WP. 20
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (iii) Confidentiality aspects of statistical information taking into account register-based data

EU SILC ANONYMISATION: RESULTS OF THE EUROSTAT TASK FORCE

Invited Paper

Submitted by Eurostat[1]

[1] Prepared by Jean-Marc Museux, jean-marc.museux@cec.eu.int.

EU SILC anonymisation: results of the Eurostat Task Force

Museux Jean-Marc*

* Living Conditions and Social Protection Statistics, Unit D3, Eurostat, European Commission, L-2920 Luxembourg, jean-marc.museux@cec.eu.int

Abstract: The European Instrument on Income and Living Conditions (EU-SILC) is gathering ex post output harmonised household and individual, cross sectional and longitudinal micro data income and living conditions on collected in 25 MS and 2 EEA countries. Data are a mix of register and survey data. Anonymised micro data will be released to researchers. In order to reflect on best practices, Eurostat convened a task force of experts in data protection and/or in the EU-SILC instrument. The paper presents the results of the Task Force. The methodological problems (panel/cross sectional, register/survey, individual/household) related to the anonymisation of a European database are reviewed, problems enhanced by the multiplicity of perceptions and realities of disclosure risk encountered in the different countries. The paper details both the methodological solutions proposed by the Task Force and the strategic and operational options retained to achieve a good trade-off between, on one side, information content and usability, and, on the other side, the monitoring disclosure risk.

1 Background

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.

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 first data set to be released from EU-SILC will contain 2004 cross sectional data and will be available in March 2006.

EU-SILC is the successor of the European Community Household Panel (ECHP) which was the main source of data for Income, Poverty & Social Exclusion for the reference years between 1994 and 2001. The ECHP anonymised user data base has been widely released to researchers. In this respect, this initiative was pioneering. Many contracts have been signed between Eurostat and research bodies ruling the access to ECHP micro data. The agreed procedure for granting access to ECHP was the following:

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

(2) MS are then formally consulted on the request and have six weeks to reply. For ECHP, only PT and BE are consulted a priori. Other MS are informed a posteriori through the annual report prepared for the Committee on Statistical Confidentiality

In order to come up with best practices and recommendations for anonymisation of EU-SILC user data base (UDB), Eurostat has convened a Task Force (TF) bringing together experts in the domain of anonymisation and experts in the SILC instrument (B. Bruno (Eurostat), L. Coppola (Istat), P. Feuvrier (INSEE), Ph. Gublin/J. Longhurst (ONS), N. Jukic (Stat Of. Slovenia), H. Minkel (Stat Bun), JM Museux (Eurostat), E. Schulte Nordholt (CBS), H. Sauli (Stat Fin)). The mandate states that the anonymisation of the EU-SILC data set both for release to researchers and to the public should be addressed. The public release should be conditioned to the quasi absence of risk of disclosure. In this matter, the role of the TF can be considered as a pilot for Eurostat and can serve as a basis to define more precisely the condition of public release of household micro data. This paper covers the work of the TF on the research release. Only headline of public release are reorted. The TF is currently keeping on elaborating on this issue. More account with be given during the oral presentation.

The objective of this paper is to give an account of the conclusions of the TF. It is the reference document to which Eurostat has refered to when making a proposal on the anonymisation and release of EU-SILC user data base to the Working Group on Living Conditions

2 Main orientations

The work of the TF initially drew upon three different approaches for assessing the disclosure risk when releasing micro data :

(3) the ONS approach based on population uniques and sample uniques (Elliot and Skinner) implemented in the SUDA software package,

(4) the CBS approach based on sample counts and implemented in the software package Mu-Argus.

(5) the ISTAT approach based on individual and household measure of disclosure risk (Franconi and Polettini, 2004), also implemented in Mu-Argus

The objective of the TF was not to reconcile the different methods but to benefit from each of them in order to issue recommendation for the anonymisation of the EU-SILC micro-data sets.

EU-SILC is a typical household survey. By nature key income variables are measured at household level. The household dimension carefully developed in EU-SILC makes it a primary source for household studies. It is characterised by the presence of hierarchies in the micro data file where the link between individuals and household is always present. From a disclosure point of view, this structure complicates the disclosure risk assessment since individuals can be disclosed through the identification of household characteristics and vice versa, households can be disclosed by characteristics of individuals. Given this hierarchical structure, it was necessary to consider the disclosure risks at both the individual and household level when assessing the disclosure risk of EU-SILC database.

One characteristic of EU-SILC instrument is the coexistence of a longitudinal dimension together with a cross sectional component. Although, the framework does not impose these to be linkable, many MS have used an integrated design (a rotational panel as recommended by Eurostat) and in this case, the components are not independent. Therefore it was necessary to ensure consistency between the anonymisation methods for the longitudinal and the cross-sectional data files.

The TF has pointed the specificities of so called register countries. For these countries, some of the income variables available in the EU-SILC may come directly from registers (DK, NO, SE, FI, SI, IS). If this register information is available together with direct identifiers is available to external users, the risk of disclosure is greatly increased. A specific section of this report is dedicated to it.

The recommendations proposed in this document draw upon the analysis of the disclosure risk in the EU-SILC 2003 data base for two countries. Analysis has been conducted for one small country, Luxembourg which gets the highest sampling fraction and in one medium size country, Greece, which gets a small sample fraction.

