

**UNITED NATIONS STATISTICAL COMMISSION and EUROPEAN COMMISSION
ECONOMIC COMMISSION FOR EUROPE STATISTICAL OFFICE OF THE
CONFERENCE OF EUROPEAN STATISTICIANS EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UN/ECE and Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (ii): New data release techniques

ACCESSING MICRODATA VIA THE INTERNET

Invited Paper

Submitted by the Central Bureau of Statistics, Israel¹

I. Introduction

Microdata files released at the Israel Central Bureau of Statistics currently undergo a series of ad-hoc decisions with regard to protecting the confidentiality of statistical entities. In general, there are three levels of microdata files: PUF which is released upon request to legitimate researchers after they outline the details of their research, MUC which is available to researchers in universities and approved research institutions that have a special contract with the CBS, and files available in an on-site research facility at the CBS where researchers can work and access original files (without direct identifiers) for data analysis. Because of the current lack of a systematic way of evaluating the risks of the files, mistakes are common. One example is the release of multiple PUF files for the same survey or census, each having detailed coding for different sets of variables. These files can easily be linked using common variables and the entire original file can be recovered. The fact that up till now PUF files were generally not given to the wide public, but were granted to researchers with specific requests, gave a false sense of security with regard to the level of protection needed on the PUF file.

In recent years, the Israel CBS has faced increasing demands for microdata files, especially with regard to providing remote access to users over the internet. Some of the microdata files of the CBS are in the process of being placed on the internet for this purpose. This is being implemented in two parallel systems:

1. A multi-facet system for internet access is being developed at the CBS that allows users to build custom-made tables and/or download files, depending on the level of security of the file. With this new application, a decision has to be made as to what type of file should be in the background in order to produce the tables and whether downloading all or parts of the file will be allowed. A file defined as PUF from which tables can be generated will a priori produce tables that are safe and there is no need for any further restriction or licensing. However, the data generated under such restrictions might have insufficient detail to have any real value to researchers. Another possibility is to have the original file in the background of the system from which tables can be generated, but this means placing strict restrictions on the tables themselves, such as a minimum number of cases to a cell, a maximum number of dimensions to the table, disclosure control software to evaluate and detect unsafe combinations, and prior registration of the users.

¹ Prepared by Natalie Shlomo (natalies@cbs.gov.il).

2. The Israel CBS has an agreement to place microdata files in a depository managed by the Hebrew University of Jerusalem. Some of the current PUF microdata files have been put on an internet website of the depository to provide remote access to registered researchers at universities and research institutions of Israel for producing custom-made tables and carrying out data analysis. There are currently no restrictions on the tables and files can be downloaded as well.

With the implementation of the new internet systems, the CBS recognized the need to develop methodologies to reassess the level of security of the microdata files that it releases. A working group was set up at the CBS to address these issues under the leadership of Prof. Yosi Rinott of the Hebrew University of Jerusalem. The purpose of the research is to provide better tools for defining the levels of security of microdata according to its intended usage. The ultimate goals will be to develop methodologies for evaluating the risks and categorizing the files according to risk measures; to determine thresholds and accepted levels of risk; and to develop methods of disclosure control that best suit the growing needs and policies of the CBS. In addition, there is an immediate need to take disclosure control action during the interim period, in particular for the files that will be placed on the internet.

In section II of the paper we compare two methods of disclosure control for a simple theoretical example using the R-U confidentiality map developed by Duncan, et.al. (2001). To build an empirical R-U confidentiality map for a real data set, we examine in section III different methods for measuring the disclosure risk based on a realistic attack scenerio by linking the data set to the National Population Register (NPR) and also calculate estimated global risk measures based on the sampling design and the individual risk methodology (Benedetti, et.al. (2003), Rinott (2003)). In section IV information loss measures for the different levels of disclosure risk will be calculated and in section V a discussion for determining safe microdata files, especially for use with remote access via the internet.

II. R-U Confidentiality Map

Duncan, et.al. (2001) develop the means for examining the balance between information loss and disclosure risk through the use of an R-U confidentiality map. R is a risk measure for the data file and U is the utility of the data file. The map is a function of the parameters used in the disclosure control method. For example, adding random noise to a variable affects its variance. As the variance of the random noise increases, the risk of disclosing statistical entities decreases but the analytic properties of the variable can be seriously compromised. With the use of the R-U confidentiality map, a decision theory can be developed based on a given risk threshold, and optimal parameters of the disclosure control methods can be determined that maximize the data utility and minimize the disclosure risk.

