# *Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki*

## *Administrative data and registers in EU-SILC*

*Rihard Tomaz Inglic*
*The Statistical Office of the Republic of Slovenia*
*rihard.inglic@gov.si*

# *Chapter 1. Background of EU-SILC (EU Statistics on Income and Living Conditions)*

The implementation of EU-SILC was a big challenge to the Statistical Office of the Republic of Slovenia (SORS) due to the fact that SORS is register-oriented, which offered us the possibility to use a lot of data from registers and other administrative sources and at the same time to reduce the response burden of the households to the minimum.

The main definition of EU-SILC is written in the frame regulation: EU-SILC is a survey about income and living conditions and is conducted on the basis of the European Commission regulation. The aim of this regulation is to establish a common framework for the systematic production of Community statistics on income and living conditions, encompassing comparable and timely cross-sectional and longitudinal data on income and on the level and composition of poverty and social exclusion at national and European levels. EU-SILC is a new panel survey. Before, in EU-15, ECHP was conducted from 1994 to 2001, after that EU-SILC was introduced. The main difference between ECHP and EU-SILC is that ECHP was inside harmonised, which means that questionnaires are harmonised, but EU-SILC is outside harmonised and this means that outputs are harmonised.

EU-SILC is a harmonised survey, this means that all countries which conduct the survey have to respect the harmonised methodology, or better - certain rules, which ensure common, comparable results. The questionnaires are not harmonised, but the variables to be reported to Eurostat are defined precisely. The majority of countries use only the survey questionnaires to collect the required variables, but some countries combine survey questionnaires with registers and other administrative sources, and this is also the practice in Slovenia. EU-SILC covers different areas. According to the regulation we must collect basic household data, total household income, gross and net income components at household level, housing and non-housing related arrears, variables for measuring social exclusion, physical and social environment, child care, dwelling type, tenure status, amenities in the dwelling, housing costs. On the individual level the following areas must be covered: demographic data, gross personal income, education, labour information, health, and every year an ad hoc module is added. At SORS we analysed all available registers and other administrative sources and defined the variables which can be taken from these sources; after a demanding analysis a set of variables connected mainly to the incomes of the persons/households was defined, which enabled us to create a short and friendly questionnaire for households/persons. The questionnaire consists mainly from the questions connected to the living conditions, housing and opinion questions.

# Chapter 2. Development of EU-SILC in Slovenia

## Chapter 2.1 Legal grounds for using registers in EU-SILC

The decision on broad use of registers and administrative data was possible only because of the Slovenian National Statistical Act provides the legal base for such an approach. In the National Statistics Act, Article 4 defines that the reporting units shall be holders of official and other administrative data collections (records, registers, databases, etc.), and also natural and legal persons that are defined by the programme of statistical surveys as data providers.

According to the stated Act, official collections shall be data collections, established by regulations or general acts of public power holders, on the basis of which certificates and public documents shall be issued.

According to the stated Act, the administrative collections shall be other data collections, which are kept and maintained by the holders under the previous paragraph hereof. [2)]

According to the National Statistics Act, the Statistical Office has the right to get access to all administrative sources in Slovenia and to use them for statistical purposes.

## Chapter 2.2 Pilot Survey in 2003 and 2004

As the first step in implementing EU-SILC we conducted two waves of pilot surveys (in 2003 and 2004, respectively) with a sample of 300 households. In the pilot survey we tested questionnaires, CATI interviewing (computer assisted telephone interviewing) and some possibilities to extract some data from the registers. After having conducted the pilot, we established that the burden on households could be reduced even more by using register data and data from other administrative sources. In pilot surveys the imputations according to Eurostat rules for pilot surveys were not performed and because of this we did not discover that some data were missing in a larger extent. The test was successfully finished according to the knowledge we had at that time. We analysed the data from the pilot survey and we found out that it would be possible to conduct EU-SILC by using registers.

