Contributions to Methodology and Statistics
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 10, Ljubljana: FDV 1995

Disclosure Risk and Sample of Anonymized Records

Mirna Macur and Vasja Vehovar[1]

Abstract

The disclosure problem relates to the possibility of identifying individuals in the released statistical information. The paper evaluates the disclosure risk on a 3% sample of individual data from the Slovene 1991 Population Census. The concept of uniqueness is used for this purpose. The level of regional aggregation, the number of identifying variables and the grouping of the categories are discussed as the parameters of the disclosure risk.

Keywords: Sample; Confidentiality; Disclosure.

1 Introduction

The term *confidentiality* typically relates to the freedom of the individual to decide how much of the self is to be revealed to others, when and to whom. On the other hand, confidentiality has also become an important feature of the data because, in general, not all the information about individuals can be released without restrictions. The issue of confidentiality is especially important in the context of "the information society", as new technologies create new and powerful means of using (and misusing) the individual data.

The issue of confidentiality is of great relevance for statistical agencies since they collect individual data. Of course, the statistical use of the data means that the interest is not in individual values themselves but in aggregates or parameters (averages, totals, correlations, trends). However, care should be taken whenever the information which is based on individual values is released.

Most often, the statistical data are presented in the aggregated form, that is, in the form of cross-tabulations. The alternative is the form of the microdata file in

[1] Faculty of Social Sciences, University of Ljubljana, P.O. Box 47, 61109 Ljubljana, Slovenia. The authors wish to thank the Statistical Office of the Republic of Slovenia for the data and for the assistance needed for this research.

which the individual records are released to the public. Of course, in both cases there exists the possibility of identification of individual data. The recognition of the individual would be extremely inconvenient since the individual data are, in principle, confidential. For this reason the issue of confidentiality is generally regulated by law.

- In the USA, for example, The Privacy Act of 1974 protects all forms of individually identifiable information. By this Act all agencies holding individual data are obliged to report their data-base system to the Federal Register. The Privacy Act also gives the right to persons whose records are on the file to examine and correct the data. The penalties for misuse have been also greatly increased (Cecil, 1993).
- In Italy, until 1989, the legal protection of personal data was essentially regulated by the law governing official secrets. The new Statistical Act of 1989 states that the individual data gathered by the statistical office can be used only for statistical purposes and cannot be disseminated to any external agent. Thus it is not possible to obtain any individual data, and the external dissemination of individual data is, in general, forbidden (Biggeri and Zannella, 1991).
- In Germany *absolute anonymity* is required for the release of the statistical data. The microdata may be released only if identification is absolutely impossible. However, the Federal Act on Statistics, 1987, introduced an exception: the use of microdata for scientific purposes where the *factual anonymity* is required. Factual anonymity means "that the data can be linked to the respondents only by employing an excessive amount of time, expenses and power" (Luttinger et al., 1993: 218), so that the identification risk is extremely low. A similar exception for release of the individual data exists in the practice of many countries, including USA and Italy.

Slovenia is one of the few countries that has data protection entrenched in the Constitution. The corresponding legislation is also relatively restrictive. Specifically, *The Law on the Protection of Individual Data*, passed in 1991, says "... the holder of a data base can only give the requested data in a form that does not enable identification of individuals, to whom the data refer...." (Article 10). Similar restrictions are included in the Act on State Statistics which is to be passed in 1995. However, there does exist an exception - similar to other countries - with regard to the scientific use of the data.

Despite the fact that the legislative acts generally forbid the release of data in a form that enables identification, it is extremely difficult to make this rule operationally effective. For example, in a Census the data may be collected about the land owned by individuals. Even if the data were published in the aggregated form on the level of settlements, the individuals could be identified when knowing there was only one person from a settlement owing a certain type of land. Similarly, a risk of identification exists if the data are released in a microdata file with the identifying variables removed.

We are thus faced in practice with the problem of the highest level of acceptable risk related to the identification of the individual data. Even more important, however, is the problem of defining and measuring this risk. In the remainder of this paper we will discuss these problems, though restricting the discussion only to the microdata form of the released statistical information. The problems with the aggregated data will not be addressed here.