Duncan, et. al. (2001) presented a simple theoretical example based on the perturbative method of additive random noise to normally distributed variables to show the use of the R-U confidentiality map. We will elaborate and compare this method to global recoding of the variable which is a non-perturbative method of disclosure control.

Let $X_1, X_2, \dots, X_n \sim N(\mathbf{q}, 1)$ and let Y_1, Y_2, \dots, Y_n be the masked data after applying a disclosure control method. In general, a user will estimate a parameter \mathbf{q} of the distribution based on the released data, and will most likely estimate the parameter while ignoring the disclosure control method. A more sophisticated user that knows that masking of the data has taken place will want to estimate the true parameter \mathbf{q} of the distribution. This can be carried out by moment or maximum likelihood estimates that take into account that the released data has been coarsened or perturbed. The use of the EM algorithm, for example, is a general method for finding the maximum likelihood estimate of the parameter of the underlying distribution from the given data which has been altered.

The two disclosure control methods analyzed:

1. Additive random noise to the variable, $\hat{X}_i = X_i + \mathbf{e}_i$, $\mathbf{e}_i \sim iid(0, \mathbf{I}^2), i = 1..n$.
2. Global recoding or coarsening the variable by publishing categorized groupings designated by cut off values of the distribution, (a_i, a_{i+1}) $i = 1..g$ where $a_0 = -\infty$ and $a_{g+1} = \infty$

Assuming that the data user is interested in estimating the population mean \mathbf{q} , the MLE of the parameter for additive random noise is the sample mean since no bias was introduced. The MLE of the parameter for global recoding can be estimated by the E-M algorithm though this will not be shown here.

For the two disclosure control methods, we can apply the R-U confidentiality map. The utility of the data will be measured as the reciprocal of the variance of the sample mean after the application of the disclosure control method.

For the first method of additive random noise, the variance of $\hat{\mathbf{q}}$ will be equal to:

$$var(\hat{\mathbf{q}}) = var(\bar{Y}) = \frac{\mathbf{s}_y^2}{n} = \frac{1 + \mathbf{I}^2}{n}. \text{ For the second method of global recoding, the utility of the MLE}$$

of $\hat{\mathbf{q}}$ can be calculated by the Fisher Information Matrix $I_n(\mathbf{q})$:

Defining, $\mathbf{j}(X) = \frac{1}{\sqrt{2\mathbf{p}}} e^{-\frac{X^2}{2}}$ we obtain:

$$E_{\mathbf{q}} \left(\frac{d}{d\mathbf{q}} \log p_{\mathbf{q}}(Y = y_i) \right)^2 = \sum_{i=1}^g \left(\frac{\frac{d}{d\mathbf{q}} p_{\mathbf{q}}(Y = y_i)}{p_{\mathbf{q}}(Y = y_i)} \right)^2 p_{\mathbf{q}}(Y = y_i) = \sum_{i=1}^g \frac{\left(\frac{d}{d\mathbf{q}} p_{\mathbf{q}}(Y = y_i) \right)^2}{p_{\mathbf{q}}(Y = y_i)}$$

For element i :

$$\frac{d}{d\mathbf{q}} p_{\mathbf{q}}(Y = y_i) = \frac{d}{d\mathbf{q}} \int_{a_i}^{a_{i+1}} \mathbf{j}(x - \mathbf{q}) dx = \int_{a_i}^{a_{i+1}} \frac{d}{d\mathbf{q}} \mathbf{j}(x - \mathbf{q}) dx = \int_{a_i}^{a_{i+1}} (x - \mathbf{q}) \mathbf{j}(x - \mathbf{q}) dx = \mathbf{j}(a_i - \mathbf{q}) - \mathbf{j}(a_{i+1} - \mathbf{q})$$

And the final result is $I_1(\mathbf{q}) = \sum_{i=1}^g \frac{[\mathbf{j}(a_i - \mathbf{q}) - \mathbf{j}(a_{i+1} - \mathbf{q})]^2}{\mathbf{F}(a_{i+1} - \mathbf{q}) - \mathbf{F}(a_i - \mathbf{q})}$. For n variables, $I_n(\mathbf{q}) = nI_1(\mathbf{q})$.