## Chapter 2.3 Conducting the main survey in 2005 and further

In February 2005 we began with the implementation of the regular annual survey. In the first year, when a large sample of approximately 13,500 households was chosen, we interviewed all persons face to face with the paper questionnaires. We know that each household will participate several years in the survey. When the household participates in the survey, it participates 4 consecutive years. We decided that from all households, from which we would acquire their telephone number and would be interviewed within the next years by CATI, only those households which would participate for the first time in the survey are to be interviewed in the field (CAPI – computer assisted interviewing). The survey can be conducted by phone, because the survey was shortened due to use of registers. We began to use such a mode of interviewing in 2006. In 2006 approximately 6,000 households were interviewed by CAPI and 8,000 by CATI. The data from 2006 will be available at the end of 2007.

Due to the fact that we could use a lot of administrative data (especially data on income) we were in the position to design a relatively friendly and short survey questionnaire. The paper questionnaire was used for all households in 2005. In 2006 households which participated in the survey a year before were interviewed by phone and only for the “newcomers” in the survey CAPI (computer assisted personal interview) was used.

## *Chapter 2.3 Sources of the data for EU-SILC*

EU-SILC is the first sample survey where we used different sources and combined them. Some of these sources were acquired from different institutions; some of them were obtained from different surveys which are conducted by SORS. In the sources from outside, there were included all the persons living in Slovenia, but in EU-SILC only 8,287 households participated. The frameworks which define the persons, who are included into the database, are questionnaires. For all persons who participated in the survey, all the data from other sources were merged into the database. In EU-SILC we used the following sources:

| **Institution** | **Source** |
|---|---|
| The Statistical Office of the Republic of Slovenia | • Questionnaires |
| Tax Authority | • Tax income register<br>• Tax register for income from self-employment |
| Ministry of Labour, Family and Social Affairs | • Family allowances (parental allowance, childbirth allowance, child allowance, large family allowance, allowance for care of a child needing special care and protection, part payment for lost income and compensation for childbirth leave)<br>• Social allowances<br>• Housing allowance (up to 2006, after that this register will be kept in the Ministry for Environment and Spatial Planning |
| Pension and Disability Insurance Institute | • Old age benefits<br>• Survivor's benefits<br>• Disability benefits |
| Employment Service of Slovenia | • Register of unemployed persons<br>• Unemployment benefits |
| Health Insurance Institute | • Activity status for inactive persons |
| Central Register of Population | • Addresses (for sampling)<br>• Marital status<br>• Birthday<br>• Country of birth |
| The Statistical Office of the Republic of Slovenia | • Statistical register of employment |
| The Statistical Office of the Republic of Slovenia | • Survey on scholarships |

At first glance the use of administrative data seems attractive and simple, but in reality this means that many agreements with different institutions have to be agreed upon, prepared and signed. At the same time the protocol for technical structure and transmission of the data has to be agreed.

The advantages of using registers are as follows:

1. A shorter questionnaire and consequently less time used for interviewing.
2. Skipping the most difficult and sensible questions about income.
3. More accurate data.
4. Answers are to a lesser extent affected by forgetfulness of interviewers.
5. Item non-response as well as unit non-response is lower.
6. Use of administrative data means lower costs for conducting the survey.

Of course, using registers has also significant disadvantages:

1. It is more difficult to compose all data.
2. A lot of work is required to ensure logical integrity of data.
3. Cleaning and editing the data take much more time.
4. Some persons are not in registers which causes another level of problems.
5. The technical processing of data is much more demanding and time consuming.
6. Timeliness is a problem, because some registers are not available on time.
7. Administrative sources can change each year in the sense of variables as well as their definitions.

