WP. 27
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Analysis of information loss in European data due to confidentiality

Prepared by Mihaela Agafitei and Daniel Defays (Eurostat)

Analysis of information loss in European data due to confidentiality

Mihaela Agafitei and Daniel Defays

Eurostat, European Commission, L-2920, Luxembourg,
e-mail: mihaela.agafitei@ec.europa.eu; daniel.defays@ec.europa.eu

Abstract. Eurostat and Statistical Agencies face increasing demands to release more and more detailed statistical data. This requires a high attention to respect the right to privacy of individuals and businesses while maximising the informative power of the European statistics released to both the general user and research community. A balance between the cost of implementing new modes of data access and the expected gains in information has to be kept. The paper will report on an attempt to quantify the informative power of released data compared to the full/complete datasets owned by the data provider. It provides a review of some information loss measures proposed in the literature and discuss how they may offer a global and suggestive assessment of the informative power of the European statistics.

Keywords: information loss measures, privacy, statistical disclosure control

1 Introduction

Statistical Disclosure Control (SDC) methods aim to ensure that statistical outputs provide as much values to the users while protecting the confidentiality of information concerning economic and social entities (e.g. businesses, individuals). If, in terms of protecting confidential data, things are clear being regulated by the European law, defining, measuring and maximising the value of the data disseminated put problems. Considering the fact that the range of users and of their interests is so diverse, it is very difficult to develop a general concept of data utility. On the contrary, we could try to quantify the information loss (IL) due to SDC methods by measuring the differences in joint distributions and assess whether conclusions based on data analysis of the original and perturbed data are or not significantly different.

This paper focuses on measuring the information loss for EU statistics based on a complex survey. We study the effect of SDC methods on the "European Union Statistics on Income and Living Conditions" (EU-SILC) data. The EU-SILC is the main source for the compilation o f comparable indicators on income distribution and social inclusion at European level. In particular, it provides the underlying data for the calculation of poverty indicators.

EU-SILC data are available on the Eurostat website and comprise multi-dimensional contingency tables and policy indicators. Access to anonymised microdata is possible for scientific purposes only, under specific conditions.

For this paper we used the 2008 cross-sectional data and we analyse the information loss by comparing different statistics based on original and perturbed/anonymised data. The selected variables for this exercise together with the definitions of their categories are presented in Table 1-1.

Table 1-1 Variables, categories and perturbation methods applied

Code	*Variable*	*Categories*	*Perturbation method*
CTR	Country	NUTS	None
REGION	Region	NUTS (2 digits)	Global recoding
URBAN	Degree of urbanisation	densely populated area	Top coding
		intermediate area	
		thinly populated area	
AGE	Age at the end of the income reference period		Top coding
GENDER	Gender	male	None
		female	
DWELLING	Dwelling type	detached house	Top coding
		semi-detached or terraced house	
		apartment or flat in a building with less than 10 dwellings	
		apartment or flat in a building with 10 or more dwellings	
		some other kind of accommodation	
NBROOM	Number of rooms		Top coding
CTRBIRTH	Country of birth		Global recoding
CITSHIP	Main citizenship		Global recoding
EDU	Highest education level attained	pre-primary education	Top coding
		primary education	
		lower secondary education	
		(upper) secondary education	
		post-secondary non tertiary education	
		first stage of tertiary education	
		second stage of tertiary education	
ACTIVITY	The economic activity of the local unit of the main job for respondents who are currently at work (NACE rev2)		Global recoding
ACTSTA	Status in employment	self-employed with employees	Derived indicator
		self-employed without employees	
		employee	
		family worker	
HHTYPE	Household	One person household	Derived

Code	*Variable*	*Categories*	*Perturbation method*
	type	2 adults, no dependent children, both adults under 65 years	indicator
		2 adults, no dependent children, at least one adult 65 years or more	
		Other households without dependent children	
		Single parent household, one or more dependent children	
		2 adults, one dependent child	
		2 adults, two dependent children	
		2 adults, three or more dependent children	
		Other households with dependent children	
		Other (these household are excluded from Laeken indicators calculation	
HHSIZE	Household size		Derived indicator
EQVINC	Equivalised disposable income		Rounding, micro-aggregation
AROP	At-risk-of-poverty	equivalised disposable income is greater or equal than 60% of median equivalised disposable income.	Derived indicator
		equivalised disposable income is less than the 60% of median equivalised disposable income	

