EU-SILC provisional results available two months after data collection. The case of INE-Spain

Antonio ARGÜESO

Director of Social and Demographic Statistics. INE-Spain

Abstract: The current economic crisis is having an important impact on social surveys. There is a pressure on statistics to provide updated information to monitor the extent of the crisis in the social field. It is the case of EU-SILC, the main source of comparable information on income and living conditions across Europe. Official statistics must provide solid data, and it takes time, especially when microdata eventually become public for researchers. But timeliness is also one of the most important components of quality. The information demanded for decision making often doesn't need to offer the maximum level of detail, nor even to be completely accurate but to give a timely and reliable picture of the situation. In Spain it takes around 16 months to release EU-SILC after the fieldwork period (from June year-t to October year-t+1). Now, what happens if we apply just some automatic treatment to data collected and calculate the main results? We can compare these results, which could have been available two months after data collection, with those actually published. This study has been made in INE-Spain using files from recent years and the conclusion has led to the decision of publishing provisional data for EU-SILC 2010 before the end of this year.

Key words and phrases: Survey on Income and Living Conditions, provisional indicators.

1. Introduction

The Survey on Income and Living Conditions (SILC) in Spain is an annual household survey carried out by the National Statistics Institute (INE). The general objective of this survey is the production of statistics on the level and distribution of income, and the living conditions. This statistical operation, which has been harmonized for EU countries by a Community Regulation, provides comparable data regarding the level and composition of poverty and social exclusion.

This regulation specifies the deadlines for data availability. Regarding the cross-sectional component, the deadline for the transmission of micro-data to Eurostat is 30 November (year N+1) for Member States where data are collected at the end of year N (or through a continuous survey or

through registers) and 1 October (N+1) for the rest. Together with the micro-data files, Member States transmit the harmonized European Social Cohesion Indicators.

From the data collection to the delivery of the final micro-data, Member States carry out the data checking, cleaning and editing, the imputation of data in relation to income and the weighting of the data files. After all this process good quality data and indicators are produced.

But timeliness is also one of the most important components of quality, being today a serious limitation for EU-SILC. The need for timely data is also reinforced by the current economic crisis. Policy-makers need information to be able to design effective responses and often the information demanded for decision making doesn't need to have the maximum level of detail, nor even to be completely accurate but it has to give a timely and reliable picture of the situation.

In this paper we calculate some provisional indicators using the original data collected in fieldwork and compare them with those obtained after the whole cleaning process, i.e., one year later. To build these provisional figures only some automatic treatments are applied to the original data. This study has been made in INE-Spain using files from years 2007 and 2008. The conclusion was definite: it was decided to publish provisional figures for EU-SILC 2009 in March 2010 (six months in advance) and the release of provisional figures for 2010 is already scheduled for November 2010, 4 months after data collection is over (and one year in advance from original schedule). Additional time-saving operations can be incorporated so that provisional data availability two months after data collection (in September) could be reached in 2011.

2. From data collection to final results. Current situation

The Spanish SILC is conducted every year during the period April 1st – June 30th. The data management in EU-SILC is quite complex. There is a panel component that makes the operation more difficult, due to the longitudinal consistency that is required. Another difficult task is the collection of income data through personal interview. This type of variables usually must be checked and imputation is needed in case of partial non-response to calculate the total disposable household income.

The current data process has the following main steps:

- Data collection (CAPI survey). Due to the use of laptops, automatic controls can be included at this stage to control the follow-up of units, the household composition and the basic demographic information (structural consistency) and the other variables of the survey (variables consistency).

- Centralized structural data checking. After fieldwork the data are checked to ensure the structural consistency. The follow-up of households and persons is checked carefully. In this stage, normally the analysis of each case is quite time consuming although, after the inclusion of automatic controls during fieldwork, there are few cases. After this step good quality weights can be obtained.
- Centralized general data checking. Income and non-income variables are checked. This phase normally takes a long time. There are many cases to check.
- Imputation of income variables. After the cleaning of the data, income variables are automatically imputed in case of partial non-response.
- Weighting. Final weights are calculated.
- Production of the Eurostat files and calculation of the indicators.

This process takes a long time, several months, affecting the timeliness of the operation. The wave of 2007 was published in November 2008 (16 months after the end of data collection) and the one of 2008 in October (15 months of delay).

3. Comparing provisional and definitive results

We can compare the data already published (*final*) with those indicators that we would have been obtained just after data collection applying only automatic treatments (*provisional*) and assess the quality of the provisional indicators. The test has been made with data of the years 2007 and 2008. The conclusions are similar in both cases.