3 Identifying variables in EU-SILC UDB and intruder scenarios

For research release, the list of variables and the corresponding structure of the User Data Base (UDB) is likely to be very close to the structure of the data bases transmitted to Eurostat and described in Commission Regulation N°1983/2003. The data base is likely to be complemented with several derived variables (e.g. household size measured with the modified OECD equivalence scale, household activity status).

When releasing micro data files, statistical offices want to protect against standard disclosure risk scenarios by which an intruder, possessing a few variables (called identifying variables) about individuals in the population, is able to re-identify individual records from the micro data file thereby disclosing the content of other variables. These other variables can be classified as “sensitive” if they are perceived as (strictly) confidential.

The initial EU-SILC variables have been reviewed with respect to the identifying potential (the availability for intruders). They have been classified as Identifying, Sensitive or Others. Some variables have been classified as problematic regarding their specific nature: design weights and strata can lead to disclosure of detailed geographic information; precise timing variables such as month of birth or month on moved in or out are likely to created to fine classification which can lead to rare combinations; detailed fieldwork information, although not strictly disclosive, contains personal information. In addition, according to the CBS methodology, the identifying variables are grouped into **E**xtremely identifying, **V**ery identifying, (simply) **I**dentifying. This grouping refers to the specific methodology used by CBS and aims to introduce a hierarchy (in terms of availability) between the different variables. The subgroups are nested: one extremely identifying variable is considered automatically as very identifying and so on. This grouping exercise was based on the experience of the TF members.

Annex 1 gives the list of variables considered as Sensitive, Identifying and problematic present in the EU-SILC data base. Because, in some sense, it is difficult to unify national disclosure perception and the availability of variables, the subset of identifying variables can be seen as the union of national identifying variable sets.

The potentially identifying variables are then grouped into different "scenarios" which represent the information set the intruder has to hand to attack the database in different situations.

The TF has collected thirteen specific scenarios based on national experience. They can be found in annex 3 and are classified according to

- whether they relate to research release only or to both public and research release (by definition public scenarios apply to the release to researchers)

- whether they consider attack at household level or only at individual level

In order to assess the disclosure risk of research release of the EU-SILC database, 3 generic scenarios have been generated representing the common core of the different scenarios collected by the TF.

EU1 (Simple attack with HH information (individual and household level))

REGION x SEX x YEAR OF BIRTH x MARITAL STATUS x HH SIZE x HH TYPE

EU2 (Nosy neighbour individual attack– minimum scenario for nosy neighbour)

REGION x URBANISATION x SEX x YEAR OF BIRTH x BASIC ACTIVITY STATUS x BATH OR SHOWER x DO YOU HAVE A CAR? x EDUCATION x OCCUPATION x SECTOR OF ACTIVITY x HH SIZE x HH TYPE

EU3 (Occupational group address book individual attack)

REGION x URBANISATION x SEX x YEAR OF BIRTH x MONTH OF BIRTH x STATUS IN EMPLOYMENT x OCCUPATION x SECTOR OF ACTIVITY

These scenarios include quite a large number of variables. Their conjoint availability in attackers' data base depends on the national situation. The disclosure risk analysis relies on detection of rare combination in the population based on sample estimates

The CBS (NL) methodology for assessing the disclosure risk for research release does not use scenarios but considers all 3 way-combinations of one Extremely identifying

variable, one Very identifying variable and one Identifying variable focusing on rare (unique, two's, three's) combinations in the sample

4 Measure of risk and threshold

The three different approaches for assessing disclosure risk introduced in Section 2 can be distinguished by the measure of risk they rely on.

The ONS method is based on a measure of risk developed by Elliot and Skinner: the probability that a unique match of identifying variables with a sample unit is correct. This risk is estimated by computational intensive resampling methods implemented in the SUDA software. Special uniques in the sample (sample uniques which correspond to population uniques up to a high level of aggregation) are detected by the software. In addition to the individual and global measures of risk the results of the method also provides variables and value contribution to the risk Up to now, the ONS has only implemented this method for assessing the risk of microdata samples from the population censuses.

In the CBS approach, considering research release only rare combinations of identifying keys are considered as problematic. The following table gives the threshold for the number of replications of the given combination of the identifying keys above which the records are considered as potentially disclosive,e.g in the majority of MS a record is considered potentially disclosive if it is unique in the sample

Sampling fraction	Countries	Threshold
1/50 – 1/2	LU (f=2.5%)	5 (1+114 f)
1/100 – 1/50	MT, IS, CY	3
1/200 – 1/100	EE, SI	2
< 1/200	All other 21 MS	1

ISTAT approach provides a measure of risk for individual record based on the scenario where an attacker has a database with identifiers and key variables and tries to link it with the records in the sample. A match is given when the combination of the key variables is the same in both the sample record and the attacker's data base. The disclosure risk is defined as the probability that a match is correct (i.e. the sample record actually corresponds to the individual in the attacker's data base). This risk is estimated taking into account the sampling design, and is implemented in Mu Argus. Once the risk is estimated, a threshold has to be fixed to decide whether a record has to be considered safe or unsafe. Such a threshold is chosen depending of several aspects: (i) availability of data base providing identifiers and key variables; (ii) data base level of completeness and quality; (iii) comparability between the data base and the micro-data to be released (in terms of classification of the key variables); (iv) sampling fraction…etc.

Even though the three methods are based on different risk measures, it is expected that they produce datasets protected against the main disclosure risk. .