The disclosure risk will be measured as the reciprocal of the MSE between a specific target variable X and the released variable obtained from the masked sample \hat{X} . A small MSE between the target variable and the released variable will increase the likelihood of reidentification.

For the first method of additive random noise, the MSE is: $E(\hat{X} - X)^2 = E(\mathbf{e}^2) = \mathbf{I}^2$ and the disclosure risk function is: $R^a = \frac{1}{\mathbf{I}^2}$.

For the second method of global recoding, we can assume that the sophisticated attacker will try and identify the target variable by:

$$\hat{X}_i = E(X_i / X_i \in (a_i, a_{i+1})) = \frac{\int_{a_i}^{a_{i+1}} x \mathbf{j}(x - \mathbf{q}) dx}{p(X_i \in (a_i, a_{i+1}))} = \frac{\int_{a_i}^{a_{i+1}} (x - \mathbf{q}) \mathbf{j}(x - \mathbf{q}) dx}{p(X_i \in (a_i, a_{i+1}))} + \mathbf{q}$$

From here, we calculate the mean square error:

$$E(\hat{X} - X)^2 = \sum_{i=1}^g E((\hat{X}_i - X_i)^2 / X_i \in (a_i, a_{i+1})) p(X_i \in (a_i, a_{i+1})) = \sum_{i=1}^g \int_{a_i}^{a_{i+1}} (\hat{x} - x)^2 \mathbf{j}(x - \mathbf{q}) dx$$

By replacing \hat{X} with the above formula, and by noticing that $\sum_{i=1}^g a_i \mathbf{j}(a_i - \mathbf{q}) - a_{i+1} \mathbf{j}(a_{i+1} - \mathbf{q}) = 0$

and $\sum_{i=1}^g \int_{a_i}^{a_{i+1}} \mathbf{j}(x - \mathbf{q}) dx = 1$ we obtain the MSE :

$$E(\hat{X} - X)^2 = 1 - \sum_{i=1}^g \frac{[\mathbf{j}(a_i - \mathbf{q}) - \mathbf{j}(a_{i+1} - \mathbf{q})]^2}{\Phi(a_{i+1} - \mathbf{q}) - \Phi(a_i - \mathbf{q})} = 1 - I_1(\mathbf{q}),$$

