STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 41
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics.

CONFIDENTIALITY PROBLEMS RELATED TO SURVEY DATA IN NORWAY AND SOME POSSIBLE APPROACHES

Contributed paper

Submitted by Statistics Norway[1]

Keywords: Income data, disclosure, rank matching, rank swapping.

I. BACKGROUND

1. Contrast to practice in many other statistical offices, Statistics Norway does not release microdata files from surveys and censuses for public use. Such datasets are released to researchers on contract only. Formal identifiers and geographical details have then been removed, but otherwise the datasets are usually released with full detail. This detail includes variables linked to samples from registers by means of our general personal identification number. Among such variables are all the common demographical variables, type and level of education, income based on the assessment and economic transfers. In some surveys the entire or parts of the components of the income tax returns from the assessment register is linked. Among the surveys being linked that way are the now annual Survey of Living Conditions (SLC) and the consumer expenditure survey. This way of supplying information to the survey produces microdata of high detail and quality for the researchers getting access to them, but make the data sets more vulnerable to disclosure in the hands of a disloyal researcher.

2. The assessment in Norway is open. This means that when the assessment is finished, the printed assessment lists with name, address, taxable incomes, net property value of every assessed individual are made public for three weeks in the local assessment offices, and anyone can go and look up what their neighbour earned last year. The lists can freely be copied during a three-week period. In the last few years the lists for the entire country have been printed on a CD. A private company distributes this CD on license to licensed users. Among such users are for instance credit information firms.

[1] Prepared by Johan Heldal, Statistics Norway and University of Oslo. Johan Fosen, Statistics Norway, did the computing with the SLC survey.

3. The openness about the assessment causes income, as presented in the lists, not to be a sensitive variable in itself in Norway. However, the existence of the published lists and of a commercially available CD makes the assessment a powerful key for identification of individual respondents.

4. The concepts of income in the SLC are somewhat different than the concepts used in the assessment. Thus, the two sets of data are not quite compatible. Nevertheless, Statistics Norway wanted to find out if the variables in the two datasets could be manipulated to create parallel variables in the two datasets or in any used for disclosure. The scenario is a researcher having access to SLC and assessment wanting to use the assessment to disclose the identity of as many members of the survey as possible. It turned out that with our strategy, a unique and credible match was obtained for about 36 percent of the adult members of the sample. These were individuals for whom the definitional differences did not apply, mostly people over 60 years old or younger than 25. In addition to income variables, the matching strategy included the variables age (in one year groups) and sex. In order to check out false matches, household composition in the survey was checked against the set of individuals sharing address with the matched individuals on the CD. There is an even greater potential for identity disclosure in the data, and more refined efforts to make such disclosures would most likely lead to a higher proportion of disclosures.

5. Having made the disclosures, the next step was to see what could be done with the survey data to prevent the possibility of making them. We had in mind methods like global recoding (on age), microaggregation (Defays and Anwar (1998) and references therein), perturbation methods (Kim (1986), Paass (1988), Sullivan and Fuller (1989, 1990) and Fuller (1993) and others) and data swapping (Spruill (1983)). We also had in our minds in our minds the question of whether the access to register data could offer us offer us some extra opportunities.

6. In the process of preparing the methods mentioned above, we came up with some new methods which we found exiting and wanted to give priority. These methods, which we have coined *rank matching* and *rank swapping*, will be presented and discussed in section III and IV. During preparation of this paper we became aware that another paper to this conference (paper no. 5) also described rank swapping with the same term and with reference to an unpublished paper by Moore (1996).

7. Experimenting with rank matching and rank swapping on the real survey is an ongoing project that we had hoped would be brought so far that we could publish interesting results in this paper. For practical reasons this has not been possible. However, the basic ideas, some theory and simulation results comparing the two methods are presented. Comparison with other methods will be given in later papers. A description of the real data that entered our disclosure experiment and on which we are still working on is given in section II.

II. DESCRIPTION OF THE DATA

8. We used the 1997 version of SLC, which consists of 3363 primary respondents from a national sample. The primary respondents answered the survey questionnaire. The questionnaire contained both questions affecting only the primary respondents and questions concerning his or her household. In addition to the primary respondents, all other household members were mapped and fully identified in the interview. They will be termed the secondary respondents. Demographic variables, register information on education, the income tax return and several kinds of economic transfers like for instance family allowance, disablement benefits and social security were linked to all secondary as well as primary respondents in the survey. The income tax returns were edited to reflect welfare-oriented income. Not every detail in the original income tax returns was kept in the final survey dataset. It was not, based on the survey-data, possible to recalculate the original assessment data with certainty from the survey. The variables central to matching were *Total income* (TI), *Income after tax* (IAT), *Age* and *Fylke* (County). Total amount of tax paid was be inferred from SLC as TI minus IAT. There are certainly other routes to disclosure using other variables such as for instance education. This paper however, considers only the assessment data.