2 Information loss measures

We propose to review three categories of information loss measures proposed in the literature and discuss how we can get a global measure of information loss related to a specific survey (EU-SILC) and a specific objective (see chapter 3 - Discussions). The first category is based on Shannon’s entropy which quantifies the expected value of the information (i.e. the original value) given specific data (i.e. the perturbed values). The second category uses the Hellinger distance as a tool for measuring differences in distributions of two datasets: original and protected data. The third category focuses on continuous variables and seeks to capture the discrepancies between correlations, covariances and factors obtained through principal component analysis.

2.1 Entropy-based information loss measure (EBIL)

In the information theory, the Shannon entropy is a measure of the uncertainty associated with a random variable. In other words, it quantifies the average information content that a receiver loses when not knowing the value of random variable. We apply the particular case of the conditional entropy where we quantify the remaining entropy/uncertainty of a variable in the original data given the value of the same variable in the perturbed data. Briefly, the smaller conditional probability for a variable in the original data given its values in the perturbed data, the larger is the information loss.

Let *V* be a discrete variable in the original data *O*, taking *K* categories and *V'* the corresponding variable in the protected data *P*, taking *L* categories. As the protected data is less detailed as the original data, we state that $L \le K$. According to the conditional entropy developed by Shannon, we can express the conditional uncertainty of *V* given *V'* as:

$$H(V/V^{'}=j) = -\sum_{i=1}^{K} p(V=i/V^{'}=j)\cdot \log p(V=i/V^{'}=j) \tag{1}$$

where

$$\sum_{i=1}^{K} p(V=i/V^{'}=j)=1 \;\; (\forall)j \in \{1,2,...,L\} \tag{2}$$

The total information loss based on Shannon's entropy (EBIL) is obtained by summing up the conditional uncertainties of all individuals *r* in the perturbed data *P*.

$$EBIL = \sum_{r\in P} H(V/V^{i}=j_r) \tag{3}$$

where j_r is the value taken by *V'* for the individual *r* in the perturbed data *P*.

This entropy function takes its largest value when all possible values of *V* have the same probability of being observed and the smallest when all the probability mass is concentrated on a single value:

$$0 \le EBIL \le N_P \cdot \log(K) \tag{5}$$

where N_P is the total number of individuals in the protected data *P* and *K* is the number of categories taken by the variable *V*.

2.2 Hellinger distance as an Information Loss measure (HDIL)

Distance metrics are used to measure distortion to distributions. Thus, we use the Hellinger distance to quantify the similarity between probability distributions of original and perturbed data. Let *X* be a variable and *o* its density function in the original data and *p* its density function in the perturbed data. The Hellinger distance is expressed as a standard calculus integral:

$$H(P,O) = \sqrt{\frac{1}{2}\cdot \int \left(\sqrt{o(x)} - \sqrt{p(x)}\right)^2 dx} \tag{6}$$

which satisfy the property $0 \le H(P,O) \le 1$ (7)

Since this measure is defined on continuous variables, we need to discrete it in order to apply on categorical variables. So, if we assume a countable space, the Hellinger distance between a variable *V* in original data and the corresponding variable *V'* in perturbed data is:

$$HD(V,V^{'}) = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{p(V=i)}-\sqrt{p(V^{'}=i)}\right)^{2}} = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Oi}}{N_O}}-\sqrt{\frac{n_{Pi}}{N_P}}\right)^{2}} \qquad (8)$$

where *K* is the total number of cells in the contingency table, n_{Oi} is the frequency of cell *i* in the original data *O*, n_{Pi} is the frequency of cell *i* in the perturbed data *P* and *N* is the total size of the countable space (i.e. the total number of individuals in the perturbed data).

The maximum distance 1 is achieved when *P* assigns probability zero to every set to which *O* assigns a positive probability, and vice versa.