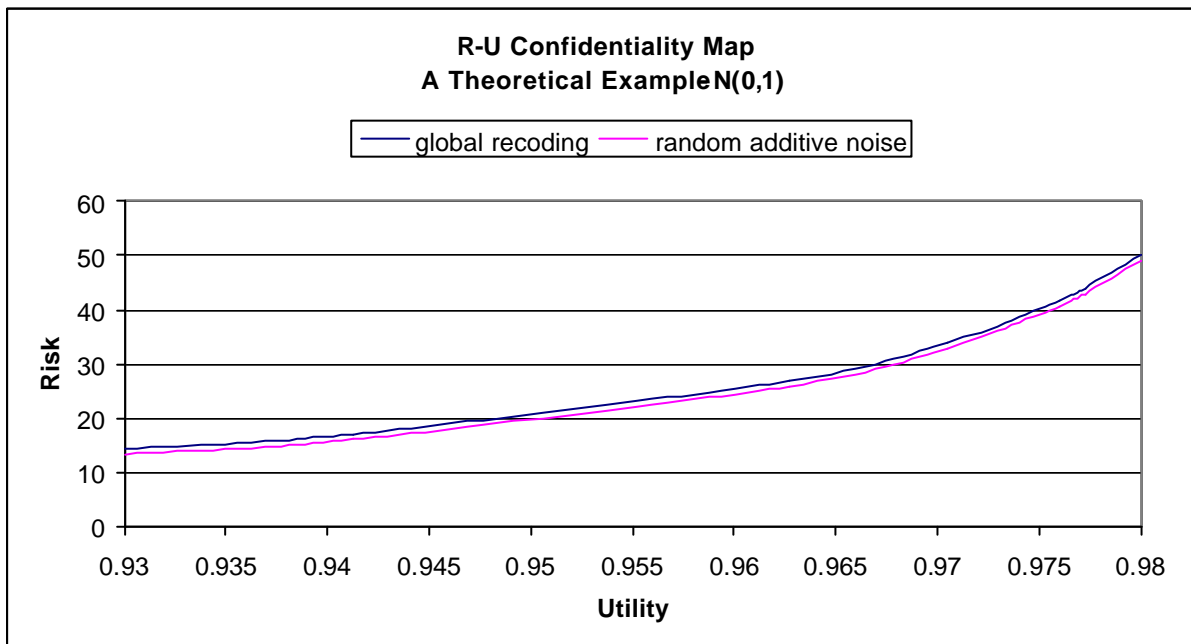
and the disclosure risk function,

$$R^c = \frac{1}{1 - I_1(\mathbf{q})}.$$

For any parameter \mathbf{q} , we compare the risk measures of the two methods by setting the utilities to be equal, ie. $I_1(\mathbf{q}) = \frac{1}{1 + I^2}$, and obtain: $\frac{1}{R^c} = 1 - I_1(\mathbf{q}) = 1 - \frac{1}{1 + I^2} = 1 - \frac{1}{1 + \frac{1}{R^a}} \Rightarrow R^c = R^a + 1.$

Thus for all parameters of the disclosure control method, and for the same utility, $R^c > R^a$ and the risk is always greater for global recoding than for additive noise. This is an interesting result on the basis of this simple model and shows that further theory is called for.

Figure 1: R-U confidentiality map for variables distributed $N(0,1)$



Israel CBS has maintained a policy of releasing unperturbed data and has never considered random noise or other perturbative methods (Kamen (2001)). The sole method for providing disclosure control at the moment is by global recoding of the categorical demographic and geographic identifying variables, ie. year of birth, country of birth, locality code, etc. or through eliminating identifying variables altogether from the file. We will continue with more experiments and theory on random noise and other perturbative methods to see if these are preferable and whether the policies at the CBS should be changed.

The following sections of the paper describe disclosure risk and information loss measures for a real data set using the non-perturbative method of global recoding in practice at the CBS. An empirical R-U confidentiality map can be built to determine optimal parameters for disclosure control depending on the different levels of security needed for the file.

III. Measuring Disclosure Risk

In determining the risks of the microdata file we will use two methods, both based on the probability of obtaining a correct match in the population for a given sample unit. The underlying and realistic disclosure scenerio is that a potential data snooper has access to the NPR which includes demographic and geographic information for all citizens in the State of Israel. We assume also that the data snooper is interested in the uniques of a sample defined under a key made up of common variables to both the NPR and the sample. By focusing on the sample uniques, the data snooper increases his chances of a successful link.

The following evaluations are performed on a real data set, the Israel Income Survey 2000 (IS). The sample file contains 32,869 records that were sampled at about 1:126.

The disclosure risks of the data set were evaluated in two ways:

1. Estimation of a global risk measure defined as the expected number of correct matches to the population for the sample uniques, using the sampling design and the individual risk measure methodology (Benedetti, et.al.(2003), Seri, et.al. (2003) and Rinott (2003)). The individual risk measure methodology assumes that population frequencies F_k for a key k , given the sample frequencies f_k , follows a negative binomial distribution with success probability p_k and the number of successes f_k ie., $F_k/f_k \sim NB(f_k, p_k)$. The individual risk measure r_k is calculated for every cell k in the key and represents the probability that any sample unit in cell k can be correctly matched to the total population, ie. $r_k = \frac{1}{F_k}$. The estimate of the risk measure is:

$\hat{r}_k = E_{p_k} \left(\frac{1}{F_k} / f_k \right)$ under the negative binomial distribution. The estimate depends on p_k , which

is estimated by: $\hat{p}_k = \frac{f_k}{\hat{F}_k} = \frac{f_k}{\sum_{i:i \in k} w_i}$. Note that \hat{p}_k is based on the weights that were assigned to

each unit in the sample at the estimation stage of the survey processing. Weights are typically calculated using a calibration method by benchmarking the inflated sample to known population totals, such as geographical areas, age and sex distributions. Thus by utilizing the weights of the survey, the population frequencies, F_k can be estimated. Since r_k is the probability of a correct match to the population for any unit in the cell k , we can derive a global measure which is the expected number of correct matches for the entire file, $t = \sum_{k=1}^K f_k r_k$ and is estimated by:

$\hat{t} = \sum_{k=1}^K f_k \hat{r}_k$. In our scenerio, we are interested in the sample uniques so the global measure is:

$t = \sum_{k=1}^K I(f_k = 1) r_k$ and its estimate: $\hat{t} = \sum_{k=1}^K I(f_k = 1) \hat{r}_k$.