The ONS method has not been used for this analysis because the software was not available at Eurostat in the short term (license fee), the approach implemented did not take into account the sample design weights and the computational complexity was known to be rather high (one run could last for one week on a PC).

To assess the disclosure risk of EU-SILC, it is useful to compare to its predecessor: ECHP. ECHP anonymisation rules were based only on global recoding but no objective measure of risk was considered at that time. Ex post studies have however shown that the level of risk run when releasing ECHP anonymised user data base was quite important. However, up to now and despite a wide release of anonymised data files, there has been no striking problem of breach of confidentiality. The community of researchers has proved to be quite reliable and the actual level of risk when releasing micro data files to researchers under license restriction is likely to be very limited. This fact might be taken into account when a decision has to be taken on the release of anonymised EU_SILC micro data files.

5 Household and individual records

In EU-SILC data base, two levels of information (household and individual) will always coexist. Even if household identifiers are removed from individual files, the presence of household information (e.g. equivalised income) at individual level allows for household clustering.

The ISTAT approach allows for estimating individual risk of disclosure (named BIR in Mu-Argus) as well as household risk of disclosure (named BHR in Mu-Argus). To some extent, BHR implicitly takes into account the individual level of risk because it is based on the assumption that an household member is identifiable through the identification of another member of the same household. The ISTAT approach takes into account the household size through including it in the set of key variables. Moreover, an individual is considered safe only if BHR is lower than a chosen threshold, and at the same time BIR is lower than the same threshold divided by the household size. This implies an increasing level of protection, according to the household size. Such a strategy is sensible when the scenario of attack assumes that the intruder tries to match individual records, knowing which household each individual in the file to be released belongs to.

In other words, in order to integrate both individual and household levels, disclosure risk analysis is carried out on a file of individual data where identifying household variables are collected at individual level. Alternatively, a household file where data on individuals (such as age and sex of household members) are brought together at household level might be considered.

On the basis of the latter approach, UK studies on disclosure risk of household data have shown that most households of size 5+ are unique in the UK population and that they are a non-ignorable part of the population, The UK pattern has been reproduced for Slovenia and is likely to occur for all MS. These studies underline the fact that the disclosure risk of large households when using the ISTAT approach is likely to be underestimated. The household character of EU-SILC data are taken into account in scenario EU1 which take into account household and individual level attack and partly in scenario EU2 where household characteristics have been included at an individual level. The ISTAT model for household risk estimation, as implemented in Mu-Argus 4.0, has been used for assessing the risk.

It is proposed to control the actual disclosure risk linked to large households by carrying out uniform recoding as described in the section 7 and to only release household files under strict licence as was the case for ECHP.

Others solutions could have been:

- removal of households above a threshold and some uniform recoding

- perturbation with some uniform recoding

Given large households are crucial for the analysis of poverty their removal is not conceivable for research release.

Perturbation was not adopted here since the method requires a high level of competence both for designing a perturbed dataset and for analysing perturbed datasets. In addtion the method risks altering key indicators derived from the perturbed dataset.

6 Longitudinal data

One component of the EU-SILC instrument is longitudinal: individuals (and corresponding households) are traced for a minimum period of four years.

The release of cross sectional and longitudinal components to researchers is crucial.

Many countries are using an integrated design where the longitudinal and the cross sectional component may share the same individuals, as well as the same variables. Therefore, the matching between the files will be possible even with the absence of common identifiers. So there is a need to have a common strategy of anonymisation for both datasets so that one cannot be used to disclose the other (and vice versa).

The tracing of individuals over time increases the disclosure risk because transitions in identifying variables are likely to determine rare patterns. This type of risk has been highlighted for the different identifying variables and is detailed in the table of annex 4.

At the same time, this presupposes that the attacker is able to detect this change equally. For registers it is actually difficult to get access to data that exactly correspond to the reference period of EU-SILC data

In conclusion, it is very difficult to fully protect longitudinal data and keep relevant information for researchers at the same time. Following practice in MS, the TF recommends to consider the release of longitudinal files only under strict license.

7 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, it is recommended to remove the design information from the file. This issue if further addressed when discussion researchers needs in section 11.

8 Global recoding

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. The TF 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 can be harmonised for all MS and also are more easily implemented at a centralised level. The harmonisation of the anonymisation methods is crucial for usability and usefulness of the released database. The details of the recodes 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 choices made are benchmarked against each other using the number of remaining unsafe

records. On the basis of the analysis carried out with the software Mu-Argus (4.0) and presented in section 12, the TF proposes the following recoding as a first stage of risk reduction:

Label	Code	Global/top coding 1st step
REGION	DB040	not considered at the first step
DEGREE OF URBANISATION	DB100	not considered at the first step
SEX	RB090	None
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[1]	RB080	Bottom recode (1923 and before)
MONTH OF BIRTH		Removed
DWELLING TYPE	HH010	Modality 5 put to missing
TENURE STATUS	HH020	None
NUMBER OF ROOMS AVAILABLE TO THE HOUSEHOLD	HH030	Top coding (6 and more)
BATH OR SHOWER IN DWELLING	HH080	None
DO YOU HAVE A CAR?	HS110	None
MARITAL STATUS	PB190	None
CONSENSUAL UNION	PB200	None
EDUCATION (ISCED)	PE040	Isced 5 and 6 regrouped
ECONOMIC STATUS	PL030	None
STATUS IN EMPLOYMENT	PL040	None
OCCUPATION (ISCO-88 (COM))	PL050	None
NACE	PL110	Regrouped at 1 one letter (19 level)
HOUSEHOLD TYPE		Derived
HOUSEHOLD SIZE		Derived from hierarchical structure

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.