The missing frequencies in the perturbed data are estimated by assuming an equal distribution of the perturbed category across the corresponding possible set of categories in the original data. For instance, we consider the variable "country of birth" which is perturbed by general recoding as follows: "*local*" (the country of birth and the country of residence are the same), "*born in EU*" (the country of birth is different from the country of residence but within European Union) and "*born outside EU*" (country of birth is outside the European Union). If we are looking at category "*born in EU*", we assume an equal distribution of "*born in EU*" people across all 26 remaining countries.

2.3 Information loss measure for continuous data

Domingo-Ferrer, Mateo-Sanz 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 featured 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 find small differences between the statistics computed on the original and protected data we could asses 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).

Table 2-1– Information loss measures for continuous variables

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

Assume a set of microdata with *N* individuals and *W* continuous variables. Let denote *O* the matrix representing the original set of microdata and *P* the matrix representing the perturbed set of microdata. Based on the two set of microdata we can compute the following statistics:

- covariance matrices: COV_O on *O* and COV_P on *P*;
- variance matrices: VAR_O on *O* and VAR_P on *P*;
- correlation matrices: R_O on *O* and R_P on *P*;
- correlation matrices RF_O (respectively RF_P) between the *W* variables and the W factors obtained through principal component analysis;
- factor score coefficient matrices: F_O on *O* and F_P on *P*;
- commonalities C_O (respectively C_P) 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;

3 Discussions

When deciding on data access modes and SDC methods, data disseminators needs (1) to ensure the protection of confidential information in accordance to the existing regulation and (2) to ensure a right balance between data utility and cost. The global difference in the distributions of the original and perturbed data plays a central role in assessing the utility of the disseminated statistical information.

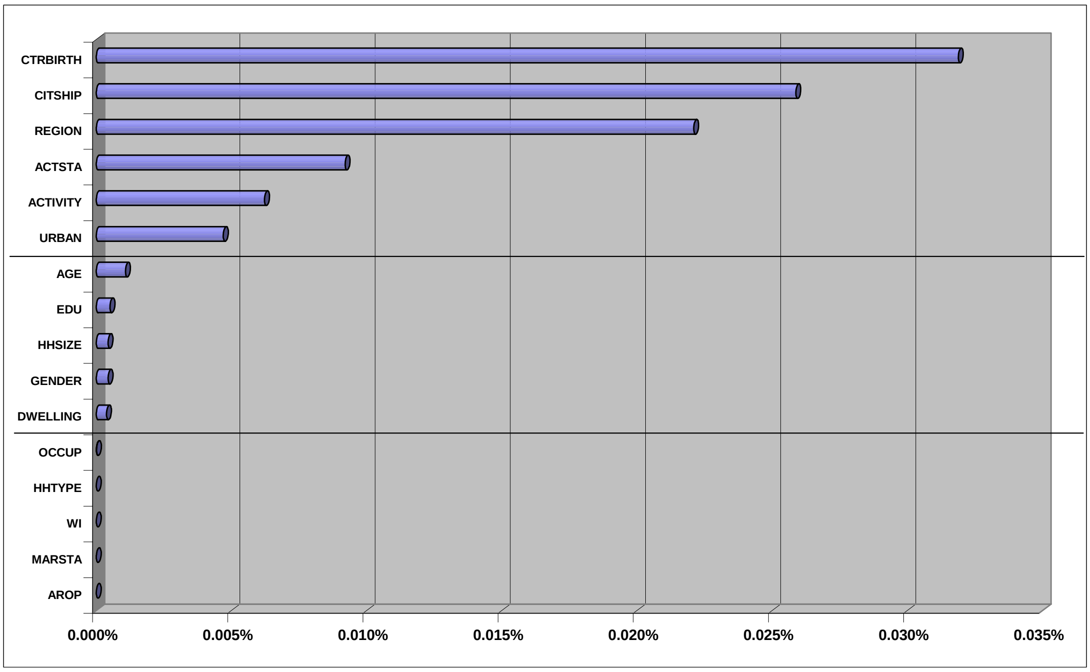
In this paper, we have presented three categories of information loss measures. The challenge is to develop a general measure of information loss over the data as a whole, to see whether the utility of the proposed release is “good enough" for the majority of users. In other words, the disseminator would like to quantify how much utility is lost when a particular pattern of data access mode and SDC method is used. So, the concept of “good enough” has no link to the quality/accuracy of the original data but to the bias/uncertainty introduced by SDC methods to the original data.

