MEASURES FOR INFORMATION LOSS IN PROTECTED DATA

Mira Todorova [*]

[*]State Statistical Office of the Republic of Macedonia, e-mail: todorova.m@stat.gov.mk

Abstract: Statistical offices that disseminate their data to the public should take care of the implementation of Statistical Disclosure Control (SDC) to protect their data.

The major risk of different SDC techniques is loss of information, which, in some circumstances, may result in less useful, protected data. The paper gives an overview of different methods, which can be used in evaluating the usefulness of protected tables prior to publishing.

Besides that, the state of play of application of statistical disclosure measures for information loss in the national statistical system is given.

Introduction

The main purpose of Statistical Disclosure Control (SDC) is to allow dissemination of statistical data in such a way that they do not give away confidential information, which is related to specific individual or enterprise. SDC should modify data in such a way that the risk of disclosing information on specific respondents becomes low enough while keeping at a minimum the information loss. It is of crucial importance for statistical organisations to apply SDC, which will be able to compare and trade off information loss and disclosure risk to reach a suitable balance.

Information loss measures can be split into two classes: measures for data suppliers in order to make informed decisions about optimal SDC methods which depend on the characteristics of the tables, and measures for users in order to allow adjustments to be made when carrying out statistical analysis on protected data.

The first part of the paper reviews the methods that give adequate protection against disclosure. An overview of measuring information loss in the Macedonian statistical system is given in the second part of the paper.

1. Masking Methods for Microdata Protection

SDC should supply to the users of statistics a masked microdata file V' similar to the original file V in such a way that:

1. Disclosure risk (i.e. risk of identification of an individual) is low.
2. User analyses (means, regressions, etc.) on V' and on V result the same or at least similar results.

This section describes masking methods that can be used to produce V' from V. The main literature used for this chapter is written by Josep Domingo-Ferrer and Vicenc Torra (2001) and Willenborg and De Waal (2001).

1.1 Microdata Protection Methods

Microdata protection methods can generate the protected microdata set V':

- Either by masking original data, i.e. generating V' as modified version of the original microdata set V;
- Or by generating synthetic data V' that preserve some statistical properties of the original data V.

Masking methods can be divided in the following two categories (Willenborg and DeWaal, 2001):

- *Perturbative.* The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset.
- *Nonperturbative.* Nonperturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail on the original dataset.

From the point of view of used data, the following classification applies:

- *Continuous*. A variable is considered continuous if it is numerical and arithmetic operations can be performed with it. Such examples are income and age.
- *Categorical.* A variable is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Examples are day of the week and eye colour.

1.1.1 Perturbative Methods

Perturbative methods allow for release of the entire microdata set, although perturbed values rather than exact values are given.

Most perturbative methods stated below are special cases of matrix masking. If the original microdata set is V, then the masked microdata set V' is computed as:

$$V'=AVB+C$$

Where A is a record-transforming mask, B is a variable-transforming mask and C is a displacing (noise) mask (Duncan and Pearson 1991).

Table 1 lists the perturbative methods and indicates whether it is suitable for continuous or categorical data.

Method	Continuous data	Categorical data
Additive noise	X	
Microaggregation	X	
Resampling	X	
Multiple imputation	X	
PRAM		X
Rank swapping	X	X
Rounding	X	

Table 1.1: Perturbative Methods Versus Data Types

1.1.2 Nonperturbative Methods

Nonperturbative methods do not rely on distortion of the original data but on partial suppressions or reductions of detail. Some of the methods are usable on both categorical and continuous data, but others are not suitable for continuous data.

Table 2 lists the nonperturbative methods and indicates whether it is suitable for continuous or categorical data.

Method	Continuous data	Categorical data
Sampling		X
Global recoding	X	X
Top and bottom coding	X	X
Local suppression		X

Table 1. 2: Nonperturbative Methods Versus Data Types

2. Information Loss Measures

To evaluate the information loss caused by an SDC method on a microdata set, we should assess how different the masked data set is from the original data set. The statement that there is little information loss is convenient if the structure of the masked data set is very similar to the structure of the original data set. In fact, the main purpose for preserving the structure of the data set is to ensure that the masked data set will be analytically valid and interesting.