9. In our experiment we used the variables available on a file equivalent to the above mentioned CD. They were: *Name*, *address*, *municipality of living*, *age* (in one year groups), *sex*, *net income*, *net asset* and *total amount of tax paid*. Municipality implied "Fylke". None of the income variables on the CD were exactly compatible to any variables in SLC, but as demonstrated below, *Total amount of tax paid* (TP) was to be the central matching variable.

10. TP on the CD differed from TP in SLC in two respects: In the SLC variable *Income after tax* (IAT), mandatory job-related pension insurance was subtracted from income and so were child support payments. For individuals having neither of these, IAT is the same in both files. For the majority however, TP in the SLC should be somewhat higher than TP in the CD.

11. There were also some differences between the CD and the file actually used: We did not have access to Name. As a replacement we had sex, which for matching purposes was the primary feature which could be deduced from the name. Address, which was in full text on the real CD, was replaced by a matriculation code. This code identified the house where the household lived, but not necessarily the household. We will not go through the disclosure procedure in detail.

III. THE RANK MATCHING METHOD

12. Consider a sample $\boldsymbol{s}$ of size N drawn from the finite population. This sample gives rise to a dataset $\mathbf{X}$ with records $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK}),\ i = 1, \ldots, N$ that will here be considered as generated by a cumulative superpopulation distribution $F(\boldsymbol{x})$. Assume (for convenience) that the components X_{ik} are all continuous with a joint density $f(\boldsymbol{x})$ and that none of them are functionally related to any other (or at least to no more than one other component). To keep concepts as simple as possible, assume also that the sampling design is simple random so that the $\boldsymbol{X}_i$'s are independent.

13. Next, add noise to the observed $\boldsymbol{X}_i$ vectors, transforming them randomly to vectors $\boldsymbol{Y}_i$ according to a conditional density $g(\boldsymbol{y} \mid \boldsymbol{x})$, producing a new set of observations having density

$$h(\boldsymbol{y}) = \int g(\boldsymbol{y} \mid \boldsymbol{x})\, f(\boldsymbol{x})\, d\boldsymbol{x} \qquad (1.1)$$

If $g(\boldsymbol{y} \mid \boldsymbol{x})$ represents a transition density for a stochastic process with $f(\boldsymbol{x})$ as a stationary distribution, we will have $h(\boldsymbol{y}) = f(\boldsymbol{y})$. In such a case, the noise addition generates a new sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ with the same father-distribution F as the original sample, being statistically equivalent to the original sample. If f is known, a transition density g having this property and other desired properties can be constructed. Noting that f must be an eigenfunction for g corresponding to the eigenvalue 1, such a g can be constructed using orthogonal expansions. 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.

14. But f is not known. One solution could be to use an estimated density $\hat{f}$. The PRAM method amounts to this. However, with a large set of variables involved, the curse of dimensionality quickly makes a satisfactory estimate for f out of reach without a huge amount of data. Even for variables coming from total population registers, the number of variables will soon be too high with an increasing number of variables involved. Therefore, an alternative approach, based on a kind of resampling, will be proposed. This method is, however, not perfect in the sense above.

15. As a first approach, assume that a new sample $\boldsymbol{s}_2$ is drawn according to the same design and with the same sample size as the original sample. $\boldsymbol{s}_2$ gives rise to a new dataset $\mathbf{X}^{(2)}$ with records $\boldsymbol{X}_i^{(2)} = (X_{i1}^{(2)}, \ldots, X_{iK}^{(2)}),\ i = 1, \ldots, N$ with the same variables as before and generated by the same