Geographical information

At this first stage no geographical information (NUTS code and degree of urbanisation) is considered for inclusion in the data base. Geographical information is, in the CBS approach, pointed out as extremely identifying. When geography is not considered, the CBS approach is reduced to the examination of the crossing of one very identifying and one identifying variable. In the others approaches, the geographical variables are simply ignored

[1] Providing only AGE instead of YEAR of BIRTH would provide an additional safeguard against disclosure because it depends on the date on which it is calculated which is not always available to the attackers

EU-SILC was not primarily designed for providing regional information. Moreover, 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. The same level of NUTS code encompasses different geographic realities depending of the country. In small countries, the first NUTS breakdown is confounded with the country itself. In some other countries NUTS classes are not homogeneous and are not relevant for statistical analysis. There is thus a danger to rely only on NUTS code to define geographic desegregation. The degree of urbanisation, on his side, is a complex concept that might not be readily available to an attacker.

For some MS (most likely the large MS), the impact of reintroducing some geographical information 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 information in the research files.

Comparison with ECHP

The level of global recoding proposed so far for EU-SILC can be compared to the situation for ECHP. Unlike the EU-SILC, the month of birth was released for the ECHP Otherwise (and not taking into account geographical information), the level of information is about the same in the two data bases. The current proposal for EU-SILC aims to bring more harmonisation between MS and hence increase the usability of the EU data base. Other minor differences are also observed: ISCED and number of rooms are further regrouped in EU-SILC while on the other hand Country of Birth, NACE and ISCO are further regrouped in ECHP. The level of geographical information in ECHP depends on the country. This is in line with the recommendation provided above. More details can be found in the table of annex 4.

9 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. The information given in table of annex 3 should also be taken into account.

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 implement.

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.

10 Register information

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 TF surveyed the situation of the register countries. 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, the TF recommends applying 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 could also be considered.

These aspects require additional studies which were beyond the scope of this TF. They have to be addressed and carefully monitored at national level.

11 Research community needs

To validate the a priori choices discussed above made by the TF, Eurostat has carried out a large consultation of researcher on the basis of the TF proposal. Comments have been collected in a free format and the following needs have immerged.

Geographic information is required not only because regional analysis is important but mainly for the coupling of macro information at the level of the region (employment rates …) in order to develop explanation model of individual behaviour. NUTS1 seems a minimum requirement. The degree of urbanisation also appears as an important explanatory variable

Age and date of birth are critical variables especially for analysis of life transitions (child care, education, work, retirement). The need to locate precisely in time the event recorded by the survey is crucial. For instance, it is required to define household equivalized scales that take change of composition during the reference period. Year coarsening is far from sufficient with this respect: the possibility to identify period with a precision of a quarter seems to be a minimum requirement. With this respect the withdrawal of move in/out information would have also important consequences. The

impact of top coding of age has also to be carefully assessed on the basis of the interest of developing studies on ageing people. 80+ to coding might prevent some interesting analysis on elderly.

Researchers pay a lot of importance on the quality of the inference they can draw from observed data. With this respect, not providing design weight will not allow to develop alternative weighting schemes. Worst, the masking of clustering effect (PSU, SSU) will not allow to develop correct (embedded multi level) modelling of observed behaviour.

Masking fieldwork information will not allow detailed analysis of respondent behaviour and quality checking of the analysis.

The prevalence of ISCO on NACE with respect to level of details is validated by researchers. It is underlined that the rough coarsening of country of birth and citizenship would prevent migration analysis and the inclusion of migration trajectories in human behaviour.

12 Results of experiments

The disclosure risk of EU-SILC databases has been studied using the data available at Eurostat for Luxembourg and Greece. Luxembourg is characterised by the highest sampling fraction (f=2.5 %) among all countries. Greece is characterised by a small sampling fraction (f= 0.1 %) and should be representative of medium size countries and regions in large countries.

The impact on disclosure risk of the global, top/bottom coding described in section 8 is studied using the different approaches and scenarios described in section 3 and 4. For the ISTAT approach different levels of risk and different levels of attack (household/individual) are considered. The focus is put on the number of "unsafe" records and the structure of the local suppressions proposed by the software package.

The CBS and the ISTAT approach have been tentatively compared by first, selecting the variables that concentrates the risk in the CBS approach and then measuring the ISTAT risk associated to the combination of variables ("CBS scenario – individual attack"). The results of the two approaches are roughly comparable for individual attacks.

For the EU2 complex scenario, the software is limited and does not allow the consideration all the variables simultaneously[2]. The scenario has thus been subdivided in overlapping sub scenarios so to capture the dependencies between the set initial variables.

For Greece, for attacks developed at the individual level, the number of suppressions is always less than 1 % of the number of records in the datasets.

The household attack always implied a higher number of suppressions. For scenario 2 with a level of risk of 0.04 the number of suppression is less than 2% of the number of records. The analysis of the structure of suppression among the different types of households have shown (not visible in the excel sheet) that the suppression preferentially affects the large households (1 over 2). Some large households are considered are not protected according to the method. This could be a characteristic of the structure of household population in Greece.