2 Disclosure

There is no common definition of a disclosure, and not all researchers make their concept of disclosure explicit. Some authors define disclosure on the basis of individual data, others on aggregated data, and there exists a variety of other differences.

We will use a general definition of a disclosure proposed originally by Dalenius (1977) and slightly reworded by Steinberg (1983):

"If the release of certain statistical information makes it possible to determine a particular value relating to a known individual more accurately than it would be possible without access to the data, then a disclosure has taken place (Fienberg, 1992)."

The risk of a disclosure is therefore determined by the extent to which a released record can be linked to a respondent.

We can distinguish two major types of a disclosure (Lambert, 1993: 315):

a) *An identity disclosure* or *identification* occurs when the individual is linked to a particular record in a released file; we can talk about the recognition of the unit. Identification is also possible from the tabular data when there is only one record from the population that falls into a certain aggregated cell. Even if the intruder learns nothing new from the identification, the identification itself may compromise the security of the data file.

b) *An attribute disclosure* occurs when something new has been learned about the individual unit. An attribute disclosure may occur with or without identification. For example, we may learn about the value of the sensitive attribute by locating the unit within a very narrow aggregated group with values similar to those of sensitive variable.

We will consider only the attribute disclosure which is based on an identification. Thus, the risk of disclosure will be the risk of identifying a released record and the harm from disclosure depends on what is learned from the identification.

More formally, we can define the *disclosure* in the case of *microdata* as follows (Biggeri and Zannella, 1991). Let us have a microdata file with the following structure:

$$d_j = ((B_{j1}, B_{j2},...); (I_{j1}, I_{j2},...); (O_{j1},O_{j2},...); (S_{j1},S_{j2},...)),$$

where **d** represents a record and **j** (j = 1,2,...n) represents an individual in the microdata file. Subsets B(d), I(d), O(d), S(d) represent different types of variables in a record:

- B(d) - basic or direct identifiers of the individuals (name, address, social security number),
- I(d) quasi - identifying or surrogate variables; these are the variables which are in the public domain, such as sex, age, place of residence. The I-set may also be referred to as *identification* variables or *key* variables. They allow one to identify a record, that is, to establish a one-to-one correspondence between the record and a specific individual (Bethlehem et al., 1991).
- O(d) - ordinary or common variables; these are usually not included in the public registers or public listings, however, they also are not considered as confidential, such as, for example, a special field of study or availability of domestic appliances.
- S(d) - sensitive variables; these contain the values such as nationality, religion or social status belonging to the private domain, which the respondents would not like to be revealed.

In practice, however, no general definition exists for the variables to be considered sensitive. Sometimes, the same variable may be either identifying or sensitive depending on the cultural, social and political context of the country. Typical sensitive variables are those relating to sexual attitudes, to the use of alcohol, drugs and to information which, if released, could reasonably be considered damaging to an individual's financial standing, employability, or reputation. In the case of Census data there may exist several sensitive variables; here, we define nationality and religion as being typical.

Identification or identity disclosure occurs when key variables (I) from the record are linked with a certain individual. We assume that the identification is correct and thus the problem of false identification will not be considered. As a consequence of an identity disclosure, the sensitive variables (S) may be revealed and this can lead to an attribute disclosure. Thus we understand violation of confidentiality to have occured when data on sensitive variables are released and it is possible to associate them with a specific individual.

Disclosure depends on a variety of *circumstances and factors* (Biggeri and Zannella, 1991):

- the size of population (N);
- the type of the data collection (sample survey or complete enumeration - census);
- the type of elementary units involved (individuals, households, enterprises, etc.);
- the size of the microdata (n) set to be released in proportion to the population;
- the number and type of the key variables and their categories;
- the distribution of population over the classes, determined by categories of key variables;
- the number and type of sensitive variables;
- the relation between key, ordinary and sensitive variables;
- the type of possible intruder (a person looking for disclosure): journalist, etc;
- the possible knowledge of the intruder about the population elements;
- the ways in which the intruder tries to obtain disclosure;
- the costs and losses that the intruder incurs in attempting to obtain the disclosure;
- the potential benefits of the disclosure to the intruder;
- the extent and distribution of errors (noise and biases) in the data sets.

