Public Use Files of EU-SILC and EU-LFS data

Peter-Paul de Wolf*

* Process development, IT and methodology, Statistics Netherlands, The Hague, The Netherlands
pp.dewolf@cbs.nl

Abstract: Partly financed by Eurostat, seven member states[1] have been working on a proposal for a harmonized approach to produce public use files of the EU-SILC and EU-LFS data. The final results are due in December 2015. The current paper will describe the state of affairs and will discuss the (preliminary) results.

1 Introduction

Eurostat provides access to several microdata sets to external researchers. Some of the data is available as Secure Use Files (ScUF, on site), other as Scientific Use Files (SUF, may leave Eurostats premises). Both the secure use files and the scientific use files are made available to accredited researchers only: the data are confidential and hence the access is restricted. Since only accredited researchers who are additionally bounded by legal restrictions can access those files (with penalties when breaching those restrictions), the scientific use files are not fully anonymised. However, the possibility of disclosing sensitive information on individual respondents in scientific use files is reduced by applying some statistical disclosure control techniques, mainly global recoding and local suppression.

To apply for access to scientific use files, the organization the researcher belongs to first needs to be recognized as a genuine research entity. Then, for each project they want to start, a research proposal needs to be submitted and accepted by Eurostat. Since this accreditation process may take quite some time (sometimes up to 10 weeks), it would be beneficial to the researcher and his institute to know more about the content of the dataset in advance. That way the researcher could make a more profound decision whether it will be worthwhile to invest time in the accreditation process.

To aid the researcher in this respect, it would be helpful to have a microdata file available that could be send to the researcher in advance of the accreditation process. Since this would be a file without any control over the use of it, this should be considered to be a Public Use File (PUF) and hence treated as such. Moreover, this PUF should reflect the structure of the associated scientific use file as much as possible. That way the researcher could test scripts in advance and get a general idea about the content of the file.

[1] This paper is based on the contributions of all members of this team: Maxime Bergeat, Matthias Templ, Lydia Spies, Annu Cabrera, Péter Kristóf, Andreja Smukavec, Aleksandra Bujnowska, Peter-Paul de Wolf

In January 2015, Eurostat launched a project[2] to produce PUFs for data already available at Eurostat as scientific use files (SUFs), to facilitate researchers awaiting their accreditation. As an additional intended use of those PUFs, it was stated that these files should to be accessible in statistical trainings. At the end of this project, potential PUFs should be available along with a description how to produce those PUFs in a harmonized way. The data used for this project, was suggested by Eurostat and chosen based on the popularity of the SUFs.

The number of times that a request was made to Eurostat for access to specific microdata[3] is given in Table 1. It shows that the EU-SILC and EU-LFS data are the most frequently required datasets. Thus, in the project the EU-SILC and the EU-LFS data are used.

	ECHP	**LFS**	**SILC**	**AES**	**CIS**	**SES**	**EHIS**	**ERFT**	**CVTS**	**CSIS**
2013	7	38	45	6	13	7	3	1		
2014	41	134	164	21	31	32	13	1	9	2

Table 1: Number of requests to Eurostat for microdata access in 2013/2014.

For more detailed information on the procedures for access to EU microdata, we refer to the Eurostat paper presented at this work session (Bujnowska, 2015).

To construct the PUFs, essentially two approaches were considered: a 'traditional' approach and a (fully) synthetic data approach. In the traditional approach, only SDC methods like global recoding, local suppression and a simple form of PRAM[4] will be used. The synthetic data approach would yield the construction of a data generating process.

The EU-SILC dataset contains several income variables. Some countries do not allow for sensitive variables to appear in public use files. However, removing all income information from the EU-SILC dataset would result in a PUF that can not serve its intended purpose described above. For that reason it was decided to apply a fully synthetic data generation for the EU-SILC PUF. Moreover, some research in this direction had already been done before by one of the participating member states (see Alfons et al., 2011). For the EU-LFS dataset it was decided to use the traditional approach.

In this paper we will describe the processes to construct the PUFs. We will start with the description of the synthetic data approach for the EU-SILC in section 2. In section 3 we will describe the traditional approach for the EU-LFS data. At the time of writing this paper, the approaches have only been applied to a limited number of datasets. In section 4 we will discuss the aspect of disclosure risk related to the different approaches. To see how well the resulting public use files can be used as a preliminary dataset by researchers

[2]SGA 11112.2014.067-2014.765 under FPA 11112.2014.005-2014.533

[3]Source: Eurostat presentation at the second European Data Access Forum, 24-25 March 2015

[4]See e.g., Hundepool et al. (2012) and Gouweleeuw et al. (1998) for a description of PRAM

that await the results of the accreditation process, we defined some utility measures. In section 5 we will discuss those utility measures.

