_j} corresponding to sample unit *j* will however not be directly observable in the sample. Never the less, the structure of $\mathbb{R}$ determines the probability structure of the sample rank matrix $\mathbf{R}$. There are $(n!)^{K-1}$ possible (unordered) sample rank matrices. They define a partition of the $\binom{N}{n}$ samples in the entire sample space $\mathcal{S}$ into subsets $\mathcal{S}_{\mathbf{R}}$, some of which may be empty by the configuration of $\mathbb{R}$. Let ρ_r be the stochastic variable that maps sample rank *r* to a population rank for a variable. If for an observed matrix $\mathbf{R}$, $\mathcal{S}_{\mathbf{R}}$ is identified, then the probability $P(\rho_r = \rho \mid \mathcal{S}_{\mathbf{R}})$ that a given sample unit *j* corresponds to a given population unit *i* can be calculated exactly. However, it does not seem to be feasible to do this by formula. For large *N* and *n* efficient algorithms will be necessary to identify the partition set $\mathcal{S}_{\mathbf{R}}$ compatible with an observed $\mathbf{R}$.

With only one variable, there is only one possible sample rank matrix and one possible $\mathcal{S}_{\mathbf{R}}$. Then

$$P(\rho_r = \rho) = \binom{\rho - 1}{r - 1}\binom{N - \rho}{n - r}\Big/\binom{N}{n}$$

For two or more variables, rather than trying to develop formulae or asymptotics, I will present a small example trying to illuminate the case. For ease of notation, the ranks of the variable with $k = 1$ will be used as labels, both in the population and in the sample, meaning that $\rho_{i1} = i$ and $r_{j1} = j$.

$\mathbf{R}^T$	$\mathcal{S}_{\mathbf{R}}$	$p(i \mid j) = P(i_j = i \mid \mathbf{R})$
$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 2 & 6 \\ 4 & 5 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 7 \\ 4 & 5 & 6 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 6 \\ 2 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 7 \\ 2 & 3 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(3 \mid 1) = 1/2$ $p(2 \mid 2) = p(4 \mid 2) = 1/2$ $p(6 \mid 3) = p(7 \mid 3) = 1/2$
$\begin{pmatrix} 1 & \mathbf{2} & \mathbf{3} \\ 3 & 1 & 2 \end{pmatrix}$	$\begin{bmatrix} 1 & \mathbf{3} & \mathbf{4} \\ 4 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 2 & \mathbf{3} & \mathbf{4} \\ 5 & 2 & 3 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 1/2$ $p(3 \mid 2) = \mathbf{1}$ $p(4 \mid 3) = \mathbf{1}$
$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$	$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 4 \\ 4 & 5 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 5 \\ 4 & 5 & 1 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 5 \\ 2 & 3 & 1 \end{bmatrix}$	$p(1 \mid 1) = 3/4,\ p(3 \mid 1) = 1/4$ $p(2 \mid 2) = 3/4,\ p(4 \mid 2) = 1/4$ $p(3 \mid 3) = p(4 \mid 3) = 1/4,\ p(5 \mid 3) = 1/2$
$\begin{pmatrix} 1 & \mathbf{2} & \mathbf{3} \\ 1 & 3 & 2 \end{pmatrix}$	$\begin{bmatrix} 1 & \mathbf{6} & \mathbf{7} \\ 4 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 2 & \mathbf{6} & \mathbf{7} \\ 5 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 3 & \mathbf{6} & \mathbf{7} \\ 2 & 7 & 6 \end{bmatrix},$ $\begin{bmatrix} 4 & \mathbf{6} & \mathbf{7} \\ 3 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 5 & \mathbf{6} & \mathbf{7} \\ 1 & 7 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = p(3 \mid 1)$ $= p(4 \mid 1) = p(5 \mid 1) = 1/5$ $p(6 \mid 2) = p(7 \mid 3) = \mathbf{1}$
$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 3 & 6 \\ 4 & 2 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 3 & 7 \\ 4 & 2 & 6 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 6 \\ 4 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 7 \\ 4 & 3 & 6 \end{bmatrix}$ $\begin{bmatrix} 1 & 5 & 6 \\ 4 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 5 & 7 \\ 4 & 1 & 6 \end{bmatrix}, \begin{bmatrix} 2 & 3 & 6 \\ 5 & 2 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 3 & 7 \\ 5 & 2 & 6 \end{bmatrix}$ $\begin{bmatrix} 2 & 4 & 6 \\ 5 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 4 & 7 \\ 5 & 3 & 6 \end{bmatrix}, \begin{bmatrix} 2 & 5 & 6 \\ 5 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 5 & 7 \\ 5 & 1 & 6 \end{bmatrix}$ $\begin{bmatrix} 3 & 5 & 6 \\ 2 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 3 & 5 & 7 \\ 2 & 1 & 6 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 6 \\ 3 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 7 \\ 3 & 1 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 3/8$ $p(3 \mid 1) = p(4 \mid 1) = 1/8$ $p(3 \mid 2) = p(4 \mid 2) = 1/4$ $p(5 \mid 2) = 1/2$ $p(6 \mid 3) = p(7 \mid 3) = 1/2$
$\begin{pmatrix} 1 & 2 & \mathbf{3} \\ 3 & 2 & 1 \end{pmatrix}$	$\begin{bmatrix} 1 & 3 & \mathbf{5} \\ 4 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 4 & \mathbf{5} \\ 4 & 3 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 3 & \mathbf{5} \\ 5 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 4 & \mathbf{5} \\ 5 & 3 & 1 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 1/2$ $p(3 \mid 2) = p(4 \mid 2) = 1/2$ $p(5 \mid 3) = \mathbf{1}$