Obviously, there are many factors influencing the disclosure risk. Some of these are difficult to control or predict, especially the characteristics typical of intruders. However, we will not consider the aspect of intruders in our discussion. This will further restrict our analysis to the following ways in which a disclosure may occur (Biggeri and Zannella, 1991):

- *matching* of the records of the released microdata set with those of an external microdata set derived from any register;
- *spontaneous* recognition with reference to the so-called 'rare persons';
- using *response knowledge*, i.e. the knowledge of persons who have participated in a survey and whose data are contained in the released microdata file.

Disclosure risk we understand as being the likelihood of disclosure occurring. The following discussion will be restricted to the risk of identity disclosure. We will further limit our discussion to identity disclosure in the case of spontaneous recognition since this is the most common way disclosure occurs when microdata from the Census are released. We will use the concept of *uniqueness* to evaluate the above mentioned disclosure risk from microdata.

As mentioned, the key variables are the variables which - taken together - may contribute to the linking of a record to its respondent. Thus, the categories of key variables define a variety of very specific cells, for example: a cell comprising exclusively woman, living in Ljubljana, 40-50 years old, graduated, having three children. If there is only one unit (person) in such a cell, we denote this unit as *unique*. Of course, there exists a danger of spontaneous recognition (identification) of such person, and this may be followed with the linking to other sensitive data from the same record, for example the religion of such a person. Obviously, such units possess a high disclosure risk.

We define a *population unique* as a record which stands out alone in a combination class determined by categories of the key variables. Of course, if we have a microdata file that contains only a sample of individuals from the population, we have to distinguish also the *sample unique* record. Though the sample uniques in a microdata file are not necessarily the population uniques, they still possess a high disclosure risk.

We will use the concept of sample uniqueness to quantify the disclosure risk. The percentage of the records that are unique will thus serve as a specific measure of the disclosure risk. Of course, since we only have a sample of the population (i.e. Census) the disclosure risk based on sample uniques will overestimate the disclosure risk of the population uniques. We will discus the relation between the sample unique and the population unique in the conclusion.

In the next section we present the empirical evaluation of this risk in a case of a 3% sample from the Slovene Census.

3 Empirical evaluation

3.1 Sample from census data

The Population Census is an extremely precious source of information. Its basic advantage compared to the other survey data is the fact that in the Census the whole population has been surveyed. This enables one to make a reliable analysis of the small areas and also a detailed analysis of marginal subgroups.

Despite the enormous amount of information collected in a Census, the statistical offices generally have to grapple with the fact that the Census data are not optimally exploited. Since the Census is extremely expensive, new ways are being sought for using and marketing the Census product. According to the United Nations 1990 survey among 97 countries (Dekker, 1991) performing the 1991 Census round, there are two basic forms of revealing the Census data: the aggregated form and the microdata form. Most often the standard aggregated form consists of cross-tabulations, however, more and more countries also reveal the Census data in the form of small-area aggregates. Generally, the smallest aggregate is a census enumeration area, however, even the aggregates at the level of five households are sometimes released. On the other hand about 20% of all countries also use a microdata form to release the Census data, i.e. they create a sample of anonymized individual records from the Census.

The microdata form is often preferable for many users because it provides them with full flexibility in performing analyses: they can create their own cross-tabulation and they can build their own models. Thus, there are many possible applications for a sample of anonymized records from the Census (Marsh et al., 1991), such as economic policy and labour market analysis, demography, population forecasting and social policy, health policy, housing research, marketing research. Well known countries which release samples from the Census are USA, Canada, Great Britain and Italy. The usual sample size is about 2%.

3.2 Data and method

We will use a 3% systematic sample of the individual records from the Slovene Census 1991. The sample consists of n=60,578 persons. Due to missing values we additionally exclude some records, so that we end up with n' = 58,704 complete units. The variables name, address and identification number were omitted from the records, and the categories of age and occupation were further collapsed as described below. The structure of the record in the microdata corresponds to the Census form and can be found in the Appendix.