For Luxembourg, the results clearly demonstrate that the level of coding considered is not enough to declare the file as safe. The level of "unsafe" records can reach 10%

[2] In addition, the whole set of variables might not be available at the same time in all MS

depending on the approach and level of risk. The impact of recoding age by 5 years classes is briefly studied and seems to improve the situation. Further studies might be needed. The low performance of the coding for Luxembourg is due to the relatively high sampling fraction in that country. This corresponds to a more important disclosure risk typical of small regions/countries.

In conclusion, for large countries (sampling fraction lower than 0.01%), the global recodings described in section 8 are likely to significantly decrease the disclosure risk (measured in terms of local suppression) regardless of the approach (sampling fraction lower then 0.1 %). Further coding should be envisaged for small countries (LU (f=2.5 %) EE (f=:0.6%) CY (f=1.1%) MT (f= 1.8%) SI (f=:0.5%) and IS (f=1.3%)

13 EUROSTAT strategy for the design of Anonymised EU-SILC UDB for research release

In order to implement the here above recommendations, trying to find the right balanced between the need for harmonisation on one side and the need for some flexibility to adapt to MS sensitivity., Eurostat proposes the following strategy emphasising the prevalence of objective risk measure and good practices.

It can be split in two steps:

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 are not unlikely, most likely for the variable Year of Birth;
- At least NUTS1 and degree of urbanisation geographic coarsening should be introduced whenever the level of risk remain limited. Homogeneity of geographical grouping should prevail and more detailed breakdown could also be envisaged. It is likely that region of size equivalent to small countries can be made explicit in the files.
- 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;
- 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 level of risk which are relevant in their national context.
- The safety of the file is then assessed on the basis of the number 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.

In view of the researchers needs, Eurostat advocates to maintain the a minimum design information (anonymised PSU, anonymised SSU, order of selection and rotational group). However, neither strata nor design weights should be released. Fieldwork information such as contact information, proxy would be released in order to allow respondent behaviour studies which could help to improve the instrument.

In the longitudinal files, month of move in/out would be released because they are crucial longitudinal information and the increase of risk is inherent part of the risk of releasing longitudinal information.

The increase of risk induced by these decisions can be monitored by the contractual link and the close follow up mechanism Eurostat has put in place. 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

14 Public files

The TF has not considered the research release and the public release simultaneously. The TF contribution to public release is limited to a list a problematic issues to be addressed. The TF proposes to hold a 4th meeting dedicated to these issues. This meeting could come up with a practical proposal for a public use file for EU-SILC in line with current best practice. This practical proposal could help to assess the real need for such a file not taking into account the political and legal issues raised by such a proposal. These issues are far beyond the scope of this TF and have to be discussed at the appropriate level.

Issues concerned with the release of public EU-SILC files

- For the reasons explained in section 6, the longitudinal aspects of the EU-SILC survey are not compatible with public release. Only cross sectional files can be publicly released. Moreover, subsequent cross-sectional files can be easily matched because they include partly the same individuals. Hence, the annual release of cross-sectional files is also not compatible with the public release. The timeframe for the public release of cross sectional EU-SILC files must thus be carefully thought through.
- The hierarchical structure of the EU-SILC files allows, even without common keys, to reconstruct households and their detailed compositions. Following the UK studies mentioned in section 5, the identification risk of large households is very important and is thus incompatible with public release. The removal of large households information will definitively hamper the interest in the file. Alternatively, purely individual and/or household files can be released. In household files, the level of coding of the individual variables must be adapted to public release.
- The level of coding necessary to ensure the required safety of the public files might result in a level of information loss that causes the file to have no more

statistical interest. In many countries a public use file of the EU SILC data could only serve educational purposes but not advanced research requests. The interest might thus only rely on the visibility made to the survey and to the Official Statistical Office as public service provider.

- The public release of anonymised files with some sensitive information seems incompatible with situation where register information is publicly available
- The TF noted that the public release might be in contradiction with some National legislations.

References

Franconi, L. and Polettini, S. (2004). *Individual risk estimation in μ-Argus: a review*. In: Domingo-Ferrer, J. (Ed.), Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer, 262-272

Annex 1 : Classification of EU-SILC variable with respect to disclosure control

N°	X	L	name	Extremely Identifying	Very Identifying	Identifying	Problematic	Sensitive	Label
	HOUSEHOLD REGISTER (D-FILE)								
1	X	L	DB010						YEAR OF THE SURVEY
2	X	L	DB020						COUNTRY
3	X	L	DB030						HOUSEHOLD ID
4	X	L	DB040	**X**					REGION
5	X	L	DB050				**X**		PRIMARY STRATA
6	X	L	DB060				**X**		PSU-1 (FIRST STAGE)
7	X	L	DB062				**X**		PSU-2 (SECOND STAGE)
8	X	L	DB070				**X**		ORDER OF SELECTION OF PSU
9	X	L	DB075				**X**		ROTATIONAL GROUP
10	X	L	DB080				**X**		HOUSEHOLD DESIGN WEIGHT
11	X	L	DB090						HOUSEHOLD CROSS-SECTIONAL WEIGHT
12	X	L	DB100	**(X)**	**X**				DEGREE OF URBANISATION
13		L	DB110						HOUSEHOLD STATUS
14	X	L	DB120				**X**		CONTACT AT ADDRESS
15	X	L	DB130				**X**		HOUSEHOLD QUESTIONNAIRE RESULT
16	X	L	DB135				**X**		HOUSEHOLD INTERVIEW ACCEPTANCE

