

From Survey to Social Indicators: The Data Management System of EU-SILC in Austria (preliminary draft, version of 30/10/2006)

Nadja Lamei*

*Statistics Austria, Directorate Populations Statistics - Social and Housing Statistics, Vienna, Austria (nadja.lamei@statistik.gv.at)

Abstract: This paper gives an insight into the data management system used for EU-SILC in Austria, from the time a question is asked in a household and keyed into the CAPI-laptop to the final statistical analysis. Four years of EU-SILC in Austria as a survey and the requirement to produce consistent micro-datasets, fully checked, imputed and weighted have brought up numerous challenges. An integrated system of data management was developed to meet these. It aims at fulfilling the following criteria: highest possible automation, full transparency, and expandability. The practical implementation was done in SPSS-code. Special emphasis in this system is put on ways to streamline the data editing process and become more and more efficient each year whilst at the same time improving data quality.

An article in German together with Richard Heuberger dealing with this topic will be published in "Statistische Nachrichten" Nr. 11/2006.

1. EU-SILC in Austria

EU-SILC (Statistics on Income and Living Conditions) was launched in Austria in 2003 as a cross-sectional survey. In 2004 the rotating panel with a cross-sectional and a longitudinal component was introduced: every year about a quarter of households comes newly into the sample, three quarters are reinterviewed. According to the EU-SILC regulation a minimum effective sample size of 4.500 households per year has to be reached, in the longitudinal component between two survey years it is 3.250 households.

As there are no administrative data sources to build a reliable household income at the moment, all data are surveyed. At the moment the field work is outsourced. The mode used is face-to face CAPI (*computer assisted personal interviewing*). Major advantages compared to PAPI (*paper and pencil interviewing*) are electronic data collection instead of manual data entry after the survey, automated filters, and the possibility of integrated data checks. CAPI is not a necessary prerequisite for good data, but it helps to reduce measurement errors and inconsistencies at an early stage and thus reduces the effort to be put in the actual data editing process.

Data management, as we see it, begins with building checks that run during the survey and continues with control mechanisms that are activated as soon as the first data come in. On basis of those early checks deficiencies to the data are reported back to the field work institute and call-backs are started. Our aim is to do as much as we can in the field and leave as little as possible to post-field work data editing.

When the process of data collection and early checking is finished data have to be edited, imputed, weighted, analyzed and indicators computed. As those tasks are nearly the same each year it was a logical decision to once build up a data management system that can more or less be used for following waves.

2. The data management system in theory

2.1. Demands: A complex survey requires a simple and transparent data management

The requirement to send fully consistent, checked, imputed, and weighted micro-data sets to Eurostat each year with the slightest possible time lag to the survey as well as indicators of social cohesion, have made it necessary to standardize and structure the work as much as possible. Much effort is put into the development of editing steps. Those steps should ensure internal and external cohesion and thus data quality in general. Also the questionnaire and the editing steps have to respond to changes in national social transfers systems. Plus, due to the longitudinal component more and more information becomes available each year that helps data editing. Data management thus has to account for highest possible standardization on one side but on the other side remain as open and amendable as possible. Documentation for all editing steps either in written descriptions or in program code is also a key element of the data management system.

EU-SILC is different from other surveys in many ways and those particularities determine the demands towards the data management process. To sum up, particularities of EU-SILC are:

- the emphasis put on the household context and the interdependence between personal and household data
- the longitudinal component that plays an important role
- the subject matter, i.e. the living standard of private households, and changes of the object of investigation – the income and its different components

Household context: The importance of the household context can be seen in the indicators based on EU-SILC, e.g. the at-risk-of poverty rate which is based on household income and composition. Therefore it is not sufficient to have consistent personal data; data for all persons of one household have to be “in line”. Data editing therefore has to take all the other personal data into account as well as the household as a whole. To give a simple example: when a retired woman living alone does not have any income at all, this is highly implausible and most likely a surveying error. But when this woman lives together with a partner or the family of her children she does not necessarily have to have an income of her own. Those differences of the household context have to be taken into account already in the survey and then again in the data editing process.