2 The approach taken for the EU-SILC data

EU-SILC (the EU Statistics on Income and Living Conditions) is a cross-sectional and longitudinal sample survey, coordinated by Eurostat, based on data from the EU member states, some EFTA countries and some EU-candidate countries. EU-SILC provides data on income, poverty, social exclusion and living conditions in the European Union. There are two data scopes:

- Cross-sectional data pertaining to fixed time periods, with variables on income, poverty, social exclusion and living conditions, and
- Longitudinal data pertaining to individual-level changes over time, usually observed over four years.

To our knowledge, no synthetic data generation method is readily available to produce a fully synthetic longitudinal dataset, consistent with the related synthesized cross-sectional datasets and still close to the SUF. We therefore decided in the limited time frame of this project to concentrate on the production of a PUF of the cross-sectional part of the EU-SILC data only.

2.1 The synthetic approach

For a detailed description of the used methodology we refer to Alfons et al. (2011). For a detailed description of the application to the EU-SILC dataset specifically, we refer to one of the deliverables of the project (due December 2015). In this section we will try to summarize the main issues of the approach.

The main goal for this dataset is to produce a synthetic dataset that is 'close' to the SUF provided by Eurostat. Generally speaking, constructing a *fully* synthetic dataset (all variables will be 'fake') will result in a 'safe' PUF. However, we will discuss the disclosure risk issue in more detail in section 4.

To be able to construct a PUF that resembles the structure of the SUF as well as possible and at the same time contains information as close as possible to the information in the original sample, the models to produce the synthetic data will be based on the raw data. That is, we use the dataset that is send to Eurostat before it is transformed in a SUF by Eurostat. This will result in a synthetic dataset that resembles the structure of the raw data. Hence, we will also need to apply the transformation used by Eurostat to that synthetic dataset, to obtain the structure of the SUF.

The EU-SILC dataset is essentially a household survey. This means that some household structure is embedded into the dataset. Using a synthetic data approach, we will keep the household structure in our PUF.

The general idea is to first generate a synthetic population, using models that are estimated with the raw data. Then from that synthetic population a sample is drawn of the same size as the raw data. Finally that is transformed to reflect the structure of the SUF. To be able to simulate a full population we will use the cross-sectional weights that are included in the raw data.

The synthetic data simulation framework consists of four steps, applied to each (regional) stratum independently:

1. Setup of the household structure
2. Simulation of categorical variables
3. Simulation of (semi) continuous variables
4. Splitting (semi) continuous variables into various components

The first step, setting up the household structure, comprises to the following idea. First the Horvitz-Thompson estimator is used to estimate the number of households of each household size in the population. Then the synthetic population is constructed by producing exactly that number of households. For each household of size l in that synthetic population, the household structure (age and sex distribution within the household) is drawn from the structures of households of size l apparent in the raw data, i.e., by re-sampling. This is done to prevent the construction of illogical household structures. However, this also means that whenever a certain household size is sample-unique in its stratum, its structure (age sex distribution) will always be reproduced for each simulated household of that size within that stratum. On the other hand, it will be possible that multiple households of that size will be simulated within that stratum.

In the next step, the categorical variables will be simulated for each household in the synthetic population. This is performed in a sequential way (each variable will be simulated, conditional on the previously simulated variables). Using the raw data, a multinomial logistic regression model is fitted for each categorical variable, with the previously simulated variables as predictors. Then a score is drawn from the multinomial distribution with the estimated (conditional) probabilities. The variables that are simulated in this way are (in this order): self defined economic status, citizenship, marital status, education, occupation (1 digit, the second digit is drawn randomly conditional on the first digit), NACE (1 digit).

In the third step the (semi) continuous variables are simulated. A continuous variable is simulated in two stages. First the variable is mapped to a categorization of the variable (e.g., income classes). Then the same approach is used as in the case of 'real' categorical variables as just explained. Finally, a random value within the income class is drawn to get a continuous score.

In case of an income variable, the total income needs to be split into different components. To this end, the distribution over the different components is done by donor

imputation where one record of the same stratum is taken for each simulated record and the simulated record gets all proportions of this real donor record. This is done independently for household and personal income components. Note that only the *proportions* of the income variable of a donor record are used.

Finally, the PUF is drawn from the synthetic population. This is done using stratified random sampling with replacement, where the region is the stratum variable and the household is the sample unit. I.e., per region the same number of households as in the SUF is drawn (but possibly with different household sizes) from the synthetic population, with replacement.

2.2 Some practical issues

When applying the synthetic data generation process as just described, we came across some practical issues. It turned out that some variables were rather sparsely distributed within some strata. This affects the model estimation. In some cases, we then decided to estimate the model for the whole country at once.

Another practical issue concerns the size of the population. The general idea of generating the full population is not feasible for all countries (depending on their actual population size and the available computing power). In that case, it suffices to generate a synthetic population that is smaller than the true population, but substantially larger than the SILC sample size. Obviously, one has to take this into account when dealing with the weights.