distribution F. For variables attached to the samples from registers, this is practical thing to do without doubling the survey. From the original dataset, extract the rank matrix $\mathbf{R} = (\boldsymbol{R}_1', \ldots, \boldsymbol{R}_N')'$ whose element in row i and column j, r_{ij}, is the rank of the i-th observation of variable j, x_{ij}. Let $x_{(r)j}$ denote the r-th ordered observation of variable j. Replace the observed values $x_{ij} (= x_{(r_{ij})j})$ with the value $x_{(r_{ij})j}^{(2)}$ having the same rank on the same variable in $\mathbf{X}^{(2)}$. This produces a synthetic dataset $\mathbf{X}^{(2*)}$ with observed elements $x_{ij}^{(2*)} = x_{(r_{ij})j}^{(2)}$. In multivariate terms, this can be expressed as $\boldsymbol{x}_i^{(2*)} = \boldsymbol{x}_{(\boldsymbol{r}_i)}^{(2)} = (x_{1(r_{i1})}^{(2)}, \ldots, x_{K(r_{iK})}^{(2)})$ where $\boldsymbol{r}_i$ is the vector of ranks corresponding to the vector observation $\boldsymbol{x}_i$ in the original dataset. The distribution of $\boldsymbol{x}_i^{(2*)}$ will depend on the original $\boldsymbol{x}_i$ through its rank vector $\boldsymbol{r}_i$. Let H be the joint cumulative distribution of the rank vectors $\boldsymbol{R}_i$. The density of a row vector $\boldsymbol{X}_i^{(2*)}$ in this dataset is

$$f^*(\boldsymbol{x}_i^{(2*)}) = \int f^*(\boldsymbol{x}_{(\boldsymbol{r}_i)}^{(2*)} \mid \boldsymbol{r}_i) dH(\boldsymbol{r}_i) . \qquad (1.2)$$

Let $f(\square \mid \boldsymbol{r}_i)$ be the common conditional density of $\boldsymbol{x}_i^{(2)}$ and $\boldsymbol{x}_i$ given $\boldsymbol{r}_i$. It could be tempting to believe that $f^*(\square \mid \boldsymbol{r}) = f(\square \mid \boldsymbol{r})$, in which case the unconditional densities $f^*(\square)$ and $f(\square)$ for $\boldsymbol{x}_i^{(2*)}$ and $\boldsymbol{x}_i$ would be the same. If so were the case, $\mathbf{X}^{(2*)}$ and $\mathbf{X}$ would be statistically equivalent samples from F and no loss of information would have taken place in the replacement process. Unfortunately, it is not quite so. But $f(x_{ij}^{(2*)} \mid r_{ij}) = f(x_{(r_{ij})j}^{(2)} \mid r_{ij})$ is not the same as the distribution $f(x_{(r_{ij})j}^{(2)} \mid \boldsymbol{r}_i)$. The conditional distributions of $x_{ij}^{(2)}$ and x_{ij} given a rank vector $\boldsymbol{r}_i$, can depend also on other than components of $\boldsymbol{r}_i$ than r_{ij}. The information in this dependence is lost. For each marginal component, the distribution of $x_{ij}^{(2*)}$ is the same as for the original x_{ij}, but when it comes to the joint behaviour only the information contained in the ranks is preserved.

16. This resampling-replacement procedure is what in this paper is termed *rank matching*. The procedure preserves the rank structure of the original data matrix $\mathbf{X}$. The best way to measure the loss of information incurred by the method is by using the Kullback-Leibler information distance

$$I(f, f^{(2*)}) = -\int \log(f^{(2*)} / f) dF .$$

17. In most cases, the variables available from registers only make up some of the variables in a survey dataset. Denote the observation vector of non-register variables by $\boldsymbol{Z}_i$. These variables are not affected by the resampling and replacement procedure. An observation vector $\boldsymbol{x}_i^{(2*)}$ will consist of components taken from many different units in $\boldsymbol{s}_2$ and does not represent any physically existing sampling unit. This is what makes it difficult to reveal the original identity i of a rank-swapped record in dataset and thereby the values of $\boldsymbol{Z}_i$. This will be illuminated below with a simulation experiment. Unfortunately, rank matching is not applicable to census type data making up the entire finite population.

18. A simulation study was performed to investigate the joint statistical properties of a rank swapped dataset with 6 correlated variables following a jointly multinormal distribution and 1000 observations. All continuous variables can be transformed to a normal scale marginally. The effect of rank matching on the estimated correlations in $\mathbf{X}^{(2*)}$ was considered.