2. Linking the sample uniques in the dataset to the NPR using keys with common variables and calculating the average number of correct matches for the sample uniques under the above scenerio. This method is evaluated by comparing names for those with a one-to-one matching status.

The original variables in the key common to both the NPR and the IS were: district (24 categories), type of locality, ie., urban according to population sizes and rural according to the type (16 categories), locality code (215 categories), religion (2 categories), gender (2 categories), year of birth (85 categories), marital status (5 categories), country of birth including country of birth of father for those born in Israel (130 categories), ethnic group (2 categories) and year of immigration (85 categories).

Six other keys were developed for the evaluation, each key being more coarse than the previous one. The variables that were recorded in the building of the keys were: type of locality bottom coded for up to 50,000 persons in the locality, districts collapsed to regions, elimination of locality code, groupings for year of immigration and year of birth, country of birth collapsed to continent of birth. The keys are defined as Key1 to Key6, and it should be noted that the current definition of the PUF file for the IS is Key1.

III.A. Estimating the Global Risk Measure

The following are the results obtained for the estimated global risk measure, ie. the expected number of correct matches of the sample uniques to the population, and its percentage out of the total sample size :

Table 1: Results of the estimated expected number of correct matches for the IS

	Full Key	Key1	Key2	Key3	Key4	Key5	Key6
Number in Key	28,658	23,264	17,469	16,694	12,627	6,669	4,726
Number of Uniques in Sample	26,121	19,049	12,523	11,07	8,128	3,319	2,083
Estimated Expected Number of Correct Matches \hat{t}	1,219.3	884.4	590.2	526.6	388.5	158.0	98.2
Max r_k	22.3%	22.3%	22.3%	22.3%	22.3%	15.5%	22.3%
Percentage out of the total sample size	3.7%	2.7%	1.8%	1.6%	1.2%	0.5%	0.3%

Note that some local suppressions have to be undertaken to reduce the high individual risk measures, r_k for cell k of the key.

The following bar charts note the differences in the distributions of the individual risk measures according to the full key and the most collapsed Key6. Groups of individual risk measures were defined where the first group has the smallest risk (up to 0.03%) and the last group has the most risk (over 13.5%):

Figure 2: Number of records in groups according to individual risk measures for original key - IS