Table 4 The partitions of $\mathcal{S}$ generated by the sample rank matrices and the induced disclosure probabilities.

***Example*:** Assume that the population rank matrix

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

meaning that $N = 7$ and $K = 2$. Consider samples of size $n = 3$. The sample space consists of 35 such samples. There are 6 possible sample rank matrices $\mathbf{R}$. The 6 sample rank matrices, their associated partition sets and the probabilities $p(i \mid j) = P(i_j = i \mid \mathcal{S}_{\mathbf{R}})$ are given in table 5.2.

Table 4 shows a large variation of the number of samples in each partition. The cases where $p(i \mid j) = 1$ (boldfaced) define identity disclosure with probability one. They make up nine samples in three partition subsets, meaning that before sampling the probability of at least one disclosure is 9/35.

5. Summary and loose ends

This paper is only a brief taste of some ideas that need further research. It indicates that rank matching can be a useful method for confidentiality protection when it can be assumed that the number of accurately measured key-variables available to an intruder is small, even when the intruder knows that the target unit is in the dataset. When the number of key variables is large, rm seems not always to be sufficient alone. A more theoretical development of the properties of rank matching will be given in a later paper.

There are several loose ends that can be topics for future research. Among such I will mention rm with nominal or ordinal discrete variables. This will require an artificial ordering of nominal categories and of the records with like values. The problem is to avoid impossible and very unlikely combinations to occur in a rank matched sample and guarantee that cell frequencies are retained in a rank matched sample, at least in expectation. Another topic is the effect of rm on statistics for sub-domains. Since rm retains sample ranks also within sub-domains I conjecture that this will not pose big problems. A third topic is how rank matching will behave in more complex designs with stratification, two-stage sampling and unequal inclusion probabilities. Here, the formulation of the problem in terms of model and sampling distribution may make the problem easier to handle.

REFERENCES

Duncan, G. T. and Lambert, D. (1986): *Disclosure-Limited Data Dissemination.* J. of the Am. Stat. Assoc. vol 81 no. 393 pp 10-18

_____ (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.

Gouweleeuw, J. M., Kooman, P., Willenborg, L.C.R.J. and de Wolf, P.-P.*: Post Randomisation for Statistical Disclosure Control: Theory and Implementation*. J. of Official Statistics, vol 14 no. 4 pp 463-478

Kruskal, W. H. (1958): *Ordinal Measures of Association.* J. of the Am. Stat. Assoc. vol 53 no. 284 pp 814-861

Moore, R. (1996): *Controlled Data Swapping Techniques for Masking Public Use Microdata Sets*. U.S. Bureau of the Census (unpublished manuscript).

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

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

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

Spruill, N. L. (1982): *Measures of Confidentiality*. in Statistics of Income and Related Administrative Record Research: 1982. Washington DC: US Dept. of Treasury, Internal Revenue Service, Statistics of Income Div. pp 131-136.

Federal Committee on Statistical Methodology (FCSM) (1994): *Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation methodology*.

Sullivan, G. and Fuller, W.A. (1989): *The Use of Measurement Error to avoid Disclosure*. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 802-807.

Willenborg, L. and de Waal, T. (1996): *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics no. 111. Springer Verlag.