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
	PERSONAL REGISTER (R-FILE)								
1	X	L	RB010						YEAR OF THE SURVEY
2	X	L	RB020						COUNTRY
3	X	L	RB030						PERSONAL ID
4		L	RB040						CURRENT HOUSEHOLD ID
5	X		RB041						PERSONAL ID
6	X		RB050						PERSONAL CROSS-SECTIONAL WEIGHT
7		L	RB060						PERSONAL BASE WEIGHT
8	X	L	RB070				X		MONTH OF BIRTH
9	X	L	RB080			X			YEAR OF BIRTH
10	X	L	RB090		X				SEX
11		L	RB100				X		SAMPLE PERSON OR CO-RESIDENT
12		L	RB110						MEMBERSHIP STATUS
13		L	RB120						MOVED TO
14		L	RB140				X		MONTH MOVED OUT OR DIED
15		L	RB150						YEAR MOVED OUT OR DIED
16		L	RB160						NUMBER OF MONTHS IN HOUSEHOLD DURING THE INCOME REFERENCE PERIOD
17		L	RB170						MAIN ACTIVITY STATUS DURING THE INCOME REFERENCE PERIOD
18		L	RB180				X		MONTH MOVED IN
19		L	RB190						YEAR MOVED IN
20	X	L	RB200						RESIDENTIAL STATUS
21	X	L	RB210			X			BASIC ACTIVITY STATUS
22	X	L	RB220						FATHER ID

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
23	X	L	RB230						MOTHER ID
24	X	L	RB240						SPOUSE/PARTNER ID
25	X	L	RB245						RESPONDENT STATUS
26	X	L	RB250				X		DATA STATUS
27	X	L	RB260				X		TYPE OF INTERVIEW
28	X	L	RB270				X		PERSONAL ID OF PROXY
29	X		RL010-020			X			EDUCATION AT PRE-SCHOOL /COMPULSORY SCHOOL
31	X		RL030-060			X			CHILD CARE VARIABLES
35	X		RL070						CHILDREN CROSS-SECTIONAL WEIGHT FOR CHILD CARE

HOUSEHOLD DATA (H-FILE)

1	X	L	HB010						YEAR OF THE SURVEY
2	X	L	HB020						COUNTRY
3	X	L	HB030						HOUSEHOLD ID
4	X	L	HB040				X		DAY OF HOUSEHOLD INTERVIEW
5	X	L	HB050						MONTH OF HOUSEHOLD INTERVIEW
6	X	L	HB060						YEAR OF HOUSEHOLD INTERVIEW
7	X	L	HB070				X		PERSON RESPONDING THE HOUSEHOLD QUESTIONNAIRE
8	X	L	HB080				X		PERSON 1 RESPONSIBLE FOR THE ACCOMMODATION
9	X	L	HB090				X		PERSON 2 RESPONSIBLE FOR THE ACCOMMODATION
10	X	L	HB100				X		NUMBER OF MINUTES TO COMPLETE THE

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
									HOUSEHOLD QUESTIONNAIRE
11	X	L	HH010			**X**			DWELLING TYPE
12	X	L	HH020			**X**			TENURE STATUS
13	X	L	HH030			**X**		**X**	NUMBER OF ROOMS AVAILABLE TO THE HOUSEHOLD
14	X	L	HH031						YEAR OF CONTRACT OR PURCHASING OR INSTALLATION
15	X	L	HH040						LEAKING ROOF, DAMP WALLS/FLOORS/FOUNDATION, ….
16	X	L	HH050						ABILITY TO KEEP HOME ADEQUATELY WARM
17	X	L	HH060						CURRENT RENT RELATED TO OCCUPIED DWELLING
18	X	L	HH061						SUBJECTIVE RENT
19	X		HH070						TOTAL HOUSING COST
20	X	L	HH080			**X**			BATH OR SHOWER IN DWELLING
21	X	L	HH090						INDOOR FLUSHING TOILET FOR SOLE USE OF HOUSEHOLD
22	X	L	HS010						ARREARS ON MORTGAGE OR RENT PAYMENTS
23	X	L	HS020						ARREARS ON UTILITY BILLS
24	X	L	HS030						ARREARS ON HIRE PURCHASE INSTALMENTS OR OTHER LOAN PAYMENTS 176
25	X	L	HS040						CAPACITY TO AFFORD PAYING FOR ONE WEEK ANNUAL HOLIDAY AWAY FROM HOME
26	X	L	HS050						CAPACITY TO AFFORD A MEAL WITH MEAT, CHICKEN, FISH EVERY SECOND DAY
27	X	L	HS060						CAPACITY TO FACE UNEXPECTED FINANCIAL EXPENSES
28	X	L	HS070						DO YOU HAVE A TELEPHONE (INCLUDING MOBILE PHONE)?
29	X	L	HS080						DO YOU HAVE A COLOUR TV?