Longitudinal component: Information from previous or subsequent waves can help solve problems. If the old woman from the example above did have an old-age pension in 2004 and 2005 but reports none in 2006 this is highly unlikely. Data checks that run during the field work should detect this inconsistency and a call back should be made to either correct or verify this information. But: as matter of fact information is not always that clear and data are not always available at the time needed, e.g. data from the 2006 survey could only be used to solve few cases in the 2005 cross-sectional data, as the survey was still ongoing in the field when data for 2005 had to be finished. Also, our experience shows that it is often necessary to have data from three different waves to validate information.

Subject matter and changes: The need to assess the households’ living standard realistically poses great challenges to field work, but also to the data editing process. And, as mentioned above, the system has to be flexible enough to react to changes in the social transfers system or to changes because a question seems to be faulty.

With regard to those necessities the major advantages of our data management system prove to be that it

- helps standardize the checking, editing, imputation and finally the analysis process as much as possible,
- documents each single change to the data, so that the whole process is transparent,
- can be amended wherever and rerun whenever necessary.

Of course, the data management system we set up once is not nearly finished yet and since its first implementation has undergone many changes. But that ability to make amendments works as a constant invitation to all people involved to improve the system. Furthermore, as this is not necessarily the case with other surveys, we should add that the whole process is managed by the EU-SILC team, except for the weighting which is done by the methodology department, no external programmers are involved. The main advantage of this is that all decisions can be taken with the subject matter in mind rather than from a technical perspective only.

2.2. Build-up: A modular structure of interdependent programs

What needs to be done to the data, determines the structure of the data management system. The data have to undergo the following procedures:

1. Checking raw data and feeding back inconsistencies to field work
2. Editing income-variables
3. Editing non-income variables¹
4. Imputations
5. Weighting and non-response analysis
6. Calculating Eurostat target variables
7. Calculating Laeken indicators and other key indicators
8. Creating tables for publication
9. Creating a data set for external data users

Those modules are from a contextual perspective strongly related and are processed in the order cited above. But from a program perspective they are relatively independent. Changes in one module affect of course the outcome of the other modules and the final results, but there is no need to change the program code of the other modules. Also, module order is not as strict as it may seem. It may e.g. be necessary to calculate Eurostat target variables at an earlier stage as listed here, to do some tests. Or to calculate household income when not all personal income components have been finished yet. That is, of course, possible and this flexibility is much used and was foreseen in the technical implementation.

¹ We distinguish between income and non-income variables because of the significance of income variables in EU-SILC and the necessity to have data “fully checked, edited and imputed in relation to income” (EU-SILC Regulation L165/2).

3. The data management system in practise

3.1 From raw data to checked, imputed and consistent data files

As mentioned earlier we try to have as many data checks as possible during the field work period. Nevertheless, measurement errors to some degree remain and have to be treated in post-field work procedures – when they are known they can be corrected. False or inconsistent values in the data can stem from surveying errors (effects of interviewers on respondents or respondent errors due to lack of understanding or errors due to the questionnaire), errors in data entry or errors from data editing².

Several methods are used to detect errors and verify information, from checkup-calls to consistency checks between waves and comparisons to external sources. Missing values are a special problem; in relation to income it is especially important to have values for each income component to build household income and thus make inferences about the living standard of the household possible³. To correct errors or input data where they are missing we can distinguish

1. cross-sectional and longitudinal methods
2. deductive and deterministic/stochastic methods.

Those can usually be combined i.e. deductive methods can use information from previous waves and thus run on longitudinal level or be cross-sectional and take information from the current survey. Which methods is used depends upon availability of information and “best practice” (e.g. from the ECHP)⁴.