Defining what a generic information loss measure is can be a tricky issue. In general, it should capture the amount of information loss for a reasonable range of data uses. There is a little information loss if the protected dataset is analytically valid and interesting according to the following definitions by Winkler (1999):

- A protected microdata set is *analytically valid* if it approximately preserves the following with respect to the original data:
    1. Means and covariances on a small set of sub domains
    2. Marginal values for a few tabulations of the data
    3. At least one distributional characteristic.
- A microdata set is *analytically interesting* if six attributes on important sub domains are provided that can be validly analysed.

More precise conditions for analytical validity and analytical interest cannot be stated without taking specific data uses into account. The above definitions suggest some possible measures:

- Compare raw records in the original and the protected dataset. The more similar the masking method to the identity function, the lesser the impact. This requires pairing records in the original dataset and records in the protected dataset. For masking methods based on the original data, each record in the protected dataset is naturally paired to the record in the original dataset it originates from. For synthetic microdata preserving only some features of the original data, pairing is more artificial.
- Compare some statistics computed on the original and the protected datasets. The above definitions list some statistics, which should be preserved as much as possible by a masking method.

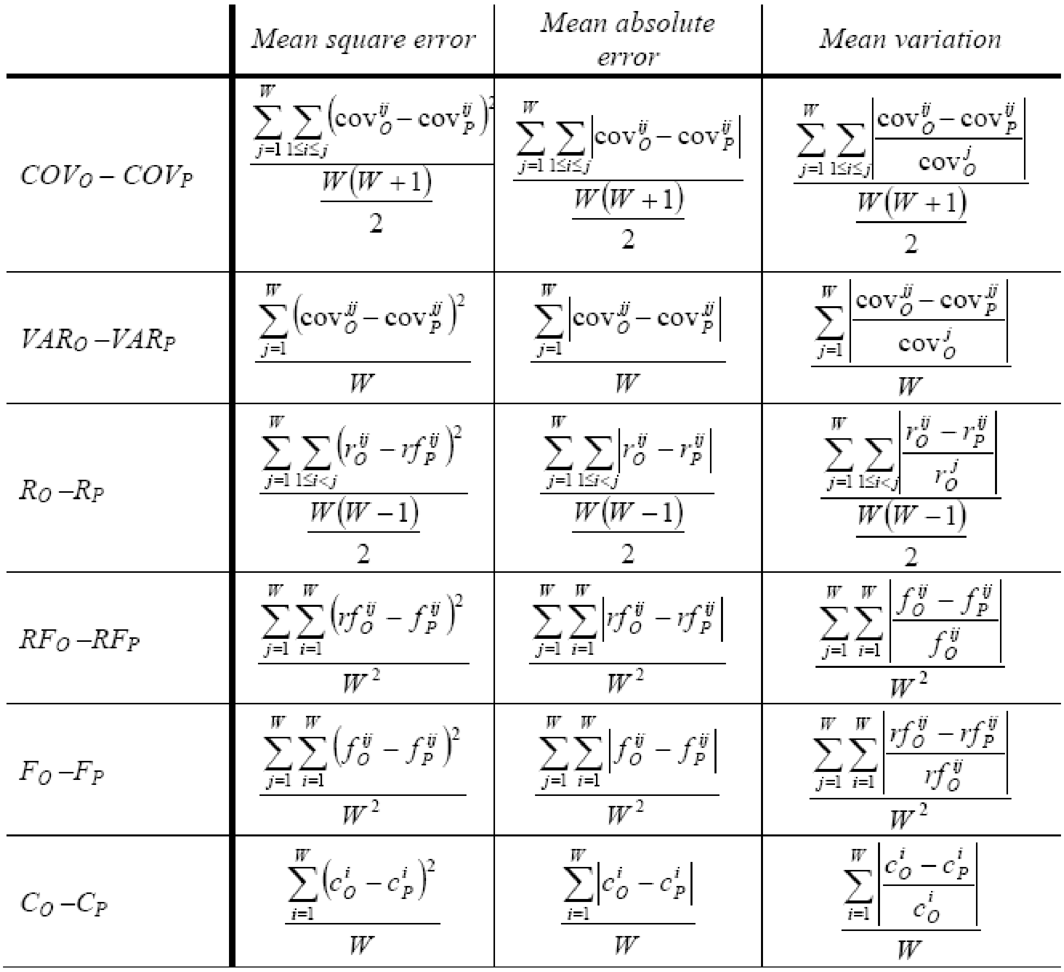
1.2 Information Loss Measures for Continuous Microdata