Variable numbers	Variable numbers 1	2	3	4	5	6
1	1.0000	0.1961 *0.1698* 0.2476	0.4472 *0.4826* 0.4715	0.7071 *0.7072* 0.7207	0.8944 *0.8926* 0.8960	0.9806 *0.9821* 0.9800
2	*0.1698* 0.1653 0.1667	1.0000	0.0877 *0.1229* 0.1568	0.1387 *0.1021* 0.1908	0.1754 *0.1370* 0.2145	0.1923 *0.1710* 0.2461
3	*0.4826* 0.4811 0.4767	*0.1229* 0.1241 0.1211	1.0000	0.3162 *0.3169* 0.3705	0.4000 *0.4407* 0.4153	0.4385 *0.4694* 0.4554
4	*0.7072* 0.7036 0.7002	*0.1021* 0.0951 0.0986	*0.3169* 0.3174 0.3116	1.0000	0.6325 *0.6439* 0.6376	0.6934 *0.6942* 0.7029
5	*0.8926* 0.8894 0.8876	*0.1370* 0.1364 0.1348	*0.4407* 0.4377 0.4333	*0.6439* 0.6399 0.6371	1.0000	0.8771 *0.8749* 0.8825
6	*0.9821* 0.9812 0.9788	*0.1710* 0.1683 0.1678	*0.4694* 0.4687 0.4634	*0.6942* 0.6921 0.6847	*0.8749* 0.8710 0.8689	1.0000

Table 1. Correlations between variables in a simulated dataset. In the upper triangle the upper entries show the true correlations, the middle entries *(italic)* show the empirical correlations in $\mathbf{X}$ and the lower entries show the empirical correlations in $\mathbf{X}^{(2)}$. In the lower triangle the upper entries *(italic)* are the same as the entries in the upper triangle the middle entries show the correlations in the rank matched dataset $\mathbf{X}^{(2*)}$ and the lower entries show the correlations in the rank-swapped dataset $\mathbf{X}^{*}$.

19. Table 1 shows that the correlations in the synthetic dataset $\mathbf{X}^{(2*)}$ follow the correlations in the original dataset $\mathbf{X}$ much closer than the correlations in the dataset $\mathbf{X}^{(2)}$ having given values to $\mathbf{X}^{(2*)}$. This illustrates clearly that the rank structure carries the essential information about the joint distribution of the variable and that very little information is carried in the conditional distribution given the ranks. The amount of joint information lost from $\mathbf{X}$ in the rank matching procedure is insignificant compared to the random variation between the independent datasets $\mathbf{X}$ and $\mathbf{X}^{(2)}$. However, out of 15 pairs of correlation in $\mathbf{X}$ and $\mathbf{X}^{(2*)}$, 14 of the estimated correlations in $\mathbf{X}^{(2*)}$ is smaller those of $\mathbf{X}$. This indicates that the two datasets are not completely equivalent.

20. An important question is with what confidence an intruder can identify the original record number associated with the synthetic record $\boldsymbol{x}^{(2*)}$? 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 $\mathbf{X}^{(2*)}$, she uses discriminant analysis and decides for the following decision rule: Choose the record $\boldsymbol{x}_i^{(2*)}$ in $\mathbf{X}^{(2*)}$ that minimises a distance

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

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 $\hat{\mathbf{S}}^{(2*)}$. $\hat{\mathbf{S}}^{(2*)}$ is the obvious estimate of the covariance matrix based on $\mathbf{X}^{(2*)}$. All 63 possible combinations of one to six variables from $\mathbf{X}$ and $\mathbf{X}^{(2*)}$ were tested and the number of correct hits recorded. The results are summarised in table 2.

The number of variables used	Number of correct hits	The number of variables used	The number of variables used
One (of 6 variables)	6-41	Four (15 combs.)	845-989
Two (of 15 pairs)	137-545	Five (6 combs.)	983-996
Three (20 triples)	472-933	Six (1 comb.)	996

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

21. Table 2 shows a that the identifying capacity of the combinations of variables increases rapidly with the number of variables available for disclosure. 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 while those showing lowest correlations produce the highest number of correct hits.

22. The results in table 2 may seem discouraging. With precisely measured continuous variables all values in the sample dataset $\mathbf{X}$ as well as in the population will be unique. If this is also the case with the intruder's record $\boldsymbol{x}$, identification in $\mathbf{X}$ is an immediate disclosure. But for an intruder seeking in $\mathbf{X}^{(2*)}$, an exact match will not take place. If the intruder does not know that the owner of $\boldsymbol{x}$ is in the dataset, she cannot know with the confidence indicated by table 2 if the best match in $\mathbf{X}^{(2*)}$, as measured by the given distance, gives a correct disclosure. However, if $\boldsymbol{x}$ is in $\mathbf{X}^{(2)}$, she can reveal it by obtaining an exact match on at least one variable in one record of $\mathbf{X}^{(2*)}$, but this will not disclose anything about the records in $\mathbf{X}$.