After conducting the survey in 2005 with PAPI (paper assisted personal interviewing), the first task was to enter the data. After having entered all the data into the database, we had to define the key (PIN) for each person which enabled us to extract the data for these persons from the registers and other administrative sources. In Slovenia we do not have a register of households or dwellings. Because of this, we were able to sample only one person from the household, whereas all members of the household participated in the survey. Thus we collected personal data for each of them (name, surname, birthday and gender). With these data we could compose the PINs which were the key variable used to merge with all other sources. This process was composed of two stages. In the first stage we searched for the PINs by a computer program. In case the person was enlisted in the survey with completely regular data regarding his/her name, surname, birthday and gender, the computer could find his/her PIN. In this stage we found approximately 85% of PINs. For the other 15% manual searching was used. This is a relatively time-demanding process. At the end we managed to find more than 99% of PINs and only some (0,002%) of the PINs had to be imputed.

When we collected all the data, we began to compose the EU-SILC database. We found out that it can happen that a person was included in the central register of population, but s/he was not to be found in any other registers. We assume that such persons live in Slovenia and work abroad. All the data for such a person should be imputed. In EU-SILC we used income variables to calculate the imputation factors which told us the percentage of the income that was imputed.

The following table presents in how many cases (in %) no income was imputed:

Table 1: The share of imputations by income variables

| Kind of income | Source | % of cases without imputations | % of cases with partial imputations | % of cases where all income was completely imputed |
|---|---|---|---|---|
| Employee cash or near cash income | Tax records from Tax authority<br>Questionnaire | 59.9 | 34.8 | 5.4 |
| Non cash-employee income (company car) | Questionnaire | 32.7 | 39.7 | 27.6 |
| Cash benefits or losses from self-employment | Tax records from Tax authority<br>Questionnaire | 68.7 | 7.7 | 23.6 |
| Contributions to individual private plans | Questionnaire | 75.1 | 0.7 | 24.2 |
| Unemployment benefits | Register of unemployment benefits | 100.0 | 0.0 | 0.0 |
| Old age benefits | Register from Disability and Pension Institute | 100.0 | 0.0 | 0.0 |
| Survivor's benefits | Register from Disability and Pension Institute | 100.0 | 0.0 | 0.0 |
| Disability benefits | Register from Disability and Pension Institute | 100.0 | 0.0 | 0.0 |
| Education related allowances | Statistical survey on scholarships | 100.0 | 0.0 | 0.0 |
| Income from rental of a property or land | Tax records from Tax authority | 100.0 | 0.0 | 0.0 |
| Interests, dividends, profit from capital investments in unincorporated business | Tax records from Tax authority,<br>Questionnaire<br>The data are collected on household level | 88.1 | 0.3 | 11.6 |
| Family/children related allowances | Register of Ministry of Labour, Family and Social Affairs | 100.0 | 0.0 | 0.0 |
| Social classifications not else where classified | Register of Ministry of Labour, Family and Social Affairs<br>Questionnaire | 98.4 | 0.1 | 1.5 |
| Income received persons aged under 16 | Tax records from Tax authority | 100.0 | 0.0 | 0.0 |

Source: EU-SILC 2005

Table 1 illustrates that we did not perform many imputations in cases where registers were used. In cases when questionnaires were used, much more data were imputed. The main reason for this was that persons did not know the answer at the time when the interview took place. Especially in case of a proxy interview, much more data were missing. We also realized that the most difficult cases were those involving incomes from self-employment. As income from self-employment we took into account: profits, losses from self-employment and income from agriculture as well. We found out that especially the income from agriculture posed problems. We collect data on income from agriculture with a survey questionnaire and from administrative source as well. At the end we found out that the replies in questionnaires in some cases did not

provide data of adequate quality, so we decided to use the mixed manner of collecting data in future.