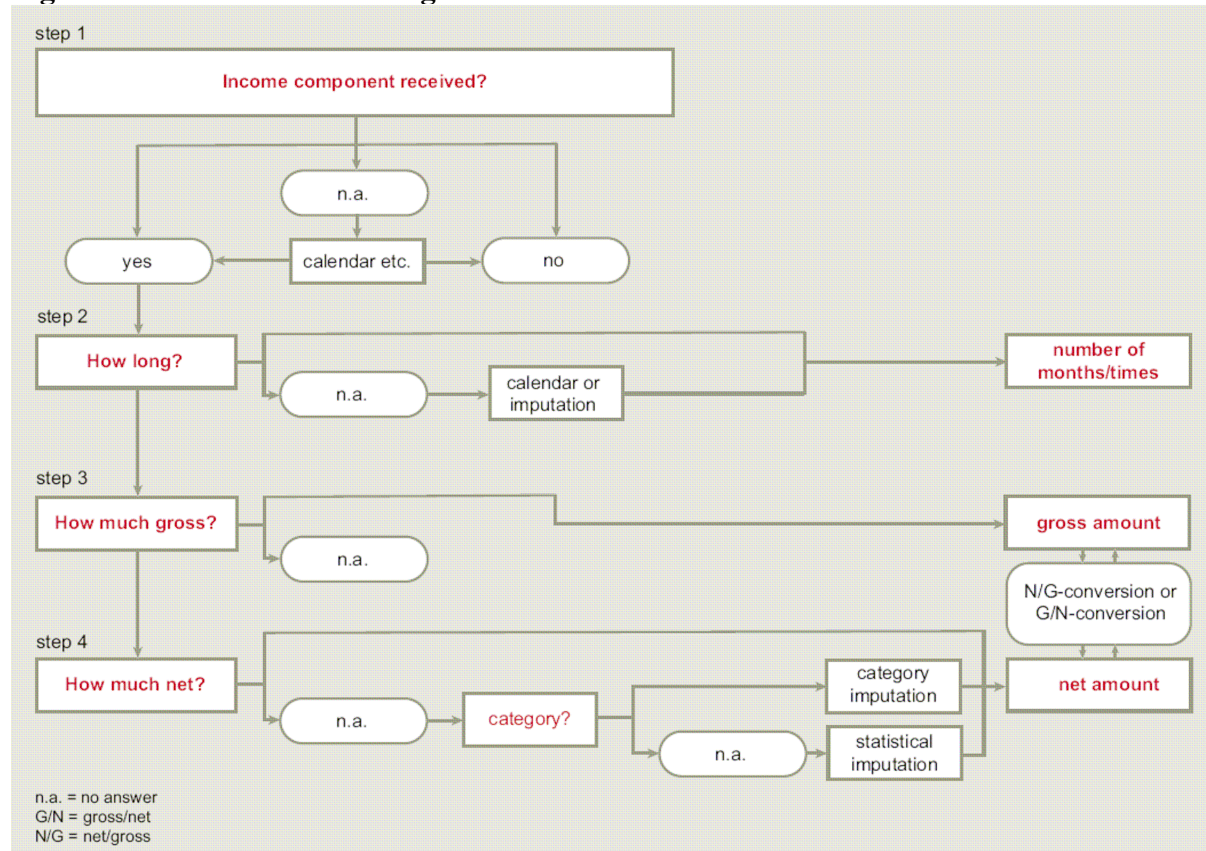
Instead of going into detail about the technical side of data imputation let us take a look at how missing data occur in EU-SILC. Missing income values can have three reasons: first the answer if an income was received can be missing, second the duration of this receipt during the income reference period can be missing, third the actual amount can be missing (see figure 1). Let us take a look at the single steps of missing income value treatment. The steps are in downward hierarchical order, which means that if a value is already missing in step 1 – income received or not – the respondent doesn’t get questions on duration or amount and these values are missing as well. But: if the answer was “yes, income was received” and the number of months is missing we go on asking for the gross and net amount. The order of gross and net questions depends on the income component. For some components only the net amount is applicable.

² A detailed description of survey errors can be found in Groves (1989).

³ This article only describes treatment of missing income values as those are the most important ones in EU-SILC.

⁴ ECHP Doc. Pan 164/00.

Figure 1: Treatment of missing income information



Ad step 1) When the answer whether an income component has been received is missing a value is imputed with a deductive method.

Longitudinal imputation: If information from a previous wave is available and if no change is indicated between the years (e.g. activity calendar, current main activity etc.) a value from a previous wave is imputed; else the missing value is treated in cross-sectional imputation. Some components are more suitable for this kind of imputation than other, e.g. pension receipt only rarely changes from “yes” to “no” between two years whereas a person having unemployment benefits may not necessarily be unemployed and receive benefits the next year.