Domingo-Ferrer, Mateo-Sauz and Torra proposed several measures to quantify the information loss for continuous variables. The general idea is based on a concept developed by Winkler where a protected data set is analytically valid if the following features are approximately preserved:

- Means and covariances on a small set of sub domains
- Marginal values for a few tabulation of the data

- At least one distributional characteristic.

Therefore, if we need small differences between the statistics computed on the original and protected data we could assess the information loss as small. For continuous variables, we might compare the mean square error or mean absolute error or mean variation between covariance matrices, correlation matrices, principal component matrices or factor matrices of the two data (i.e. original and protected).



	Mean square error	*Mean absolute error*	*Mean variation*
$COV_O - COV_P$	$\frac{\sum_{j=1}^{W}\sum_{1\le i\le j}\left(\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}\right)^2}{\frac{W(W+1)}{2}}$	$\frac{\sum_{j=1}^{W}\sum_{1\le i\le j}\left\lvert\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}\right\rvert}{\frac{W(W+1)}{2}}$	$\frac{\sum_{j=1}^{W}\sum_{1\le i\le j}\left\lvert\frac{\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}}{\mathrm{cov}_O^{j}}\right\rvert}{\frac{W(W+1)}{2}}$
$VAR_O - VAR_P$	$\frac{\sum_{j=1}^{W}\left(\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}\right)^2}{W}$	$\frac{\sum_{j=1}^{W}\left\lvert\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}\right\rvert}{W}$	$\frac{\sum_{j=1}^{W}\left\lvert\frac{\mathrm{cov}_O^{ij}-\mathrm{cov}_P^{ij}}{\mathrm{cov}_O^{j}}\right\rvert}{W}$
$R_O - R_P$	$\frac{\sum_{j=1}^{W}\sum_{1\le i<j}\left(r_O^{ij}-rf_P^{ij}\right)^2}{\frac{W(W-1)}{2}}$	$\frac{\sum_{j=1}^{W}\sum_{1\le i<j}\left\lvert r_O^{ij}-r_P^{ij}\right\rvert}{\frac{W(W-1)}{2}}$	$\frac{\sum_{j=1}^{W}\sum_{1\le i<j}\left\lvert\frac{r_O^{ij}-r_P^{ij}}{r_O^{j}}\right\rvert}{\frac{W(W-1)}{2}}$
$RF_O - RF_P$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left(rf_O^{ij}-f_P^{ij}\right)^2}{W^2}$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left\lvert rf_O^{ij}-rf_P^{ij}\right\rvert}{W^2}$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left\lvert\frac{f_O^{ij}-f_P^{ij}}{f_O^{ij}}\right\rvert}{W^2}$
$F_O - F_P$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left(f_O^{ij}-f_P^{ij}\right)^2}{W^2}$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left\lvert f_O^{ij}-f_P^{ij}\right\rvert}{W^2}$	$\frac{\sum_{j=1}^{W}\sum_{i=1}^{W}\left\lvert\frac{rf_O^{ij}-rf_P^{ij}}{rf_O^{ij}}\right\rvert}{W^2}$
$C_O - C_P$	$\frac{\sum_{i=1}^{W}\left(c_O^{i}-c_P^{i}\right)^2}{W}$	$\frac{\sum_{i=1}^{W}\left\lvert c_O^{i}-c_P^{i}\right\rvert}{W}$	$\frac{\sum_{i=1}^{W}\left\lvert\frac{c_O^{i}-c_P^{i}}{c_O^{i}}\right\rvert}{W}$