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
30	X	L	HS090						DO YOU HAVE A COMPUTER?
31	X	L	HS100						DO YOU HAVE A WASHING MACHINE?
32	X	L	HS110			**X**			DO YOU HAVE A CAR?
33	X	L	HS120						ABILITY TO MAKE ENDS MEET
34	X	L	HS130						LOWEST MONTHLY INCOME TO MAKE ENDS MEET
35	X	L	HS140						FINANCIAL BURDEN OF THE TOTAL HOUSING COST
36	X	L	HS150						FINANCIAL BURDEN OF THE REPAYMENT OF DEBTS FROM HIRE PURCHASES OR LOANS
37	X		HS160						PROBLEMS WITH THE DWELLING TOO DARK, NOT ENOUGH LIGHT
38	X		HS170						NOISE FROM NEIGHBOURS OR FROM THE STREET
39	X		HS180						POLLUTION, GRIME OR OTHER ENVIRONMENTAL PROBLEMS
40	X		HS190						CRIME VIOLENCE OR VANDALISM IN THE AREA
41	X	L	HY010					**X**	TOTAL HOUSEHOLD GROSS INCOME
42	X	L	HY020					**X**	TOTAL DISPOSABLE HOUSEHOLD INCOME
43	X	L	HY022					**X**	TOTAL DISPOSABLE HOUSEHOLD INCOME BEFORE SOCIAL TRANSFERS OTHER THAN OLDAGE AND SURVIVOR'S BENEFITS
44	X	L	HY023					**X**	TOTAL DISPOSABLE HOUSEHOLD INCOME BEFORE SOCIAL TRANSFERS INCLUDING OLDAGE AND SURVIVOR'S BENEFITS
45	X	L	HY025						WITHIN-HOUSEHOLD NON-RESPONSE INFLATION FACTOR
46	X	L	HY030G/HY030N					**X**	IMPUTED RENT
47	X	L	HY040G/HY040N					**X**	INCOME FROM RENTAL OF A PROPERTY OR LAND
48	X	L	HY090G/HY090N					**X**	INTEREST, DIVIDENDS, PROFIT FROM CAPITAL INVESTMENTS IN UNINCORPORATED BUSINESS
49	X	L	HY050G/HY050N					**X**	FAMILY/CHILDREN RELATED ALLOWANCES

N°	X	L	name	Extremely	Identifying Very	Identifying	Problematic	Sensitive	Label
50	X	L	HY060G/HY060N					**X**	SOCIAL EXCLUSION NOT ELSEWHERE CLASSIFIED
51	X	L	HY070G/HY070N					**X**	HOUSING ALLOWANCES
52	X	L	HY080G/HY080N					**X**	REGULAR INTER-HOUSEHOLD CASH TRANSFER RECEIVED
53	X	L	HY100G/HY100N					**X**	INTEREST REPAYMENTS ON MORTGAGE
54	X	L	HY110G/HY110N					**X**	INCOME RECEIVED BY PEOPLE AGED UNDER 16
55	X	L	HY120G/HY120N					**X**	REGULAR TAXES ON WEALTH
56	X	L	HY130G/HY130N					**X**	REGULAR INTER-HOUSEHOLD CASH TRANSFER PAID
57	X	L	HY140G/HY140N					**X**	TAX ON INCOME AND SOCIAL CONTRIBUTIONS
58	X	L	HY145N					**X**	REPAYMENTS/RECEIPTS FOR TAX ADJUSTMENT

PERSONAL DATA (P-FILE)

1	X	L	PB010						YEAR OF THE SURVEY
2	X	L	PB020						COUNTRY
3	X	L	PB030						PERSONAL ID
4	X		PB040						PERSONAL CROSS-SECTIONAL WEIGHT
5		L	PB050						PERSONAL BASE WEIGHT
6	X		PB060						PERSONAL CROSS-SECTIONAL WEIGHT FOR SELECTED RESPONDENT
7	X	L	PB070				**X**	**X**	PERSONAL DESIGN WEIGHT FOR SELECTED RESPONDENT
8		L	PB080						PERSONAL BASE WEIGHT FOR SELECTED RESPONDENT
9	X	L	PB090				**X**		DAY OF THE PERSONAL INTERVIEW
10	X	L	PB100						MONTH OF THE PERSONAL INTERVIEW

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
11	X	L	PB110						YEAR OF THE PERSONAL INTERVIEW
12	X	L	PB120				X		MINUTES TO COMPLETE THE PERSONAL QUESTIONNAIRE
13	X	L	PB130				X		MONTH OF BIRTH
14	X	L	PB140			X			YEAR OF BIRTH
15	X	L	PB150		X				SEX
16	X	L	PB160						FATHER ID
17	X	L	PB170						MOTHER ID
18	X	L	PB180						SPOUSE/PARTNER ID
19	X	L	PB190			X			MARITAL STATUS
20	X	L	PB200			X			CONSENSUAL UNION
21	X		PB210		X			X	COUNTRY OF BIRTH
22	X		PB220A		X			X	CITIZENSHIP 1
23	X		PB220B		X			X	CITIZENSHIP 2
24	X		PE010			X			CURRENT EDUCATION ACTIVITY
25	X		PE020			X		X	ISCED LEVEL CURRENTLY ATTENDED
26	X		PE030						YEAR WHEN HIGHEST LEVEL OF EDUCATION WAS ATTAINED
27	X	L	PE040			X		X	HIGHEST ISCED LEVEL ATTAINED
28	X	L	PH010					X	GENERAL HEALTH
29	X	L	PH020					X	SUFFER FROM ANY A CHRONIC (LONG-STANDING) ILLNESS OR CONDITION
30	X	L	PH030					X	LIMITATION IN ACTIVITIES BECAUSE OF HEALTH PROBLEMS
31	X		PH040					X	UNMET NEED FOR MEDICAL EXAMINATION OR TREATMENT
32	X		PH050					X	MAIN REASON FOR UNMET NEED FOR MEDICAL EXAMINATION OR TREATMENT