Firstly, we define the key variables. These are the variables that allow identification of an individual. We used the standard set of key variables (Biggeri and Zannella, 1991):

- Sex (male, female),
- Marital status (single, married, widowed, divorced),
- Education (14 categories),
- Occupation (58 occupational classes - two digit classification),
- Age (17 categories - 5-year bands),
- Community (62 categories),
- Regions (12 categories).

Our goal is to use the above key variables for calculating U, the number of unique records in a sample at the different levels of analysis (Slovenia, regions, communities). We were not concerned here with the problem of collapsing the categories of sensitive variables what might also become an issue when microdata would be actually released to public.

We will study the disclosure risk under the two variable conditions:

- excluding one or more key variables,
- collapsing the categories of the key variables.

3.3 Results

The tables below summarize the extensive calculations needed to evaluate U, the number of uniques. The number of the categories of a key variable can be found in brackets after the variable. We use the following abbreviations:

- A - age
- S - sex
- O - occupation
- E - education
- M - marital status

Table 1: Uniques (U, U%) when varying the number of key variables included

Level of analysis	SLOVENIA		12 REGIONS		62 COMMUNITIES	
Key variables	U	U%	U	U%	U	U%
$A_{(17)}$ * S(2) * $O_{(58)}$ * $E_{(14)}$ * MS(4)	2543	4.3	12314	20.9	22308	38.0
$A_{(17)}$ * S(2) * $O_{(58)}$ * $E_{(14)}$	1288	2.2	7686	13.1	17353	29.6
$A_{(17)}$ * S(2) * $O_{(58)}$ * MS(4)	587	0.9	3995	6.8	11463	19.5
$A_{(17)}$ * S(2) * $E_{(14)}$ * MS(4)	106	0.2	1917	3.3	7182	12.2
$A_{(17)}$ * S(2) * $O_{(58)}$	139	0.2	1863	3.1	7180	12.2
$A_{(17)}$ * S(2) * $E_{(14)}$	13	0.0	507	0.9	3022	5.1
$A_{(17)}$ * S(2) * MS(4)	3	0.0	108	0.2	941	1.6
$A_{(17)}$ * S(2)	0	0.0	0	0.0	26	0.0

In Table 1 we find U, the total numbers of units (persons) that are unique given a specific set of key variables. In other words, the U is a number of units that are alone in a combination class, determined by categories of key variables. The corresponding percentage U% is calculated as a simple proportion of the unique records in the total sample of n=58,704.

We can observe that in the case of Slovenia the number of uniques amounts at most to U% =4.3. This becomes much higher if we add regions or communities: the percentage of uniques can reach up to U%=20.9 for regions or even U%=38.0 for communities - in this case every third unit is a sample unique. On the other hand, we can observe a decline in the percentage of uniques when some key variables are excluded from the analysis. Just by excluding the occupation variable, for example, we can reduce the percentage of uniques at the level of communities from 38.0% to 12.2% and at the regional level from 20.3% to 3.3%. As the variables Occupation, Education and Community have considerable numbers of categories they determine individuals in a very narrow way and thus create a high percentage of uniques in the sample.

Another important factor influencing the increase in the number of uniques is the distribution of the population over the categories of the key variables. Symmetrical distribution of a certain variable across its categories creates a smaller impact on the risk of disclosure compared to the variables with more asymmetrical distribution. In this context, for example, the variables of sex and age are symmetrical as opposed to marital status, education and occupation which have some categories containing relatively few elements.

Immediately, of course, we face the following question: *What is the acceptable level for the percentage of uniques?* Regrettably, there is no clearly defined border. However, we can observe in the literature (Biggeri and Zannella , 1991) that the results below U%=1 are considered low and acceptable, but the percentages above U%=1 are considered as high when dealing with 2-3% samples from the Census of population.