Table 2: Share of imputations according to original data for the variable "employee cash or near cash income – gross"

| Share of income imputed according to original data | percent |
|---|---|
| Entire income imputed | 5.4 |
| 50.1 % - 99.9 % of original amount imputed | 0.6 |
| 25.1 % - 50.0 % of original amount imputed | 0.9 |
| 10.1% - 25.0 % of original amount imputed | 8.6 |
| 0.1 – 10.0 % of original amount imputed | 24.7 |
| No imputations | 59.9 |
| Amount of income decreased after imputations or editing | 1.0 |

Source: EU-SILC 2005

If we look at table 2, we can see that we collected all the data for 59.9% persons. In case of 34.8% of persons part income were imputed – this happened in case we got the data from the registers, but we did not collect the data from the questionnaire. In the questionnaire there was only a little part of this kind of income (compensation for meals and allowance for travel to/from work). Data were completely imputed only in case we found that person was in employment, but he/she did not receive any amount from there. In some cases we found out that the person was in employment and we did not manage to find him/her, so in such cases the complete income from employment was imputed.

Table 3: Percentage of households which received definite kind of income according to EU-SILC and Household Budget Survey, reference income year 2004

| Kind of income | EU-SILC | HBS |
|---|---|---|
| Employee cash or near cash income | 40 | 41 |
| Non cash-employee income (company car) | 1 | 0 |
| Cash benefits or losses from self-employment | 13 | 7 |
| Unemployment benefits | 3 | 2 |
| Old age benefits | 20 | All pensions benefits together 25 |
| Survivor's benefits | 5 | |
| Disability benefits | 9 | |
| Education related allowances | 5 | 3 |
| Income from rental of a property or land | 5 | 2 |
| Interests, dividends, profit from capital investments in unincorporated business | 27 | 12 |
| Family/children related allowances | 37 | 33 |
| Social classifications not else where classified | 14 | 11 |

Sources: EU-SILC 2005, HBS 2003-2005

The reference income period for both surveys is 2004. We must certainly point out that some categories/variables are defined differently in these surveys. In HBS we can not differ between various kinds of pensions, so we can compare only the share of persons who received it directly. We must take into account that some persons can get two kinds of pensions, so such numbers always have to be checked (or treated with some reservation). The data about non-cash employee income are calculated in EU-SILC on personal level, but in HBS on household level. It is obviously that answers from interviewing are in HBS more biased by forgetfulness of

respondents and that means that the quality of information from the registers (EU-SILC) is much higher.

Then we compared national aggregates of incomes from different sources:

Table 4: Aggregates of some kinds of income in billion SIT, Slovenia, reference income year 2004

| Kind of income | EU-SILC | HBS |
|---|---|---|
| Employee cash or near cash income + sickness benefit (net) | 1,779 | 1,611 |
| Cash benefits or losses from self-employment (net) | 133 | 130 |
| Unemployment benefits (net) | 13 | 18 |
| Old age benefits (net) | 496 | Pension benefits total: 669 |
| Survivor's benefits (net) | 91 | |
| Disability benefits (net) | 165 | |
| Education related allowances (net) | 26 | 22 |
| Income from rental of a property or land (net) | 9 | 13 |
| Interests, dividends, profit from capital investments in unincorporated business (net) | 13 | 4 |
| Family/children related allowances (net) | 86 | 73 |
| Social classifications not else where classified (net) | 35 | 25 |

Because the ranking was done according to the structure of incomes of the population from administrative sources, it is not correct to check the data from the same sources in comparison to survey data. The data were checked also with the administrative sources before the ranking was done. We found out that the data from EU-SILC were comparable to the data from the registers. Because of this, we could do the final check only with the data from HBS, where the ranking according to the structure of incomes was not done. Even before the raking was done we found out that income variables are better covered in the EU-SILC than in the HBS. In the future we will have to do different benchmarking. We will confront the HBS data with the data from the administrative sources and after that we will be able to make a real comparison of the two different sources.

Before we had published any data from EU-SILC, we did several checks of the data and calculated the social cohesion indicators from EU-SILC several times. Actually we calculated the indicators according to different versions of the data. For some incomes we namely had different sources and we had to analyse which set of the data is of higher quality for further use. In this process we found out also that by composing the database some mistakes were done. After we had corrected all the technical mistakes, we compared the social cohesion indicators from EU-SILC and those from HBS. We found out that some differences existed, but they were not significant. Also in this comparison it is important to be aware of all the methodological

differences between EU-SILC and HBS. The final comparison of some results deriving from EU-SILC and HBS is included in this paper in table 5.