Table 1. 3: Information Loss Measures for Continuous Microdata

Assume a set of microdata with *N* individuals (records) and *W* continuous variables. Let denote O the matrix representing the original set of microdata (rows are records and columns are variables) and P the matrix representing the perturbed set of microdata. Based on the two set of microdata we can compare the following statistics:

- Covariance matrices: COVo on O and COVp on P;
- Variance matrices: VARo on O and VARp on P;
- Correlation matrices: Ro on O and Rp on P,

- Correlation matrices RFo (respectively RFp) between the *W* variables and the W factors obtained through principal component analysis;
- Factor score coefficient matrices: Fo on O and Fp on P;
- Commonalities Co (respectively Cp) between each *W* variables and the first principal component.

Matrix discrepancy can be measured in at least three ways:

- Mean square error: sum of squared differences between pairs of matrices, divided by the number of cells in either matrix;
- Mean absolute error: sum of absolute differences between pairs of matrices, divided by the number of cells in either matrix;
- Mean variation: sum of absolute percentage variation of differences between pairs of matrices, divided by the number of cells in either matrix.

In practice, these calculations are done for selected continuous variables from masked data and they are used to compute the Global Information Loss for Continuous Variables (GILCV). Averaging the mean variations of covariance, variance, correlation, factor and commonalities and multiplying the resulting average by 100 compose GILCV:

$$GILCV = \left(\Delta_{COV}^{MeanVari} + \Delta_{VAR}^{MeanVari} + \Delta_{COR}^{MeanVar} + \Delta_{RF}^{MeanVari} + \Delta_{F}^{MeanVar} + \Delta_{C}^{MeanVari}\right)/\,6$$

This score is used as input for calculation of general information loss. A general score of information loss (GSIL) is calculated as:

$$GSIL = \frac{\overline{EBIL} + GHD + GILCV}{3}$$

where GHD is Global Relative Hellinger Distance, EBIL is Entropy-Based Information Loss. The score of GSIL is between 0% and 100%. A value of 0% indicates no information loss, whereas a value of 100% indicates total information loss or no similarity between the two sets of data. The interpretation of values between both scales is as follows: *small information loss* from 0% to 10%, *medium information loss* from 11% to 20%, *serious information loss* from 21% to 30% and *no data utility* for 31% and over.

1.3 Information Loss Measures for Categorical Microdata

For categorical data three kinds of information loss measures have been considered: direct comparison of categorical values, comparison of contingency tables and entropy-based measures.

1.3.1 Direct comparison of categorical variables

Comparison of matrices O and P for categorical data requires the definition of a distance for categorical variables. Definitions consider only the distances between pairs of categories that can appear when comparing a record and its masked version. When the range of a variable is an ordinal scale, the distance between category *a* and *b* is proportional to the number of categories between *a* and *b*. When the range of a variable is not ordinal, the distance is one if the values are different and zero if they are not.

1.3.2 Comparison of contingency tables

An alternative to directly comparing the values of categorical variables is to compare their contingency tables. For a given subset of variables (the original and the masked set), their corresponding t-dimensional contingency tables are computed for a file before and after applying the masking process. The number of differences between both contingency tables is denoted by CTBIL (Contingency Table Based Information Loss). As the number of cells in a contingency table depends on the number of categories in the variable, a normalised expression of CTBIL is considered. This corresponds to the former information loss measure but dividing the expression by the number of cells in all considered tables.

1.3.3 Entropy-based measures

In De Waal and Willenborg (1999) and Kooiman (1998), the use of Shannon's entropy to measure information loss is discussed for several methods of SDC (local suppression, global recoding and PRAM). Entropy is a theoretic measure, but it can be used in SDC if the masking process is modelled as the noise that would be added to the original dataset in the event of its being transmitted over a noisy channel. As this measure only depends on the masked data set and it does not account for its relation with the original data, a new information loss measure was defined. In this model V is a variable in the original data set and V' is corresponding variable in the masked data set (this set is mainly PRAM data set).

The entropy-based information loss measure EBIL is defined as:

$$EBIL(Pv,v'G)= \Sigma r \in G \; H(V'=jr)$$

Where *j* is the value taken by V' in record *r* and

$$H(V'= j)= -\sum p(V= i | V'=j)\log \; p(V=i | V'=j)$$

$p(V=i | V'=j)$ is Markov matrix.