A sample unique may occur because of the presence of the real population unique. The Slovene 3% sample thus contain (approximately) every 33rd population unique. Additional to this, a sample unique may also occur when a combination class in a population is relatively small - around 33 persons, so that only one unit from the combination class falls into the sample. In any case, having one percent of the uniques in a sample means a relatively low risk of disclosure. Additional to this, even when the sample unique is a truly a population unique (but we can not figure that out from the sample itself) there are many practical obstacles that make the actual disclosure much less possible.

We can observe from Table 1 that one strategy for lowering the number of uniques is simply to omit a certain key variable from the analysis. However, if we happen to need this variable, this will be very inconvenient. We have another strategy - collapsing the categories of the key variables. In Table 2 we can observe some of the many possibilities for collapsing the categories.

From Table 2 we can observe the percentages of unique records for different numbers of categories of the key variables. For this purpose the 17 age groups were collapsed to 7 ten-year groups. Similarly, the two digit occupation groups were collapsed to a one digit classification with 9 categories, and the categories of education were reduced to 8 instead of 14 categories.

Table 2: Uniques (U%) when varying the categories of the key variables

Key variables and their categories	SLOVENIA	REGIONS	COMMUNITIES
$A_{(17)}$ * S(2) * $O_{(58)}$ * $E_{(14)}$ * M(4)	4.7	20.9	38.4
$A_{(7)}$ * S(2) * $O_{(9)}$ * $E_{(8)}$ * M(4)	0.5	4.9	15.8
$A_{(7)}$ * S(2) * $O_{(9)}$ * $E_{(8)}$	0.1	2.0	8.8
$A_{(7)}$ * S(2) * $E_{(8)}$	0.0	0.1	1.1

The percentages in Table 2 are, of course, much lower than the corresponding percentages in Table 1. So - with the proper collapsing of the categories - we can obtain an acceptable number of uniques also at the regional level. However, the numbers of uniques are still very high at the level of communities. Thus, at the community level even the collapsing of the categories is not very helpful.

Obviously, the level of the analysis is extremely important, for the issue of disclosure and the number of units in a certain geographic region has a great impact on the number of uniques. There is another practical rule that can be observed in the literature - and also in statistical practice: the regions smaller than 100,000 persons should not be identified in a microdata file of a few percents from the Census (Griffin et al., 1991). We can confirm this rule also in the case of Slovenia, since the number of uniques is very high when dealing with communities (numbering on average population of 30,000 persons).

3.4 Usefulness of microdata

The problems of a disclosure at the regional level force us to omit some key variables and/or to intensively collapse their categories, and the high disclosure risk at the community level even prevents the release of the community level microdata. These are, of course, severe disadvantages from the point of the usefulness of the data. Usefulness thus clearly contradicts the request for the confidentiality.

To observe the notion of the usefulness of the microdata more closely, we will express the quality of the information from the microdata file in terms of the relative precision. The relative precision of the percentage obtained from the simple random sample can be expressed as the coefficient of variation CV(p):

$$CV(p) \approx \surd(p(1-p)/n)/p.$$

We can easily link the CV(p) with the standard ($\alpha=0.05$) confidence intervals:

$$P = p \pm 2pCV(p).$$

For example, with a percentage of unemployed in the total population $p=0.05$ we obtain $CV(p)= 1.8\%$ and a confidence interval $P=(5.0\pm0.2)\%$, however at the community level we have $CV(p) \approx 10\%$ and $P=(5.0\pm1.0)\%$ which is much less precise than in the case of the whole population. We will use the following approximate but practical rules to evaluate the obtained precision:

$CV < 5\%$ good estimate,
5 - 10% acceptable estimate,
$CV > 33\%$ totally unacceptable estimates.

In Table 3 we observe the precision of the estimates in Slovenia, in a typical region Goriška and in a community Ptuj. We observe four levels of the target percentage p.