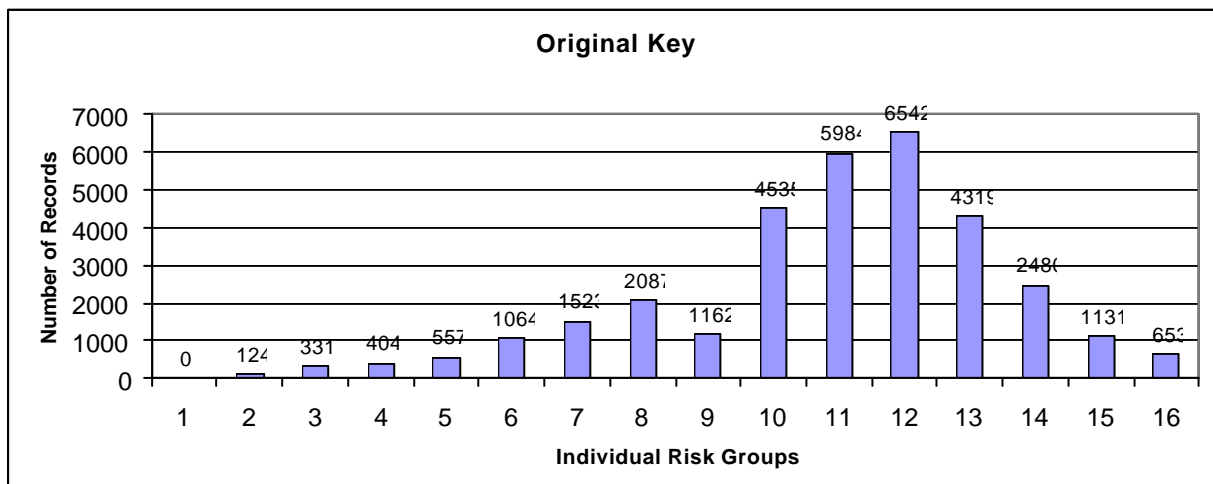
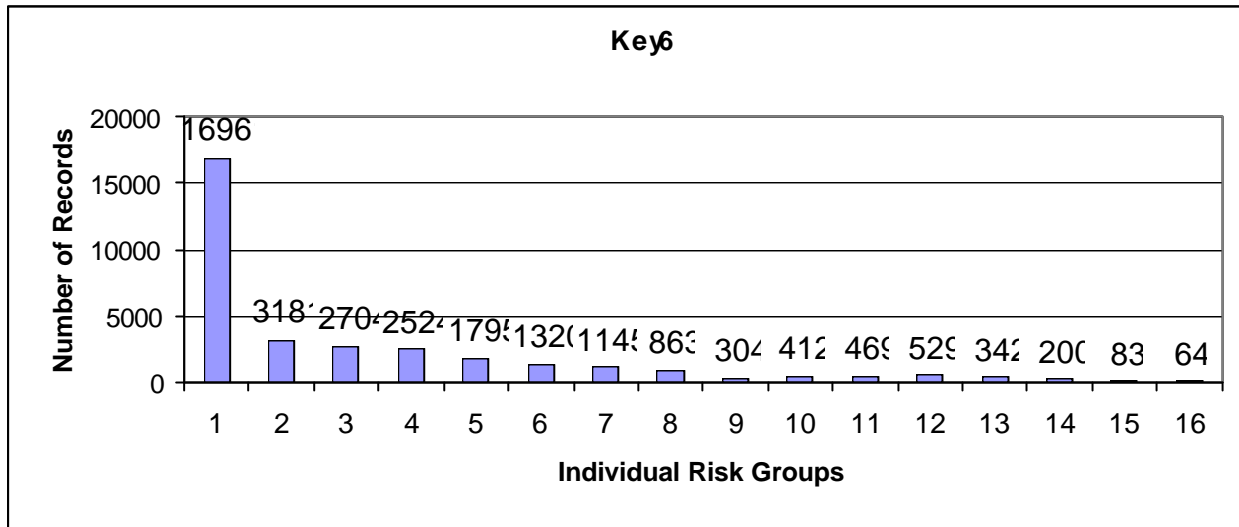


Figure 3: Number of records in groups according to individual risk measures for key6 - IS



III.B. Linking Sample Uniques to the NPR

The sample uniques of the IS for three different sets of keys were linked to the NPR. The keys used were the full key, Key1, which defines the current PUF data file, and Key4. The keys were slightly modified to accommodate the NPR.

The results of the three linkages were as follows:

Table 2: Results of linking sample uniques of the IS to the NPR

	Full Key	Key1	Key4
Number of keys in sample	24,028	19,843	9,838
Number of uniques in the sample	20,528	15,411	6,031
Estimated global risk measure - \hat{t}	973.1	726.5	291.2
Number of keys in NPR	1,348,870	467,167	107,663
Number of uniques in the NPR	942,875	213,678	32,004
Percentage of uniques in the NPR	69.9%	45.7%	29.7%
Number of sample uniques matched	14,359	13,762	5,669
Number of sample uniques unmatched	6,169	1,649	362
Average number of correct matches	4,600.3	2,118.5	411.5
Number of matches one-to-one	2,829	848	101

In table 2, the estimated global risk measure \hat{t} is largely underestimated compared to the average number of correct matches to the NPR according to this scenerio, though it does maintain its monotonic property. The comparison of the two measures, however, is problematic because of the following reasons:

1. Some of the values of the keys for sample uniques were not linked at all to the NPR, from 6% to the most collapsed key4 to 30% for the full key. This is most likely due to measurement errors and other factors in both the NPR and the sample data, which increases as the key in use is more detailed. It is important to note that the sample was drawn by sampling addresses from municipal tax records and thus is not a direct sub-file of the NPR. In addition, the NPR suffers from serious coverage problems since about 20% of the population do not reside in the same address as the one listed in the NPR. It also includes population groups not covered in the target population of the survey such as institutionalized persons.
2. For those matching one-to-one, the names on the frame which was used to draw the sample was compared to the names on the NPR in order to get some idea as to the true matching status of the keys. This comparison is also problematic, not only because of the above mentioned problems with the NPR, but because the names listed on the frame are those persons who pay the municipal tax bill and not necessarily the persons living in the dwelling or other family members. Nevertheless, with all these inconsistencies between the data sources, it was found that among those with one-to-one matching status, about 40% -50% were identified as the same person.

In spite of the discrepancies between the data sources, it is clear that the estimated global risk measures are underestimated under this scenerio and the sample weights which are used to estimate the population frequencies do not provide the variability of the keys as compared to the population. From the point of view of the data snooper, reidentification is highly likely and any attempt to misuse the data for commercial purposes will probably be very successful.

IV. Measuring Information Loss

The information loss due to collapsing the categorical variables which make up the different keys can be measured by several methods. Since the variables that define the keys are demographic and geographic identifiers and are used by researchers mostly as explanatory variables in regression models, we will assess the information loss using two methods:

1. The loss in variance between the different groupings of the keys as they get coarser for the main variable of interest income. In other words, the loss of the predictive power of a regression model, as expressed by the R-square, where the dependent variable is income and the independent variables are the demographic and geographic variables that are collapsed.
2. The loss in information as the keys get coarser and the categorical variables are collapsed as measured by the entropy (Willenborg and de Waal (2001)).

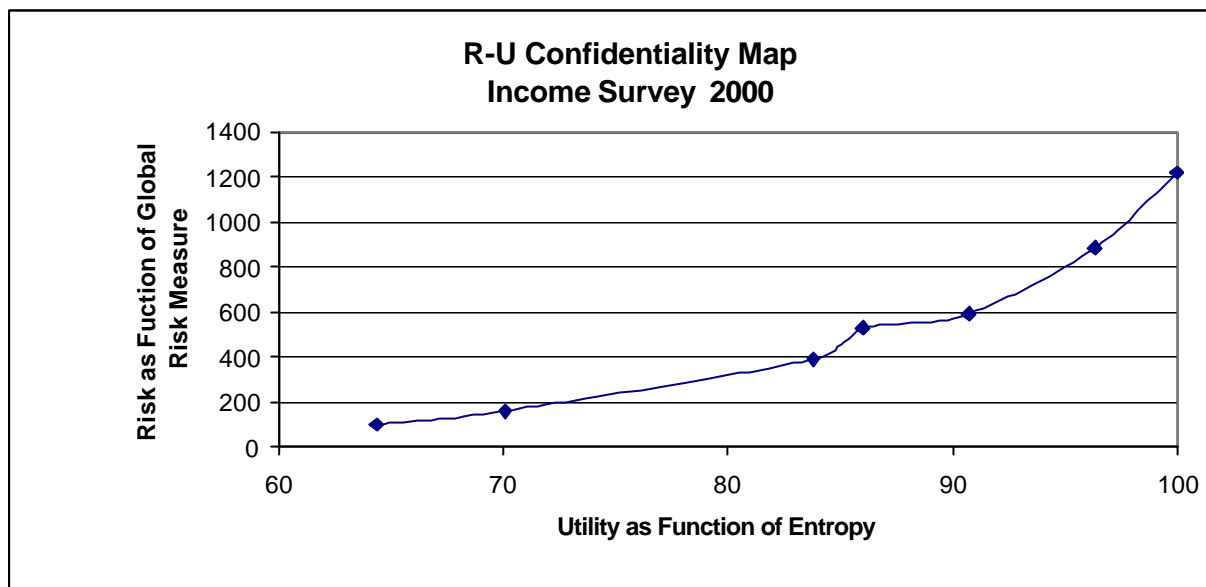
The following results are obtained:

Table 3: Information loss measures for the IS

Key	Percentage of the "between" variance out of the full key	Percentage of the entropy out of the full key
Full key	100.000	100.000
Key1	96.065	96.318
Key2	90.642	90.710
Key3	91.199	85.992
Key4	84.472	83.801
Key5	75.180	70.083
Key6	70.240	64.349

An empirical R-U confidentiality map for the IS can be built using the disclosure risks from the previous section and the loss of information in table 3 to express the data utility:

Figure 4: Empirical R-U confidentiality map for global recoding of key variables in IS



On the basis of this chart and the setting of thresholds, we can determine which of the files according to the different keys should be PUF, MUC or accessible to the public via remote access to the internet.