N°	X	L	name	Extremely	Very	Identifying	Problematic	Sensitive	Label
33	X		PH060					**X**	UNMET NEED FOR DENTAL EXAMINATION OR TREATMENT
34	X		PH070					**X**	MAIN REASON FOR UNMET NEED FOR DENTAL EXAMINATION OR TREATMENT
35	X		PL015					**X**	PERSON HAS EVER WORKED
36	X	L	PL020					**X**	ACTIVELY LOOKING FOR A JOB
37	X	L	PL025					**X**	AVAILABLE FOR WORK
38	X	L	PL030			**X**			SELF-DEFINED CURRENT ECONOMIC STATUS
39	X		PL035					**X**	WORKED AT LEAST 1 HOUR DURING THE PREVIOUS WEEK
40	X	L	PL040			**X**			STATUS IN EMPLOYMENT
41	X	L	PL050			**X**		**X**	OCCUPATION (ISCO-88 (COM))
42	X	L	PL060						NUMBER OF HOURS USUALLY WORKED PER WEEK IN MAIN JOB
43	X		PL070						NUMBER OF MONTHS SPENT AT FULL-TIME WORK
44	X		PL072						NUMBER OF MONTHS SPENT AT PART-TIME WORK
45	X		PL080						NUMBER OF MONTHS SPENT IN UNEMPLOYMENT
46	X		PL085						NUMBER OF MONTHS SPENT IN RETIREMENT
47	X		PL087						NUMBER OF MONTHS SPENT STUDYING
48	X		PL090						NUMBER OF MONTHS SPENT IN INACTIVITY
49	X		PL100						TOTAL NUMBER OF HOURS USUALLY WORKED IN SECOND, THIRD… JOBS
50	X		PL110			**X**		**X**	NACE
51	X		PL120						REASON FOR WORKING LESS THAN 30 HOURS
52	X		PL130					**X**	NUMBER OF PERSONS WORKING AT THE LOCAL UNIT
53	X	L	PL140						TYPE OF CONTRACT
54	X		PL150						MANAGERIAL POSITION

N°	X	L	name	Extremely Identifying	Very Identifying	Identifying	Problematic	Sensitive	Label
55	X	L	PL160						CHANGE OF JOB SINCE LAST YEAR
56	X	L	PL170						REASON FOR CHANGE
57	X	L	PL180						MOST RECENT CHANGE IN THE INDIVIDUAL'S ACTIVITY STATUS
58	X	L	PL190						WHEN BEGAN FIRST REGULAR JOB
59	X	L	PL200						NUMBER OF YEARS SPENT IN PAID WORK
60	X	L	PL210A-L						MAIN ACTIVITY ON JANUARY - DECEMBER
72	X	L	PY010G/PY010N					**X**	EMPLOYEE CASH OR NEAR CASH INCOME
73	X	L	PY020G/PY020N					**X**	NON-CASH EMPLOYEE INCOME
74	X	L	PY030G					**X**	EMPLOYER'S SOCIAL INSURANCE CONTRIBUTION
75	X	L	PY035G/PY035N					**X**	CONTRIBUTIONS TO INDIVIDUAL PRIVATE PENSION PLANS
76	X	L	PY050G/PY050N					**X**	CASH BENEFITS OR LOSSES FROM SELF-EMPLOYMENT
77	X	L	PY070G/PY070N					**X**	VALUE OF GOODS PRODUCED BY OWN-CONSUMPTION
78	X	L	PY080G/PY080N					**X**	PENSION FROM INDIVIDUAL PRIVATE PLANS
79	X	L	PY090G/PY090N					**X**	UNEMPLOYMENT BENEFITS
80	X	L	PY100G/PY100N					**X**	OLD-AGE BENEFITS
81	X	L	PY110G/PY110N					**X**	SURVIVOR' BENEFITS
82	X	L	PY120G/PY120N					**X**	SICKNESS BENEFITS
83	X	L	PY130G/PY130N					**X**	DISABILITY BENEFITS
84	X	L	PY140G/PY140N					**X**	EDUCATION-RELATED ALLOWANCES
85	X		PY200G					**X**	GROSS MONTHLY EARNINGS FOR EMPLOYEES

DERIVED VARIABLES

N°	X	L	name	Extremely Identifying	Very Identifying	Identifying	Problematic	Sensitive	Label
1	X		P			**X**			REFERENCE AGE (IN YEAR) AT THE INTERVIEW / AT THE END OF THE REFERENCE PERIOD
3	X		P			**X**			MOST FREQUENT ACTIVITY STATUS (EMPLOYED, UNEMPLOYED, RETIRED)
4	X		H			**X**			HOUSEHOLD TYPE (1 PERSON HH NO DEPENDENT CHILD, 2 P HH NO DCH, 1 P HH WITH DCH, 2 P HH WITH 1DCH, 2 P HH WITH 2 DCH, 2 P HH WITH >=3 DCH, OTHER HH WITH DCH)
5	X		H						WORK INTENSITY OF THE HOUSEHOLD
6	X		H			**X**			HOUSEHOLD SIZE
8	X		H					**X**	EQUIVALISED DISPOSABLE INCOME
9	X		H					**X**	TOTAL HOUSEHOLD INCOME BY COMPONENT (EMPLOYEE, SELF INCOME, SOCIAL BENEFIT, …)