EUROPEAN COMMISSION
EUROSTAT

Directorate D: Single Market, Employment and Social statistics
Unit D-2: Living conditions and social protection

Anonymisation of EU-SILC USER DATABASE for researchers (July 2005)

1. EU-SILC INSTRUMENT

The European Instrument on Income and Living Conditions (EU-SILC) is gathering ex post output harmonised micro data collected in 25 MS and 2 EEA countries. It aims to provide comparable annual cross sectional data on income and living conditions and longitudinal data on income across Europe. The main operation started in 2004 for 10 MS and will reach the almost full regime (25 countries) in 2005. The data collection is based on the European Parliament and Council Regulation n°1177/2003 concerning Community statistics on income and living conditions. The instrument allows for flexibility and MS can collect data directly from a new survey or compile data from existing surveys and registers.

The EU-SILC micro data is a unique information source for studying poverty in its relation to socio-economic variables. It will be the primary source of data used by Eurostat for the calculation of many indicators in the field of Income, Poverty & Social Exclusion such as the Structural Indicators of Social Cohesion; indicators adopted under the Open Method of Coordination such as the 'Laeken' indicators of Social Inclusion and indicators of Pensions Adequacy; Sustainable Development Indicators of poverty and of ageing; and many other indicators published on the Eurostat New Cronos database. It is therefore a key tool for policy makers in particular, for monitoring Lisbon strategy. It will be indubitably of great interest for the research community in order to carry out detailed studies on poverty and living conditions.

2. STATISTICAL CONFIDENTIALITY AND RESEARCH RELEASE

The EU-SILC data are cleaned and imputed by the MS and then individual records will be transmitted to Eurostat without any direct identifiers (e.g. name, address, fiscal numbers). MS deliver a cross sectional dataset annually and a longitudinal dataset in which up to 4 years individual trajectories are compiled.

EU-SILC individual records are likely to be considered as confidential data in the sense of Article of Council Regulation 322/97 (Statistical Law) because they would allow indirect identification of statistical units (individuals or households). With this respect they should only be used for statistical purposes or for scientific research.

Commission Regulation 831/2002 granted the Commission to provide access to confidential data in the Eurostat premises and to release anonymised micro data for instance via CD-ROM to researchers.

Anonymised micro data are defined as individual statistical records which have been modified in order to minimise, in accordance to best practices, the risk of identification of the statistical units to which they relate.

Provision for the release of anonymised micro data to researchers is already made in the EU-SILC framework Regulation n°1177/2003.

The agreed procedure for granting access to EU-SILC is the following:

(1) The requests received are technically assessed by Eurostat (need for micro data and research interest) and legally (eligibility of the requesting body).

(2) MS are then formally consulted on the request and have six weeks to reply.

3. DISCLOSURE RISK ISSUES AND RELATED PROTECTION MEASURES

3.1. Fieldwork and sampling information

The release of sampling design information is potentially problematic because it may reveal geographical information or delineate subpopulations. In a first approach, we intend to remove the design information from the file. However, this information might be necessary for researchers who need to conduct proper population inference.

Eurostat will keep on thinking about a strategy to make available variance information without providing all the sampling information. If no solution exists, researchers wanting to compute designed based variance estimates would need to access original datasets in on site secure environments.

3.2. Geographical information

EU-SILC was not primarily designed for providing regional information. The NUTS 2 information as available in the original data sets might not be useful because sample might not have been designed to be representative at this geographical level. Moreover, as it has been pointed out as extremely identifying, no geographical information (NUTS code and degree of urbanisation) is considered for inclusion in the data base.

For some MS (most likely the large MS), the impact on disclosure risk of reintroducing some geographical information (NUTS1 and urban/rural classification) might be limited. Under the hypothesis that regional information is statistically relevant and taking into account that it could be of primary importance for researchers and policy makers to carry out regional studies, these MS should have the possibility to allow for the release of this

3.3. Global recoding other removed variables

The aim of global recoding and top coding of identifying variables is to reduce the number of unsafe records by reducing the level of information that can be used to identify them.

EUROSTAT considers that an appropriate choice of global recoding could achieve a significant decrease of the disclosure risk of the EU-SILC data base. In addition global recoding methods will be harmonised for all MS. The harmonisation of the anonymisation methods is crucial for usability and usefulness of the released database. The types of recode are based on a systematic examination of the distributions of the identifying variables and the identification of rare sample combinations in 3 ways combinations of variables. The different options were benchmarked against each other on the basis of the number of remaining unsafe records. The analysis carried out with the software Mu-Argus (4.0) leads to the following recodings as a significant step toward risk reduction:

Label	Code	Status
SEX	RB090	Unaltered
COUNTRY OF BIRTH	PB210	Local/EU/non EU/world
CITIZENSHIP 1	PB220A	Local/EU/non EU/world
CITIZENSHIP 2	PB220B	Removed
YEAR OF BIRTH or AGE	RB080	Bottom recode (1923 and before)
DWELLING TYPE	HH010	Modality 5 (“Other”) put to missing
TENURE STATUS	HH020	Unaltered
NUMBER OF ROOMS	HH030	Top coding (6 and more)
BATH OR SHOWER IN DWELLING	HH080	Unaltered
DO YOU HAVE A CAR?	HS110	Unaltered
MARITAL STATUS	PB190	Unaltered
CONSENSUAL UNION	PB200	Unaltered
EDUCATION (ISCED)	PE040	ISCED 5 and 6 regrouped
ECONOMIC STATUS	PL030	Unaltered
STATUS IN EMPLOYMENT	PL040	Unaltered
OCCUPATION (ISCO-88 (COM))	PL050	Unaltered
NACE	PL110	Grouped at 1 one letter (19 levels)
HOUSEHOLD TYPE	Derived	Unaltered
HOUSEHOLD SIZE	Derived	Unaltered

For EU-SILC, it is of primary importance not to hamper the scientific interest of the data base. For this reason, special attention has been put into keeping the year of birth/age at the current level of aggregation. However, month of birth (RB070) associated with year of birth (or age) has been considered as potentially highly identifying and therefore removed from the database

3.4. Local suppressions

It is expected that the global recoding and top/bottom coding that have been proposed so far will significantly decrease the re-identification risk associated with EU-SILC. If the number of records for which the risk measure is considered too high (the so called “unsafe” records) remains limited (less than a few percents), the datasets can be released to researchers under strict licence conditions as mentioned above. Alternatively, the unsafe records can be protected by carrying out local suppression or random perturbations of key variables.

Different patterns for local suppression exist. The suppression pattern can be controlled by the use of suppression weights which can help to penalise local suppressions for some variables. Ideally, suppression should concentrate on the least crucial variables for researchers and variables that will not affect the politically relevant estimates. Age, gender, activity status, household type and tenure status are particularly important in this respect.

In addition, local suppressions may alter the comparability between output of Official Statistics providers and results of research and policy evaluation. Local suppression and basic perturbation may thus hamper the interest of researchers in the data. Local suppression will also break out the calibration of the files released. Calibration is crucial to ensure consistency with other sources (demographic …). Eventually, the coherence of the local suppression pattern between the different releases of the datasets will be very cumbersome to check.

On the other hand, local suppressions might be embedded in the bulk of the "natural" missing values in the data files resulting from item non response. In some situations local suppressions may allow the release of more detailed information for several critical variables (e.g. geographical information). The right balance between the two aspects has to be obtained on a case by case basis.

Because of selectivity of the suppression, imputation of suppressed values seems not to bring an appropriate solution to this problem.

4. REGISTER COUNTRIES

In register countries, some EU-SILC variables (mainly some income components) could come directly from register, which under certain conditions can be public or accessible to researchers.

The most difficult situation is encountered in Norway, where a public file on individuals exists in Internet available for anyone. For all citizens included in the tax register, the file contains the following variables: name, address, postcode, net assets, income and tax. The variables are not identical with the variables used in Norwegian SILC, but they can be of use for the possible attacker.

For other countries, the situation is better because the access to register information is usually restricted and controlled. Although it is possible - at least for researchers - to match different registers with identifying variables to EU-SILC files, it takes knowledge of the data sources, resources and skills to attain these registers.

When EU-SILC variables can be obtained by an attacker from register sources, we would apply rounding techniques to EU-SILC variables. For instance, the base for rounding could be tuned to the data and vary along the measurement scale. If rounding did not offer sufficient protection micro –aggregation will also be considered.

5. ACTION PLAN FOR THE ANONYMISATION OF EU-SILC UDB FOR RESARCH RELEASE

1st step:

- The global recoding envisaged so far should be carried out uniformly for all national data sets (longitudinal and cross sectional)

- For large countries this should maintain the number of records for which the risk measure is too high to a few percents of the number of records;

- For small countries, further recoding should probably be envisaged, most likely for the variable Year of Birth;

- MS should have the possibility to propose limited number of additional coding/grouping (regrouping of rare modalities) adapted to their national specificities. However for the usability of the UDB, their number and their extent should remain limited and nested;

- The impact of keeping basic geographical information (region NUT1 and degree of urbanisation) in the EU-SILC data base has to be tested on the basis of disclosure risk The level of risk obtained with the introduction of regional information in large countries is likely to be of the order of magnitude observed for small countries without geographical breakdown;

- Depending of the shape of the distribution of the income variables, grouping/top coding of these variables should be envisaged in order to protect “outliers”.

2nd step:

- MS in cooperation with Eurostat should select a few disclosure scenarios and choose the corresponding levels of risk which are relevant in their national context.

- The safety of the file is then assessed on the basis of the number of unsafe records and the pattern of the local suppression for the real datasets

- If the number of unsafe records is limited (less than a few percents), the files can be released without further protection, given that the current level of contractual arrangement is maintained by Eurostat.

- If the number of local suppressions remains important, the possibilities of further coding and of local suppressions should be balanced. In any case, the local suppressions should target the less important variables and/or the variables for which the number of item missing is significant. The impact of local suppressions on the usefulness of the data and the lack of consistency between original dataset and protected data sets should be assessed.

The procedure and the practical conditions of the data release are intimate part of the risk management and reduction. They must be considered at the same time as the methods used for protecting data. They should be part of the agreement reached between MS and Eurostat.