Cross sectional imputation: If no information from a previous wave is available or major changes have taken place in the personal or household circumstances or income components are very changeable a check of activity calendar, current main activity, other income receipts, household income and so on should clear up if an income component receipt seems plausible or not. Decisions on which variables to evaluate as relevant explicatory variables are taken for each income variable separately and put into general models – only if it is absolutely necessary single case imputations are done.

If no relevant information is available the missing answer is considered as “no”. As only about very few values are missing per component (e.g. less than half per cent with “easy” income types as pensions or employee income) these methods usually bring good results⁵.

Ad step 2) When the duration of an income receipt (usually measured in months, sometimes in times per year) is missing, we try to infer from other information, e.g. the activity calendar

⁵ Initially, i.e. in EU-SILC 2003 und 2004, item non-response was higher, but by means of better interviewer training and better compliance of the respondents these values could be decreased.

or the durations of other income receipts (some income components can run parallel some are ruled out by others). If no inference is possible we impute a random value based on the distribution for responses. Only few values for duration of receipt are missing, e.g. for unemployment income about 1 per cent of the respondents in EU-SILC 2005 who had a receipt did not tell for how long.

Ad step 3 and 4) There are different ways to answer the amount question: gross, net or if neither of them is answered a show card with categories is provided. If no actual amount is given but a category is available we have to convert this to a value. This *category imputation* is done by taking a random value in accordance to the distribution of amount responses in the relevant income category (i.e. by use of a density function). If either a net or a gross value are available we do *net/gross or gross/net conversion* based on different models of tax and social security. If no income at all is given imputations for net amounts are done and then the imputed net values are converted to gross amounts. Here we also distinguish *longitudinal* and *cross sectional imputation*. Where they are possible longitudinal imputation methods are preferred because they have the advantage of creating consistent amount between the waves⁶. Cross-sectional imputations are usually done with regressions methods, several regression models are specified for each income component using relevant predictive variables. The estimated values are added with a residual term to prevent attenuation of the variance. Sometimes, especially for income components with only few recipients, regression models are not viable and median imputations are used.

To give an example of how often statistical imputation for net income values is necessary: In EU-SILC 2005 3 per cent of the employee income had to be imputed, thereof about a quarter was imputed by longitudinal method and the rest by a cross-sectional model.

When the data are finally checked, imputed and weighted Eurostat target variables other variables for analysis and social indicators (including Laeken indicators) are computed. An extensive process of data analysis sets in which aims at assessing the quality of the data. Comparisons of results to previous EU-SILC waves and to external data e.g. wage tax statistics, National Accounts or the microcensus are important steps. Thus, analysis and indicators are not only the result of EU-SILC but they function also as a means of quality control. Also Eurostat's checking process is an important part of quality control. Further valuable input is provided by external micro-data users⁷ who comment on the data themselves but also on the meta-information provided. If at this stage errors are detected the data editing process is rerun to correct for those errors⁸.

This overview on the contents of the data management process shall now be complemented by an overview on the technical implementation.

3.2. Technical implementation

Programming

The data management in EU-SILC is implemented in SPSS for PC. Every single one of the modules named above is handled by a main-control-syntax and included where it is needed. The modules themselves are broken down into sub-tasks, each one corresponding to a syntax

⁶ The method used is row- and-column imputation proposed by Little and Su (1989).

⁷ Austrian EU-SILC micro-data sets are provided to external users who are usually affiliated to universities or private research institutes for a small fee.

⁸ This does not necessarily mean that the error stems from the data editing process itself, it can also be an interview error that has not been detected up to the stage of nearly final data.

file. Often those sub-tasks are structured according to an income component concept, e.g. there is a file for cross-sectional imputations for employee income, one for unemployment benefits, one for health related income and so on. By this breaking-down into small parts the system keeps clear despite its total size and complexity.

Valuable features in SPSS are macros because they enable even further automation and standardization. Macros are commands that create commands and run them. A simple example would be a macro that expands into a more complicated path: `!data05` stands for `P:\SILC\EU-SILC-2005\data`. During the development of data editing procedures a great number of repetitive tasks and checks were identified. For macros used more than once it has proven to be useful to separate the macro programming from the actual use of the macros and so a kind of library with all those macros was set up. Now, before running a program the first step is to include a preparation file with all those macro definitions.