V. Remote Accessing of Data via the Internet

As seen by the analysis of the IS, since current PUF files defined by Key1 had about 2.7% estimated global risk out of the total sample size, those files placed on the web site of the depository managed by the Hebrew University for remote accessing and downloading were deemed unsafe for internet access. This is true for other microdata files that were examined and are currently on the depository's web site, including the Israel Labour Force Survey and the Family Expenditure Survey. New files are being prepared based on the recoding of demographic and geographic variables. For the interim period, since the depository still has restricted access to researchers in universities and research institutions, the threshold for the estimated global risk measure out of the total sample size was set at about 1%. At this threshold, the information loss based on the entropy with respect to the full key is about 80.3% for the IS.

With respect to the system that is currently being developed at the CBS for building custom-made tables defined by the users over the internet, the question remains as to what file should be in the background of the system. A pure and safe PUF file with no risk would allow users to access any table and allow downloading of the file. This would also greatly simplify the amount of software needed to develop the system. However, as shown in the example of the IS, a file with no risk would have an information loss based on the entropy with respect to the full key at about 55.0% and would probably have little value to researchers.

More useful tables can be produced by putting a more detailed file in the background of the system, but this would mean some or all of the following restrictions depending on the initial risks in the file:

1. Minimum number of units in a cell. This is also necessary to assure the estimate's reliability in the table with respect to sampling errors.
2. Maximum number of dimensions to the table.

3. Prior registering and tracking of users.
4. Sophisticated disclosure control software for calculating the disclosure risk prior to the release of the table.

For example, if we put the file for the IS with Key1 in the background of the internet system, and allowed users to access tables of up to three dimensions, the risk of a table defined, for example, by total income according to continent of birth \times years of birth \times locality code would have an estimated global risk measure of 264.1 expected correct matches to the population for the sample uniques. The amount of protection needed for providing disclosure control of the table would be very high and after collapsing on the variables and redesigning the table would probably result in the same table that would have been obtained had a more safe file been in the background of the system. By putting the file with key2 in the background of the system, the estimated global risk measure would be at about 30 for the most elaborate three dimensional table. This would perhaps simplify some of the restrictions necessary for protecting the confidentiality of sample entities, and still allow as little information loss as possible. Thus a compromise has to be found between selecting a file for the internet system that will allow users remote access to data with high utility for building customized tables but without having to define complicated restrictions to the system or to develop elaborate software applications that would be necessary to maintain disclosure control.

VI. Acknowledgements

I wish to thank Prof. Yosi Rinott for his assistance in the paper, especially for the theoretical framework in the second section.

VII. References

Benedetti, R., Capobianchi, A, and Franconi, L. (2003) "Individual risk of disclosure using sampling design information" (forthcoming).

Domingo-Ferrer, J., Mateo-Sanz, J. and Torra, V. (2001) "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", ETK-NTTS Pre-Proceedings of the Conference, Crete, June 2001.

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001) "Disclosure risk vs. data utility: the R-U confidentiality map", Technical Report LA-UR-01-6428., Statistical Sciences Group, Los Alamos, N.M.:Los Alamos National Laboratory.

Kamen, C., (2001) "Control of statistical disclosure versus needs of data users in Israel: a delicate balance", Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, March 14-16, www.unece.org/stats/documents/2001/03/confidentiality/29.e.pdf .

Polettini, S. and Seri, G. (2003) "Guidelines for the protection of social micro-data using individual risk methodology – Application within mu-argus version 3.2", CASC Project Deliverable No. 1.2-D3, www.neon.vb.cbs.nl/casc

Rinott, Y. (2003) . "On disclosure risk measures", Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxemburg, April 7-9, www.unece.org/stats/documents/2003/04/confidentiality/wp.16.e.pdf

Willenborg, L. and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155 (Springer Verlag, New York).