In the problem that initiated this research, the intruder's problem was not to identify a given record in a given dataset, but rather to identify as many individuals from the sample as possible in a population file. The intruder will be perfectly able to identify the sample $\mathbf{s}_2$, but will not be able to disclose the values of variables not in her identfication file. She will not be able to say much about identities in $\mathbf{X}$.

23. Crucial in assuming that the X_{ij} s were all continuous, was to ensure that all the ranks were unique and that values in $\mathbf{X}$ were replaced with similar values so that the statistical properties an the data integrity were disturbed as little as possible. If a similar approach is to be attempted on discrete variables, an artificial ordering will have to be introduced between units having the same value on the discrete variable. Such orderings can be introduced using other variables in $\mathbf{X}$ and $\mathbf{X}^{(2)}$ or randomly. Care must be taken to avoid illegal combinations in $\mathbf{X}^{(2*)}$. Unlikely combinations should not be more likely as a result of rank matching. For a variable like age, which is usually given in one-year groups,

random ordering within each group may be acceptable. Random ordering can also be used to resolve ties in continuous measurements due to insufficient measurement accuracy.

24. 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. If the number of positive values in $\mathbf{X}^{(2)}$ exceeds the number in $\mathbf{X}$, some zeroes in $\mathbf{X}$ will be replaced by the smallest positive values in $\mathbf{X}^{(2)}$. Improper ordering of the original zeroes in $\mathbf{X}$ can easily introduce positive values on units that according to the values of other variables cannot be positive. Detailed discussions about how to handle discrete values will not be given here.

25. A possible variation of the rank-matching method is to draw one resample for each variable so that different sets of variables are replaced by values from different samples. Equation (1.2) will still hold. Since the synthetic dataset will take values from a larger set of physical population units, the confidential safety will increase.

IV. RANK-SWAPPING

26 Most survey variables cannot be found in registers. Some of them can be potential disclosure risks alone or in combination with other variables and some may be sensitive. Assume as a start that the sample size N is an even number. Then, in lack of registers, a possible approach is to split the original sample $\boldsymbol{s}$ in two random half-samples of equal size $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ so that the two parts are equivalent samples from F. From each of the two equivalent halves, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, of the original dataset $\mathbf{X}$, their (local) rank matrices $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ can be extracted. The rank-matching method can be applied mutually to swap the values in the two halves, generating datasets $\mathbf{X}^{(1*)}$ and $\mathbf{X}^{(2*)}$ where $\mathbf{X}^{(2*)}$ has the data from $\mathbf{X}^{(2)}$ but the rank structure $\mathbf{R}^{(1)}$ and vice versa. Then $\mathbf{X}^{(1*)}$ and $\mathbf{X}^{(2*)}$ are stacked into one combined synthetic dataset $\mathbf{X}^{*}$. We term this method *rank swapping* to distinguish it from rank matching where no half-samples are used.

27. During the swapping process, a mixing of the rank structure carried over from $\mathbf{X}$ to $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ will take place, leading to further loss of information. The extent of this mixing and its effect on some correlation structures will be studied below. Our results indicate that the effect will be negligible in large samples. On the other hand, and also unlike rank matching, all one-dimensional marginals in the original sample will be preserved exactly in data. This property is not very important for a survey, but can be highly desirable if such a method is to be applied to census data. If desired, two and higher dimensional structures of the original data can be also preserved by swapping some variables en bloc.

28. The influence of rank swapping on correlations was simulated on multinormal data. To facilitate comparisons, the same dataset was used as for rank matching. The results are shown in the lowest entries in the lower triangle in table 1. The rank swapped correlations are uniformly lower than the rank matched ones, clearly indicating that rank swapping inflicts a slightly higher loss of information than rank matching.

29. Simulations with larger sample sizes indicate that the relative loss of information, both with rank matching and rank swapping decreases as the sample size increases. The probability law of the asymptotics related to the sample size is however not clear.

The number of variables used	Number of correct hits	The number of variables used	The number of variables used
One (of 6 variables)	3-43	Four (15 combs.)	697-953
Two (of 15 pairs)	93-342	Five (6 combs.)	931-982
Three (20 triples)	253-775	Six (1 comb.)	988

Table 3. Minimum and maximum numbers of correct identifications of records in $\mathbf{X}^*$ with various numbers of identification variables.