Table 5: Basic social cohesion indicators from EU-SILC and HBS, Slovenia, income reference income year 2004

| | **EU-SILC** | | **HBS** | |
|---|---|---|---|---|
| | Income in cash | Income in cash + in kind | Income in cash | Income in cash + in kind |
| At risk of poverty rate (%) | 12.1 | 11.4 | 11.8 | 10.4 |
| At risk of poverty threshold (EUR*) | 5,278 | 5,516 | 4,615 | 4,961 |
| At risk of poverty threshold (SIT) | 1,261,821 | 1,318,908 | 1,103,450 | 1,186,065 |
| At risk of poverty threshold for a household consisting of two adults and two children (EUR*) | 11,083 | 11.585 | 9,692 | 10,418 |
| At risk of poverty threshold for a household consisting of two adults and two children (SIT) | 2,649,825 | 2.769.708 | 2,317,245 | 2,490,736 |
| At risk of poverty rate before social transfers (except old-age and survivor's pensions) (%) | 25.8 | 24.8 | 19.4 | 17.2 |
| At risk of poverty rate before all social transfers (%) | 42.2 | 40.9 | 40.6 | 37.4 |
| Inequality of income distribution: S80/S20 quintile share ratio | 3.4 | 3.3 | 3.4 | 3.2 |
| Inequality of income distribution: Gini coefficient (%) | 23.8 | 23 | 24.1 | 22.4 |

EUR rate: Eurostat, New Cronos Database.

## *Chapter 2.4 New project – Social Statistics Database (SSD)*

On the basis of experience of using administrative sources for the purpose of EU-SILC, at SORS a new project was launched in 2006 in order to solve the problems with the registers and administrative sources. The SSD is composed of 4 modules[3)]:

1. Register of persons. This module includes the majority of the basic data about persons living in the country.
2. The main task of the input database module is preparing the data from different administrative and register sources for using them for different purposes. In this module PINs will be changed into statistical personal identificators and all the data will be loaded into the database.
3. In the module for data integration and statistical processing every project – statistical survey takes the data from different sources and the database for each individual project shall be built. This module includes also possibility to take the data from the database for the purpose of other data processing (editing, imputations, weighting) and after such a process the data will be returned into the database.
4. The output of the analytical module are final data prepared for publishing via different media (internet, classical lists and tables on the paper, etc).

The aim of the SSD is to prepare the architecture and system for merging different data sources so as to easily and smoothly implement data processing; and of course to develop some analytical tools.

# Chapter 3. Conclusions

In spite of having used administrative sources and having combined them with the statistical sources in SORS for several years, EU-SILC is the first sample survey where we used the registers and other administrative sources in such a large dimension. We found out that the quality of data from administrative sources was better in comparison to the data collected only by questionnaires.

The time of interviewing was shorter, we skipped many difficult questions and thus the burden on the respondents was not so heavy. This is of high importance due to the fact that household is interviewed several times (4 waves).

We also found out that registers and administrative data are not quite complete. Some persons could not be found in any register. In such cases we had some problems, because we had to impute the data for these persons. Fortunately, the share of the persons “without« register data was quite low.

Another problem is timeliness, because we do not get immediately all the administrative sources and much more time is spent to combine all the data. In the first year of the survey we were late with the preparation of the data, causing a delay of 2 months.

Having compared the data among different sources and surveys, we found out that similar results were obtained.

Because SORS is register-oriented, we can expect that in the future our office will collect even more data from the registers and administrative sources. In this way we shall diminish the response burden of the households and at the same time we will get more reliable data.

### *References:*


1. Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC)
2. National Statistics Act (Official Journal of the Republic of Slovenia 45/1995)
3. Setting up the Social Statistics Database for the Implementation of EU SILC - inception report (S&T)