As up to five persons work on the data editing at a time this requires common rules⁹:

- Only use common paths and files. Temporary files should only be saved in a temp-folder on the local drive, files that are needed for documentation or to be further processed need to be saved on the network.
- Each program has to be fully labeled with title, function, author, and date. It should be apparent what is done and why.
- Only create new variables if necessary, do not duplicate variables, delete auxiliary variables after use.

Variable names

For a program to run for many waves it is important to have almost unchangeable variable names and variable content. As explained above changes and amendments are, however, necessary to keep the high quality standards of EU-SILC and improve them. Even in longitudinal surveys which ask for continuity slight changes to variables and questions are quite common. Changes to variables, new codes etc. are documented in lists that are practical tools to adapt the program code each year.

We distinguish between *survey* and *derived variables*. Survey variables correspond to actually asked questions or information otherwise gathered during the survey. Derived variables are built ex-post to help the checking and analyzing processes or to flag other variables (see below), also Eurostat target variables are usually computed from survey variables. Survey variables are not – like variables used for analysis – named meaningfully (“sex” for sex), but with a combination of letters and numbers. Those have of course their intrinsic meaning according to the rules for variable names used in EU-SILC.

Programming pre-conditions as starting variable names with a letter and using no more than 8 digits are the backbone of the variable name convention¹⁰. The final decision on how to build variable names, however, is strongly related to variable content. The key part of the questionnaire is about income. The structure of the questions is repetitive with only slight differences from income component to income component. This typical question order (income received – duration – amount net and gross) and the need to recognize each variable

⁹ For general advice on SPSS-syntax programming see Levesque (2003). Rules given here have been developed in the context of this particular task of the EU-SILC data management and may not generally apply to other tasks and program structures.

¹⁰ Although recent SPSS versions allow for longer variable names for reasons of compatibility only 8 digits are used.

and which type of question it corresponds to without using the codebook has triggered the following name convention (see Table 1).

Table 1: Variable Name Convention								
<i>Meaning of the digits</i>								
digit 1:	identification of the data file (D, R, H, K, oder P)							
digits 2-4:	number of the question corresponding to the questionnaire							
digits 5-6:	number of the sub-item							
digit 7:	information on variable type							
	0=no income variable							
	1-5=income variable:							
	1=component received yes/no							
	2=frequency of receipt							
	3=gross amount							
	4=net amount							
	5=category							
Digit 8:	to be filled optionally, e.g. with "F" for flag variable							
Example P057024:								
digit	1	2	3	4	5	6	7	8
variable	P	0	5	7	0	2	4	F
P.....	The variable comes from the personal questionnaire.							
P057...	It is question number 57 from the personal questionnaire.							
P05702...	It is the second item of question 57.							
P057024	It is the question for the net amount.							
P057024F	It is the flag variable of variable p057024.							

First, for each variable the questionnaire or the data file which it belongs is stated. "D" stands for the household register of all households, also the non-successfully contacted. "R" is the code for the personal register with basic information on all household members including former household members in participating households. "H" is the household questionnaire; "K" the questionnaire on child care (in Austria not only questions about children up to 12 years but also for children aged between 12 and 16, e.g. the type of school they attend, are included); And P stands for the personal questions asked to every person aged 16 or older. Digits 2 to 4 (number of the question) and 5 to 6 (number of the sub-item) facilitate the linking of questionnaire and codebook in giving the same numbers to questions and corresponding variables. Digits 7 and 8 are of great importance for efficient programming.

A good example of the use of variable names in combination with macros is a check for gross-net-relations (see Table 2). Cases shall be listed that have a gross value which is less than a net value or a net value that is less than 40 per cent of the gross value or where net and gross are equal. Listed cases have to be checked, whether this is an error or plausible. As this general check applies to all income components where there is gross and net values it would be not efficient to program this for each variable. But because gross variables always end in "3" and net variables end in "4" this can be easily automated. The only thing you have to provide for the check to run is a list of gross variables.