29. The identification exercise performed with rank matching was repeated with rank swapping. Attention was paid to the fact that no swapped record could carry any of its original values. The results are shown in table 3. The number of correct identifications is somewhat lower than with rank matching. Contrary to rank matching, there was no clear tendency in the data that the combinations with high correlations produce a lower number of correct hits.

30. Even though the number of correct hits is lower with rank swapping than with rank matching, identification in a rank swapped sample will be easier than in a rank matched sample. The reason for this is that rank matching "swaps" in data from an external sample, while the swapped data in the rank swapped sample comes from the sample itself. An intruder having exact measurements of at least one of the continuous variables will be able to reveal the presence of his friends in the sample almost certainly. This is far more difficult with rank matching where the values of the disclosure variables have been replaced by values from an external sample. In a context with a census or where the intruder has access to a complete population file, four to five variables will be sufficient to disclose the identity of a unit. The exact number of variables needed for an intruder to act upon a match also depends on the loss felt from doing a wrong disclosure compared to the gain felt from doing a correct one. The result indicates that with four or more continuous variables potentially available from disclosure, and a sample size of 1000, rank swapping alone will be insufficient for disclosure protection. Some rounding or perturbation upon the rank swapped data may be needed.

31. If the original sample size is not an even number, select one observation vector at random to be kept out of the swapping process. If this is considered undesirable, the problem can be alleviated. A variant of the method easing the problem is to draw one half-sample for each variable, eventually keeping different odd-records outside the swapping process for each variable. This will increase the number of potential synthetic samples given the original one and also to some degree the confidentiality of the sample. The one variable at the time version also provides opportunities to exercise more control over the swapping process. For variables that contain discrete as well as continuous values, it will sometimes be desirable to swap only among the continuous part of the distribution. For income variables, often containing many zeroes and some positive (or negative) values, it will be desirable not to swap zeroes with positive values to maintain the data integrity. I other cases it may be desirable to swap some variables concomitantly for the same reason.

32. Like for rank matching, rank swapping of discrete values will require an artificial ordering of like values. The same kinds of care should be taken to safeguard the data integrity. For variables that take both discrete and continuous values, it is sometimes desirable to swap only in the continuous part of the distribution. In our experiments with SLC, this was necessary with the income variables containing many zeroes. In order to manage this in relation to other variables, independent half-samples were taken among the positive values of each such variable.

V CONCLUSION

33. The simulations shown indicated that for datasets representing sample surveys, rank matching may be superior to rank swapping, both with respect to loss of information and data protecting capability whenever both methods are options. Comparisons should however also be done with microaggregation and other methods. The simulations are done with simple random sample design. In SLC and many other samples in Norway, both individuals and households are units and the samples are to some extent drawn with unequal probability. It is a topic for further research how to adapt the methods to this reality. The asymptotic properties of the methods as the sample size increases is another. To which extent will the confidentiality protection they offer and the relative information loss increase or decrease? These are some of the questions that we hope to answer in future research.

REFERENCES

Chen, G. and Keller McNulty, S. (1998). *Estimation of Identification Disclosure Risk in Microdata.* Journal of Official Statistics, 14, pp 79-95.

Defays, D. and Anwar, M.N. (1998): *Masking Microdata using Micro-Aggregation*. Journal of Official Statistics,vol 14 no. 4 pp 449-461

Duncan, G. T. and Lambert, D. (1986): *Disclosure-Limited Data Dissemination.* J. of the Am. Stat. Assoc. vol 81 no. 393 s 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*. Journal of Official Statistics, vol 14 no. 4 pp 463-478

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.

Kim, J. (1986): *A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation*. Proceedings of the Section on Survey Research Methods, American Statistical Assoc. pp 370-374.

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

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

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

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.

_____(1983): *The Confidentiality and Analytic Usefulness of Masked Business Microdata.* Proceedings of the Section on Survey Research Methods, American Statistical Assoc. pp 602-607

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

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.

de Waal, T. and Willenborg, L. (1998): *Optimal Local Suppression of Microdata*. Journal of Official Statistics, 14, pp 421-435

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

Working paper no. 5: *An Empirical Comparison of SDC Methods for Contiuous Microdata in Terms of Information Loss and Disclosure Risk.* Submitted by Universitat Rovira i Virgili, Catalonia, Spain.