The new information loss measure taking original data into account is:

$$IL \; (Pv | v',F,G)= \sum r \in g \; PRIL(Pv,v',in, jr)$$

Where *j* is the value taken by V in record *r* of F and *j* is as before, the value taken by V' in record *r* of G and

$$PRIL \; (Pv,v', i, j)= -\log P(V=i \,|\, V'=j).$$

3. Disclosure Control Methods - Practice in the Macedonian Statistical System

The Macedonian statistical system ~~of the Republic of Macedonia~~ is centralised, in which the main institution is the State Statistical Office. Other participants are: National Bank, Ministry of Finance, Ministry of Interior, Ministry of Justice, Pension and Disability Insurance Fund, Employment Agency, Institute of Public Health and Hydrometeorological Service.

In order to assess the practice of SDC inside the institutions, the SSO has conducted an ad hoc survey. The survey was conducted with the help of a short questionnaire comprising 9 questions on microdata access and usage of statistical disclosure methods for information loss measures. The text below provides a summary of the results of the survey for 7 institutions that completed the questionnaire.

PART A: Microdata Access

The survey showed that 5 of 7 institutions provide access to microdata for users. The access is ensured on the basis of signed memoranda of understanding between the data holder and the data user.

It should be mentioned that, except the SSO, all other institutions provide microdata only to legal units (not individual persons) and these data are mainly used as input for analysis.

A second question addressed the format for release of microdata and three options were given: anonymous microdata files (public use files), "Safe Room" in the premises of the institution and remote access facilities.

Only 3 institutions answered on this question stating that "Safe Room" is the format utilised for release of microdata.

Institutions were also asked to report on the statistical surveys for which microdata access is ensured and 4 of them answered with statement of relevant statistical surveys. The SSO is providing access to microdata from the following surveys: Labour Force Survey, Household Budget Survey and Survey on Income and Living Conditions.

All institutions that provide microdata for external users answered that the prerequisites needed to fulfil in order to be granted access to microdata are defined in a Memorandum of Cooperation signed between the data user and the data owner. The SSO provides access to microdata to researchers in accordance with good practices developed in the European Statistical System. Output results of the research project are checked and approved by authorised employees from the Office before being taken from the "Safe Room".

PART B: Statistical Disclosure Measures for Information Loss

The survey also tackled the issue of which statistical measures are used for calculation of information loss for "continuous data" and "categorical data", and how potential users are notified about the information loss in protected datasets.

All institutions reported that they have not developed procedures for calculation of disclosure risk on produced statistics, nor is information loss calculated.

No institution collects information about users who cancelled their request for microdata.

4. Conclusions from the ad hoc survey

The results of the survey show that in the next period more time and resources should be devoted to developing SDC practice in the Macedonian statistical system. In this sense, the State Statistical Office should update the present guidelines for SDC, including explanations for balancing usage of microdata with information loss measures, and inform other partners about the guidelines. One of the Office's tasks is coordination of the statistical system and this includes, among other things, promoting harmonised SDC practices between all institutions producing official statistics in the ~~Republic of Macedonia~~ national statistical system. Besides the other producers of statistics, the end users of statistics need to be informed to some level about the actions the SSO and other institutions take to preserve confidentiality in statistics. The reasons for practicing SDC should also be made clear and public. Statistical confidentiality must not be seen only as a factor limiting access to useful data.

References

Domingo-Ferrer, J., and Torra, V. (2001). *Disclosure Control Methods and Information loss for Microdata.* Amsterdam.

Domingo-Ferrer,J., Mateo-Sanz, J., and Torra, V. (2001). *Comparing SDC Methods for Microdata on the basis of Information loss and Disclosure Risk* . Luxembourg.

Trottini,M. (2003). *Decesion Models for Data Disclosure Limitation.* PhD thesis, Carrnegie Mellon University.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. &Roherig, S.F (2001). *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map.* Amsterdam.

Willenborg, L., and Dewaal,T., (2001). *Elements of Statistical Disclosure Control.* New York.