For this purpose, we suggest to combine the three categories of the information loss measures into a general index. To facilitate the interpretation of the discrepancy between original and perturbed data we use the relative values (percentage compared to the maximum possible information loss) for each category of IL measure.

Before presenting the findings, we need to define a specific objective for which we quantify the information loss: since the Lisbon European Council in 2000, the European Union has been committed to fight against poverty and social exclusion and developed for this purpose the Open Method of Coordination (OMC). A key element of the OMC is a set of indicators agreed upon jointly by the European Commission and all EU Member States, in order to measure progress towards the agreed EU social inclusion objectives. At the 2001 Laeken European Council, 18 indicators were adopted. The main measure of monetary poverty included in the EU list of indicators is a relative one (net income less than 60% national median), known as the “at-risk-of-poverty” rate. The “at-risk-of-poverty” rate is provided through EU-SILC survey. So, we could try to quantify the impact of SDC methods on the indicators of “at-risk-of-poverty”.

In Table 3-1 we can see that, according to EBIL measure, for “at-risk-of-poverty” rate at EU27 level, 2008 data, there is no information loss while some dimensions are affected but at a very small extent (the maximum relative EBIL by dimension is 0.03% for “country of birth”).

Figure 3-1 – The Entropy Based Information Loss (EBIL) expressed in relative value, EU27, 2008 EU-SILC data



The average relative EBIL across the variables when they are equally important is 0.006%.

We are now interested to check whether or not the distribution of population “at-risk-of-poverty” based on perturbed data is different form the distribution based on the original data. Table 3-2 and Table 3-3 show the relative HD for “at-risk-of-poverty” population by breakdowns used in disseminating the EU-SILC data, supplemented by other breakdowns considered by author as of interest for the general user. For one-dimensional breakdowns, the maximum information loss is about 32.08% for “economic activity”. As we combine different dimensions, the relative HD increases to a maximum of 34.21% for the four-dimensional breakdown by region, age, education level and status in employment.

We compute an average relative HD for each category of breakdown (i.e. $\overline{HD}_{1D}$ for one-dimensional, $\overline{HD}_{2D}$ for two-dimensional, $\overline{HD}_{3D}$ for three-dimensional and $\overline{HD}_{4D}$ for four-dimensional breakdown). Then, the global relative HD is computed as a weighted average as follows:

$$GHD = \frac{50 * \overline{HD}_{1D} + (25 * \overline{HD}_{2D} + 15 * \overline{HD}_{3D} + 10 * \overline{HD}_{4D})}{100} = 14.27\%$$

Figure 3-2 – Relative HD for "at-risk-of-poverty" population, EU27, 2008 data, by one-dimensional breakdowns