Due to the fact that there are many variables in the EU-SILC dataset, it turns out to be infeasible to simulate all variables as described above: conditioning on many (already simulated) variables increases the computation time too much. Therefore we divided the variables into two groups: one group of variables for which we condition on all previously simulated variables within that group, and another group of variables for which we always condition on a single other variable (income in 5 classes). For the latter group the distribution to simulate from is estimated on the basis of the (weighted) distribution of that variable in the SUF (conditional on the income classes). This is applied *after* the income is simulated. Hence, the simulated income is used to get the synthetic value of this variable.

For the application of this approach, some R scripts were developed. These R scripts entail some preprocessing of the data, the construction of the synthetic population, the sampling and the reformatting of the sample to concur with the format of the SUF. For the simulation of the synthetic population, the R package simPop is used.

3 The approach taken for the EU-LFS data

EU-LFS (the EU Labour Force Survey) is a cross-sectional and longitudinal household sample survey, coordinated by Eurostat, based on data from the EU member states, some EFTA countries and some EU-candidate countries. The database comprises observations

on labour market participation and persons outside the labour force. The data can be divided into quarterly datasets and yearly datasets. We decided to start with the quarterly data and then construct a yearly PUF from the four quarterly PUFs. However, some variables that appear in the yearly dataset are not available in the quarterly datasets. Hence, these variables needed to be treated a bit differently.

3.1 The traditional approach

Basically we applied three methods:

1. Removing variables (globally set to 'missing')
2. Global recoding
3. (a) Local suppression based on k-anonymity on a specific subset of all identifying variables with PRAM on the remaining ones
   (b) Local suppression based on all m-dimensional combinations of all identifying variables

The first method (removing variables) will be implemented in such a way that the structure of the resulting PUF will not differ from that of the corresponding SUF. I.e., when we mention that a variable will be removed, we mean that in all records the score on that variable will be set to 'missing'.

First the variables that could be used to (re)construct households were removed. This appeared to be obligatory for PUFs in certain member states and it turned out that the main interest of researchers was not at the level of the households but at the level of the personal information. The regional variable was globally recoded to country level, hence essentially removed. Some other variables were globally set to 'missing' (i.e., also essentially removed), because they are related in a complicated way to other variables that are going to be recoded. E.g. NACE according to an older classification.

After removing some variables, we found 12 identifying variables: Degree of urbanisation, Sex, Age, Nationality, Occupation code, Years of residence, Highest level of education, Country of birth, Nace, Professional status, Country of work and Working status. See Eurostat (2014) for more information on these variables as given in the SUF.

The global recoding was applied to several variables. The recoding is a further coarsening of the coding in the SUF.

- age into 6 classes: 0-14, 15-24, 25-39, 40-54, 55-74 and 75+
- nationality into 3 classes: Native, EU28 and NoAnswer/Other
- country of birth into 3 classes: Native, EU28 and NoAnswer/Other
- occupation into 1-digit ISCO code

- years of residence in member state into 3 classes: 0, 1-9 and 10+
- level of education into 3 classes: Low, Middle, High
- professional status: employee and family worker are grouped together
- country of work into 3 classes: Own country, EU28 and NoAnswer/Other
- degree of urbanisation: densely and intermediate are grouped together
- NACE (Rev. 2) into 7 classes: A, B-E, F, G, H-S, U and T

The initial idea was that after the global recoding has been done, the risk measure had to be checked and the remaining problems need to be solved using local suppression. However, it turned out that essentially two approaches are commonly used: k-anonymity on a limited set of key variables and checking all m-dimensional combinations (let's call this the all-m approach). In our experiments with k-anonymity we used $k = 5$ with the key variables Degree of urbanisation, Sex, Age, Nationality, Occupation code, Years of residence and Highest level of education. To the other identifying variables PRAM will be applied, with probability of not changing equal to 80% and the remaining probability equally distributed over the other categories. For the all-m approach we used $m = 4$ and a threshold of 10 in all dimensions.

3.2 Some practical issues

Some variables might give information about other variables, so they have to be treated very carefully. Examples are:

- Nationality: when nationality is suppressed, the years of residence might still give information whether the person is a foreigner or not
- Occupation: if Working Status = Employed, the Labour status during the reference week must be either 'did work during reference week' or 'did not work because was absent from his job during reference week'
- Highest level of education: if both Working status and Labour status during reference week indicate that it is about a person less than 15 years old, Highest level of education can only be one category, so suppression of Highest level of education should coincide with suppression of the other two mentioned variables

To apply the traditional approach, excluding the local suppression, both sdcMicro and μ-ARGUS can be used. As discussed before, we have proposed two risk measures: k-anonymity and the all-m approach. Depending on the chosen risk measure, the local suppression can be optimized in different ways. For the k-anonymity situation we suggest to use sdcMicro. For the all-m approach we suggest μ-ARGUS.