Table 3: Precision of the estimates CV(p)% for a 3% sample from the Slovene Census

Level of analysis	Sample size	p=0.01	p=0.05	p=0.10	p=0.50
Slovenia	n=58704	4.1%	1.8%	1.2%	0.4%
R- Goriška	n=3606	16.6%	7.3%	4.9%	1.6%
C- Ptuj	n=2012	22.2%	9.7%	6.7%	2.2%

Good estimates were obtained for the characteristics with $p=0.5$ (for example gender variable). Precision is also acceptable for $p=0.1$ (for example, Non-Slovenes) and $p=0.05$, (widowed, unemployed, families with three children, farmers). However for small subgroups $p=0.01$ (some religious or ethnic groups) the precision of the estimates in a 3% sample is no longer satisfying. To analyze small proportions we need a larger sample. As an example, we can compare in Table 4 the improved precision of the 10% sample from the Census. We can observe a considerable improvement in precision CV(p), especially in the case of small proportions, i.e. $p=0.01$.

Table 4: Precision of the estimates CV(p)% for 10% sample from the Slovene Census

Level of analysis	Sample size	p=0.01	p=0.05	p=0.10	p=0.50
Slovenia	n=58704	2.2%	0.9%	0.6%	0.2%
R- Goriška	n=3606	9.0%	3.9%	2.7%	0.9%
C- Ptuj	n=2012	12.1%	5.3%	3.6%	1.2%

4 Conclusions

- We measured a disclosure risk as a percentage of sample uniques. When using a 3% sample from the Slovene 1991 Census the disclosure risk was acceptably low at the level of Slovenia. It was still relatively low at the regional level if we collapsed the categories of the key variables. However, at the community level, the disclosure risk was much too high. If communities in a sample were wanted we should intensively collapse the categories of key variables and also exclude some of them from the microdata. However, this might be seriously damaging for the usefulness of the data.

- A larger sample would be needed for reasons of precision. The relative precision shows that for characteristics which account for only a few percent of a population - which are often of interest to us when dealing with Census data - the estimates may not be acceptable even at the regional level. It seems appropriate then to take a larger sample from the Slovene Census. Of course this may have a negative impact on the confidentiality of the data.

- We are inevitably faced with the conflict between the quest for more detailed information on the one side (community level analysis, detailed categories of key variables) and the demands of the confidentiality of the individual data on the other side. To find a practical solution to the above dilemma we found that methods for evaluating a disclosure risk lacked a more standardized approach. Specifically, we were missing the operational rules to assess the disclosure risk and a clear definition of the level of acceptability of the disclosure risk.

- We should repeat again that the smaller percentages of unique cases doesn't necessarily mean a smaller disclosure risk. The larger the subsample the more an intruder could be confident that a unique record in the sample is unique in the population. So a larger subsample with a lower percentage of unique cases could be more dangerous than a small subsample with a high percentages of unique records. Nevertheless, given a fixed set of population records and a fixed set of key variables the larger sample will definitely lower the number of sample uniques. Of course in the case of a larger sample the overestimation of the population uniques will be smaller than in the case of a smaller sample. In the paper the U%=1% level, which is implicitly used in other studies, was applied as the critical value for the percentage of the sample uniques in the 2-3% sample from the Census. Obviously, further research is needed for calculating the population uniques (Biggeri and Zannella, 1991).