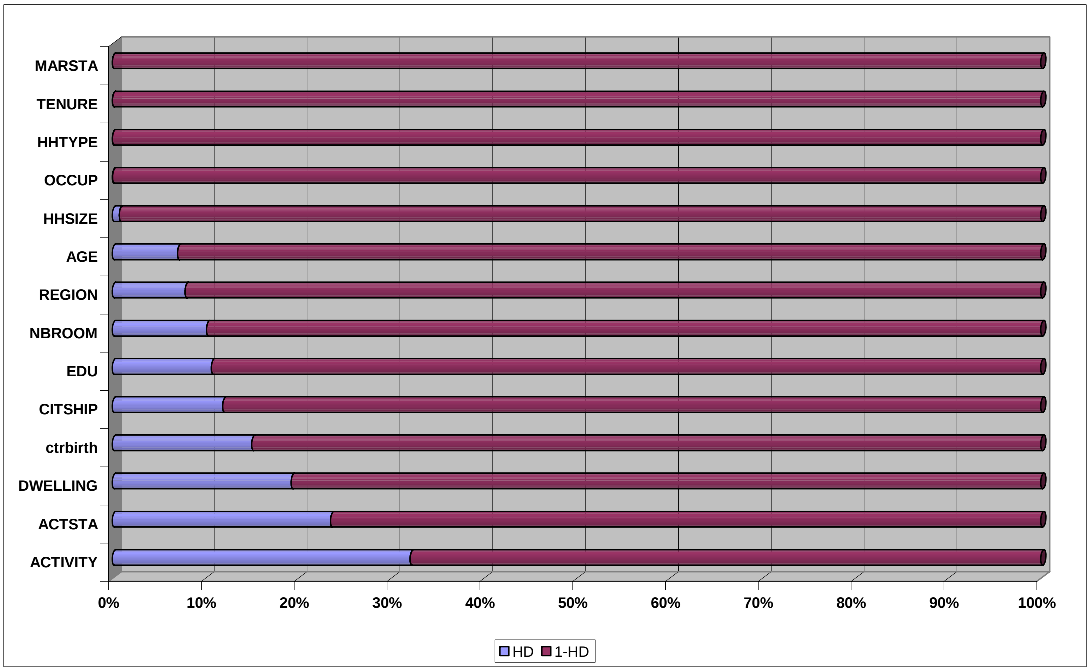
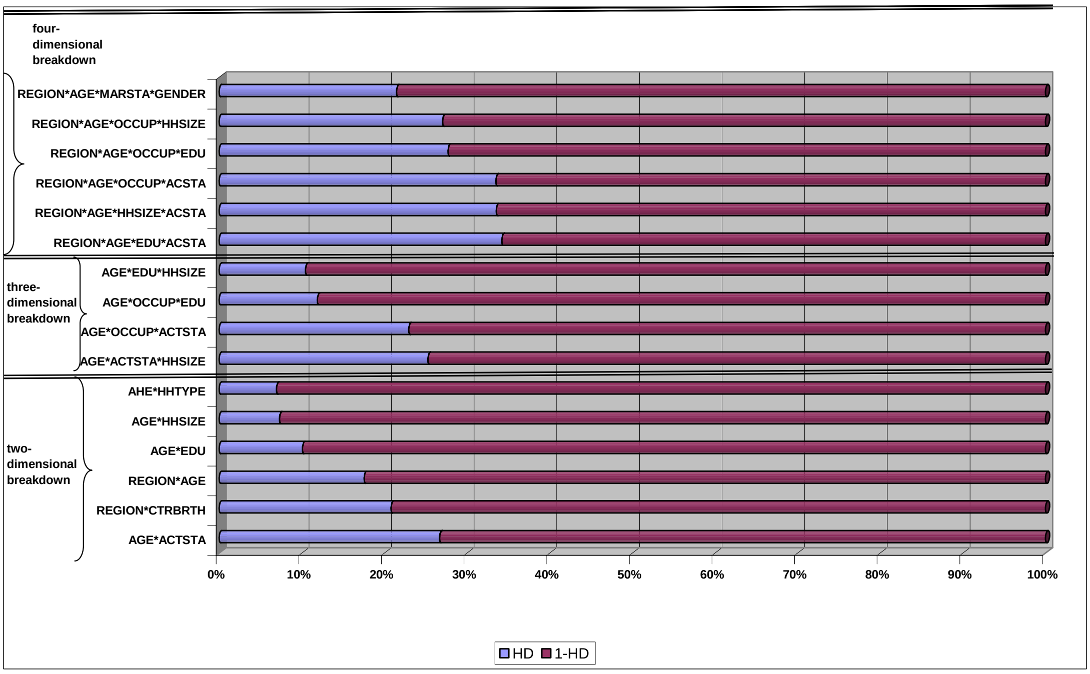


Figure 3-3 – Relative HD for "at-risk-of-poverty" population, EU27, 2008 data, by multi-dimensional breakdowns



The rationale of the above weighting is (1) to give the equal weight to one-dimensional breakdowns and to multi-dimensional breakdowns and (2) to give descending weight to multi-dimensional breakdowns.