4 Discussing the disclosure risk

The synthetic and the traditional approach differ in the way one has to look at the disclosure risk. Therefore, we will discuss both of them separately.

4.1 EU-SILC

In the synthetic data approach, the dataset is fully synthetic. That is, the data are simulated based on estimated distributions of the variables. In Templ and Alfons (2010) a general discussion on disclosure risk in case of (fully) synthetic population data is given, with an application to EU-SILC data as simulated in the AMELI project. In that paper, five disclosure scenario's are considered. The general outcome is that even in case of a very knowledgeable intruder (he has information on the data generation process that produced the synthetic data), the disclosure risk is very low. Moreover, even if the intruder is able to identify an individual, the probability that derived information is close to the original value is extremely low.

In our case, households with (close to) unique structure could be identified. For example, a large household that occurs multiple times in the PUF but always with the same structure (age and sex distribution) is likely to be a sample unique. However, the associated simulated income differs from the true income.

To reduce the disclosure risk of unique households, one might consider to remove those households from the PUF. This generally leads to a bias in estimates based on that PUF, but considering the intended use of this PUF this appears not to be a big problem.

4.2 EU-LFS

In the traditional approach, the identifying variables need to be considered in more detail. The general idea was to use k-anonymity as disclosure risk measure. However, the number of identifying variables is rather high in our case. So using all identifying variables as key variables in the k-anonymity measure would be rather problematic. Hence a subset of all identifying variables needed to be chosen.

Another approach would be to consider all combinations of identifying variables up to m dimensions. In some sense these can be considered to be the m-dimensional marginals of a k-anonymity measure with all identifying variables as key variables. Using a larger threshold for those m-dimensional combinations would give some more 'slack' on the higher dimensional combinations.

5 Usefulness of the public use files

To assess the usefulness of the PUF some utility measures were defined, measuring the relative differences between the SUF and the PUF. Essentially this could be something like

$$\frac{\text{Value(Indicator} \in \text{PUF)} - \text{Value(Indicator} \in \text{SUF)}}{\text{Value(Indicator} \in \text{SUF)}}$$

where Indicator might be something like the distribution of individuals by Sex, Age, Education, etc. to reflect some basic measures.

The Indicator in the formula could also be taken to be data dependent, using so called 'main indicators'. E.g., for the EU-SILC data one might think about the at-risk-of-poverty rate and for the EU-LFS data about the (un)employment rate.

Additionally, some model-based measures were proposed, based on the confidence interval overlap measure proposed by Drechsler (2009). As a model in case of the EU-SILC data one might think of

$$\log(\text{equivalenced disposable income}) \sim \text{age} + \text{gender} + \text{education} + \text{citizenship} + \text{hhsize}$$

and in case of the LFS data of logistic regression on 'is-employed' with age, education, citizenship and hhzise as explanatory variables.

6 State of play of the project

At the time of writing of this paper, the project team was still evaluating the proposed methods for the production of PUFs. The proposed methodology was applied to datasets of some of the partners. Most of the time was devoted to improving technical solutions to apply the methodology and to discussing the proposed disclosure risk measures and utility measures. At the end of the project (December 2015) all partners of the project are supposed to have produced prototype PUFs of EU-SILC and EU-LFS. Whether these prototype PUFs will become 'real' PUFs depends on the decision by Eurostats Working Group on Statistical Confidentiality and by the respective subject matter working groups. Note that, in case these working groups decide positively on the approaches, it will always be up to the individual member states to decide whether or not to release a PUF on EU-SILC and/or EU-LFS data.

The project will also deliver guidelines to produce the PUFs, along with technical solutions like R-scripts and other software.

References

Alfons, A., Kraft, S., Templ, M. and Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, **20**, 383–407.

Bujnowska, A. (2015). Access to EU microdata for research purposes. Presented at the UNECE/Eurostat work session on statistical data confidentiality, Helsinki, Finland.

Drechsler, J. and Reiter, J.P. (2009). Disclosure Risk and Data Utility for Partially Synthetic Data: An empirical Study Using the German IAB Establishment Survey. *Journal of Official Statistics*, **25**, 589–603.

Eurostat (2014). *EU Labour Force Survey Database User Guide*, version December 2014, http://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf P.P. (1998). Post randomisation for statistical disclosure control: theory and implementation. *Journal of Official Statistics*, **14**, 463–478.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.P. (2012). *Statistical Disclosure Control*, ISBN 978-1-119-97815-2, Wiley.

Templ, M. and Alfons, A. (2010). Disclosure risk of synthetic population data with application in the case of EU-SILC. In *Privacy in Statistical Databases, proceedings PSD 2010, Corfu, Greece*, LNCS 6344, Springer.