Appendix

Structure of the record in the 3% sample from Slovene Census

VARIABLES ON INDIVIDUAL DATA			NUMBER OF CATEGORIES
1)	OBC	community	62
2)	S	sex	2
3)	M	marital status	4
4)	OT_7	nb. of children	19
5)	NR_8	nationality	43
6)	VR_9	religion	72
7)	JZ_10	mother tongue	39
8)	PG_27	language spoken in a family	48
9)	PG_28	language spoken in the area	60
10)	E	educational level	14
11)	SO_112	the name of the school	794
12)	PI_113	literacy	3
13)	SL_12	school that he/she now attends	5
14)	OC_131	father occupation	58
15)	MA_132	mother occupation	58
16)	TU_14	foreign country	26
17)	ZD_15	nb. of years living abroad	41
18)	O	occupational group	58
19)	PL_162	(un)employed	9
20)	DH_17	salary	30
21)	PV_181	occupation - keeper	58
22)	DP_191	social status	5
23)	OL_192	a kind of a property	4
24)	DE_20	activity	XXXXXX
25)	SI_21	level of qualification, education	10
26)	DL_22	free-time activity	3
27)	DM_23	place of work/school	3
28)	MI_241	frequency	4
29)	MI_242	returning from work/school	6
30)	MI_243	spent time in minutes	300
31)	LT_25	nb.of years working abroad	40
32)	LV_26	year of return	91
33)	OD_A	reason of presence/absence	9
34)	OD_B	nb. of a family	4
35)	PC_C	relation to a family head	9
36)	A	age	18
37)	PTIP	type of area	2
38)	STP_1	type of apartment	3
39)	SUP_2	use of apartment	9
40)	SLS_9	property of apartment	4
41)	SNS_10	flat	20
42)	SLT_11	year of construction	5
43)	SGO_A	nb. of households	79
44)	SGO_B	nb. of persons in a apartment	835
45)	SGO_C	nb. of persons in a household	79
46)	GLS_1	apartment is used as	7
47)	GKG_19	farm-house	2
48)	VNAS	size of settlement	6

References

[1] Betlehem, J. G., Keller, W. J., and Pannekoek, J. (1990): Disclosure Control of Microdata. *Journal of American Statistical Association*, **85**, 38-45.

[2] Biggeri, L. and Zannella, F. (1991): Release of Microdata and Statistical Disclosure Control in the New National Statistical System of Italy: Main problems, Some Technical Solutions, Experiments. 48th ISI. Session Cairo, September 9-17, 1991.

[3] Cecil, J. S. (1993): Confidentiality Legislation and the United States Federal Statistical System. *Journal of Official Statistics*, **9**.

[4] Čebulj, J. (1992): *Varstvo informacijske zasebnosti v Evropi in v Sloveniji.* Ljubljana: Inštitut za javno upravo pri Pravni fakulteti v Ljubljani.

[5] Dekker, A. L. (1991): New or Uncommon Computer Methods in Population Census Data Processing. 48th ISI. Session Cairo, September 9-17, 1991.

[6] Duncan, G. and Lambert, D. (1989): The Risk of Disclosure for Microdata. *Journal of Business and economic Statistics*, **7**, 207-217.

[7] Fienberg, S. E. (1992): *Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality.* Paper prepared for the International seminar for Statistical Confidentiality. September 8-10, 1992, Dublin.

[8] Greenberg, B. and Voshell, L. (1992): *Relating Risk of Disclosure for Microdata and Geographic Area Size.* Paper prepared for the International seminar for Statistical Confidentiality. September 8-10, 1992, Dublin.

[9] Griffin, R. A. and Navarro, A. (1991): *1990 Census Public Use Microdata Sample Design Issues.* Paper prepared for the International seminar for Statistical Confidentiality. September 8-10, 1992, Dublin.

[10] Lambert, D. (1992): Discussion paper prepared for the International seminar for Statistical Confidentiality. September 8-10, 1992, Dublin.

[11] Lambert, D. (1993): Measures of a Disclosure Risk and Harm. *Journal of Official Statistics*, **9**, 275-313.

[12] Luttinger P., Wirth H., and Hippler H (1993): Special Aspects in Using MIcrodata: Data Anonymity and Non-Response Bias: Research at ZUMA. *Journal of Official Statistics*, **9**, 217-222.

[13] Marsh, C. et al. (1991): The Case for Samples of Anonymized Records from the 1991 Census. *Journal of Royal Statistical Society A*, **154**, 305-340.

[14] Reynolds, P. D. (1993): Privacy and Advances in Social and Policy Sciences: Balancing Present Costs and Future Gains. *Journal of Official Statistics*, **9**.

[15] Vehovar, V. (1992): Vzorec popisih podatkov - izzivi in tveganja. *Bilten Statističnega društva Slovenije*, **31**, 12-13.

[16] Wright, D., Ahmed, S. (1992): *Implementing NCES'New Confidentiality Protections*. Paper prepared for the International seminar for Statistical Confidentiality. September 8-10, 1992, Dublin.