If we are looking to the continuous variables, we can see the matrix discrepancy in Table 3-1. The selected continuous variables are: age, number of rooms, household size, total disposable income of the household and equivalised disposable income. The Global Information Loss for Continuous Variables (GILCV) is computed by averaging the mean variations of covariance, variance, correlation, factor and commonalities and multiplying the resulting average by 100.

$$GILCV = \left(\Delta_{COV}^{MeanVari} + \Delta_{VAR}^{MeanVari} + \Delta_{COR}^{MeanVar} + \Delta_{RF}^{MeanVari} + \Delta_{F}^{MeanVar} + \Delta_{C}^{MeanVari}\right)/\,6 = 11\%$$

Table 3-1 – Information loss measures for selected continuous variables, EU27 – 2008 EU-SILC data

Discrepancies in:	*Mean square error*	*Mean absolute error*	*Mean variation*
Covariance matrices	920.05	2.98	0.10
Variance matrices	2,756.13	8.47	0.08
Correlation matrices of variables	0.01	0.02	0.10
Correlation matrices between variables and their factors	0.00	0.02	0.24
Factor score coefficient matrices	0.01	0.02	0.09
Commonalities	0.00	0.01	0.02
Average	-	-	0.11

A general score of information loss (GSIL) is constructed as follows:

$$GSIL = \frac{\overline{EBIL} + GHD + GILCV}{3} = 8.43\%$$

The general score of information loss lies 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 set of data. It is more difficult to interpret value between the extremes. We proposed to use the next scale: *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.

This scale could be used either for the general score or for its components.

The score could be improved by introducing other IL measures as Shannon’s entropy for continuous variables, propensity score, association measure (Cramer’s V) and impact on variance of estimates.

References

[1] Bradshaw J. and Mayhew E., "The Measurement of Extreme Poverty In The European Union" , European Commission, Directorate-General for Employment, Social Affairs and Inclusion, 2011

[2] Domingo-Ferrer J., Mateo-Sanz J. M. and Torra V., "*Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk,*" in Proceedings of the ETK–NTTS 2001, Luxembourg: Eurostat, pp. 807–825.

[3] Domingo-Ferrer J. and Torra V.,"*A Quantitative Comparison of Disclosure Control Methods for Microdata,*" in Confidentiality, Disclosure and Data Access, 2001, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, Amsterdam: North–Holland, pp. 111–133.

[4] Duncan G. T., Fienberg S. E., Krishnan R., Padman R. and Roehrig S. F., "*Disclosure Limitation Methods and Information Loss for Tabular Data,*" in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, 2001, eds. P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. V. Zayatz, Amsterdam: Elsevier, pp. 135–166.

[5] Gomatam S. and Karr A. F., "*Distortion Measures for categorical Data Swapping*", National institute of Statistical sciences, 2003, Technical report number 131.

[6] Karr A.F., "*National Institute of Statistical sciences Data Confidentiality Technical Pa\per: Final report*", 2011, NCES 2011-608, U.S. Department of education, Washington, D.C., National Center for Education Statistics

[7] Shlomo N., "*Assessing the Impact of SDC Methods on Census Frequency Tables*", Work session on Statistical Data confidentiality, Manchester 17-19 December 2007, Eurostat

[8] Shlomo N. and Young, C., "*Information Loss Measures for Frequency Tables*", Joint UNECE/Eurostat work session on statistical data confidentiality, 2005, Geneva, Switzerland

[9] ***, "*Combating poverty and social exclusion - A statistical portrait of the European Union 2010* ", 2010 edition, Eurostat (ISSN 1830-7906),Luxembourg

[10] ***, "*Differences between data collected (as described in the guidelines) and anonymised user database*", Directorate F: Social Statistics and Information Society, Unit F-3: Living conditions and social protection statistics, 2010, Eurostat, Luxembourg

[11] ***, "*DESCRIPTION OF TARGET VARIABLES: Cross-sectional and Longitudinal 2008 operation*", Directorate F: Social Statistics and Information Society, Unit F-3: Living conditions and social protection statistics, 2010, Eurostat, Luxembourg