Table 2: Example of the use of standardized variable names in a macro

define brunet (!pos !cmd).	←	macro definition
!do !a !in (!1).	←	!a is wild-card for gross variables that will be used as input.
!let !b =	←	!b is wild-card for net variables that correspond to gross
!concat(!substr(!a,1,6),4).		variables in digits 1 to 6 but then have the number 4 in digit 7. Those are not input but called in the formula.
compute nebrpro = \$sysmis.	←	building a test variable
do if !a gt 0 and !b gt 0.	←	Test is run if gross and net variable have values greater than 0.
if !a lt !b nebrpro =1.	←	If gross variable value is less than net variable value the test variable value is set to 1.
if !b lt 0.4 * !a nebrpro = 2.	←	If net variable value is less than 40 per cent of gross variable value the test variable value is set to 2.
if !a eq !b nebrpro = 3.	←	If gross variable value equals net variable value the test variable value is set to 3.
end if.	←	end of testing
temp.		
sel if nebrpro >= 1.	←	All cases that have a test variable value equal to 1 or greater are listed.
list var pid nebrpro !a !b int.		
!doend.		
exe.		
!enddefine.	←	end of macro definition

Flag variables

Besides the identification of variable content with the help of standardized variable names it is also necessary to identify what is done to the variables in the whole process. It is important to distinguish values that originate from the survey, values that were edited and values that were imputed. This allows for the analysis of the share of imputed cases and analysis alike. After running each of the program steps intermediate data files are saved that would allow for reconstruction of this kind of information, but this would be complicated and time consuming. A flagging system where the most basic meta-information is located in the same file is more practical.

Each income variable has a „partner“ with the same name but ending in an additional „F“ for flag. These flag variable contain meta-information of the following kind:

- 2 not applicable
- 1 no answer and not (yet) imputed
- 1 value according to survey
- 2 value from category imputation
- 3 value from net-gross or gross-net conversion
- 4 value logically deduced
- 5 value statistically imputed with longitudinal method
- 6 value statistically imputed with cross-sectional method
- 7 value from survey was corrected
- 8 value computed from a monthly income (this code applies only to variables of yearly income)

A missing value will in the beginning be flagged with -1 but should transform to a non-missing value during the process of data editing and imputations and be flagged with 3, 4, 5 or 6 respectively. The majority of all values has the flag 1 (or 8 if you look at a yearly income variable that was asked as a monthly variable). Category imputation as a means to transform answers to income categories into actual values was described above and is flagged with code

2. Gross variables that are flagged with 3 were computed from an existing net value, and vice versa net variables with flag 3 from gross value – never can both, gross and net variable, have code 3. “Value logically deduced” (code 4) means that it was not asked but could be derived from other information; e.g. that is the case for child care allowance. For imputed values codes 5 and 6 are used. “Value corrected” (code 7) applies, for example, if a check of gross-net relations showed that there was a typing error: if the gross value is 1.500 Euro and the net values is 130 and we assume that it should be 1.300 we multiply the variable with 10 and code the flag with 7.

Of course, also these flags only picture an instant and they can not indicate more than one transformation to a value, but for standard procedures this is sufficient information.

4. Perspectives

Data management in the Austrian EU-SILC survey, as introduced in this article, was first implemented for the survey year 2004 and since than has been used in adapted forms for EU-SILC 2005 and 2006. Besides improving the existing programs in the months to come we will focus on the improvement of and new developments for the longitudinal component as in spring 2007 longitudinal files will be delivered to Eurostat for the first time. Also documentation of the work in textual form rather than only in SPSS-code will be advanced as well as the description of the analysis variables.

References

- European Commission (2000), *Imputation of Income in the ECHP*, Doc. Pan 164/00, November 2000.
- Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.
- Levesque, R. (2003), *SPSS Programming and Data Management. A Guide for SPSS and SAS Users*, Chicago: SPSS Inc.
- Little, R.J.A. /Su, H. (1989), *Item Non-response in Panel Surveys*, In: Kasprzyk et al., *Panel Surveys*. New York: Wiley, pp. 400-425.
- Regulation (EC) No 1177/2003 of the European Parliament and the Council of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC) and corresponding Commission Regulations.