Concerning the indicators based on non-monetary variables (subjective poverty, material deprivation) the final and provisional results are almost identical. These variables are seldom corrected in the checking phase and the differences are mainly explained by the use of provisional weights. In the case of the subjective variable "ability to make ends meet":

Households with difficulties to make ends meet *(%)*

(With great difficulty, with difficulty or with some difficulty)

	2007			2008		
	Provisional	final	Deviation (prov-final) / final	Provisional	final	Deviation (prov-final) / final
Spain	**56,7**	**56,8**	**0,18%**	**60,0**	**60,0**	**0,00%**
Regional data						
Andalucía[(1)]	**66,7**	**66,7**	**0,00%**	**70,0**	**69,9**	**0,14%**
La Rioja [(2)]	**39,6**	**39,6**	**0,00%**	**54,8**	**54,7**	**0,18%**

(1) The largest region of Spain. Population: 8 million inhab. approx.

(2) The smallest region of Spain. Population: 0.3 millions inhab. approx.

The adjustment in the case of deprivation variables is also important. For the variable "capacity to face unexpected financial expenses" the results are:

Households without capacity to face unexpected financial expenses (%)

	2007			2008		
	Provisional	final	Deviation (prov-final) / final	Provisional	final	Deviation (prov-final) / final
Spain	**30,4**	**30,5**	**0,33%**	**28,0**	**28,1**	**0,36%**
Andalucía	**44,8**	**44,8**	**0,00%**	**39,1**	**39,1**	**0,00%**
La Rioja	**22,8**	**22,8**	**0,00%**	**18,4**	**18,6**	**1,09%**

In relation to the indicators based on monetary variables the differences are higher. First, we will focus on the poverty rates. To assess these differences we must take into account the level of error. For example, for the total poverty rate, the 95% confidence interval error is about ± 1. Although these indicators are published by Eurostat rounding to integer values, we will use one decimal (as published in Spain) to facilitate the comparisons.

At-risk-of-poverty rate by regions (%)

	2007			2008		
	Provisional	final	Deviation (prov final) / final	Provisional	final	Deviation (prov-final) / final
Spain	**19,7**	**19,7**	**0,00%**	**19,5**	**19,6**	**0,51%**
Andalucía	**27,2**	**29,2**	**7,35%**	**28,9**	**28,9**	**0,00%**
La Rioja	**21,9**	**19,4**	**11,42%**	**19,9**	**19,3**	**3,02%**

At-risk-of-poverty rate by sex and age group (%)

	2007			2008		
	Provisional	final	Deviation (prov final) / final	Provisional	final	Deviation (prov final) / final
Both sexes: total	**19,7**	**19,7**	**0,00%**	**19,5**	**19,6**	**0,51%**
Less than 16	**23,4**	**23,4**	**0,00%**	**23,3**	**24,0**	**3,00%**
65 and over	**27,7**	**28,5**	**2,89%**	**27,7**	**27,6**	**0,36%**
16 and over	**19,0**	**19,1**	**0,53%**	**18,8**	**18,8**	**0,00%**
16 to 64	**16,9**	**16,8**	**0,59%**	**16,7**	**16,7**	**0,00%**
less than 65	**18,1**	**18,0**	**0,55%**	**17,9**	**18,1**	**1,12%**
Men: Total	**18,3**	**18,6**	**1,64%**	**18,1**	**18,3**	**1,10%**
Less than 16	**22,9**	**23,5**	**2,62%**	**22,1**	**23,2**	**4,98%**
65 and over	**24,8**	**26,1**	**5,24%**	**24,8**	**25,0**	**0,81%**
16 and over	**17,5**	**17,6**	**0,57%**	**17,4**	**17,4**	**0,00%**
16 to 64	**16,0**	**15,9**	**0,62%**	**15,9**	**15,8**	**0,63%**
less than 65	**17,3**	**17,3**	**0,00%**	**17,0**	**17,2**	**1,18%**
Women: Total	**20,9**	**20,9**	**0,00%**	**20,9**	**21,0**	**0,48%**
Less than 16	**23,9**	**23,2**	**2,93%**	**24,5**	**24,9**	**1,63%**
65 and over	**29,8**	**30,2**	**1,34%**	**29,9**	**29,5**	**1,34%**
16 and over	**20,4**	**20,5**	**0,49%**	**20,2**	**20,3**	**0,50%**
16 to 64	**17,8**	**17,8**	**0,00%**	**17,5**	**17,7**	**1,14%**
less than 65	**18,9**	**18,8**	**0,53%**	**18,8**	**19,0**	**1,06%**

differences higher than 3% are coloured in red

We can see that the differences are not very significant in general by sex and age groups though in the case of regions (autonomous communities) are higher. We also have to admit that these indicators are quite structural and don't change dramatically across the years.

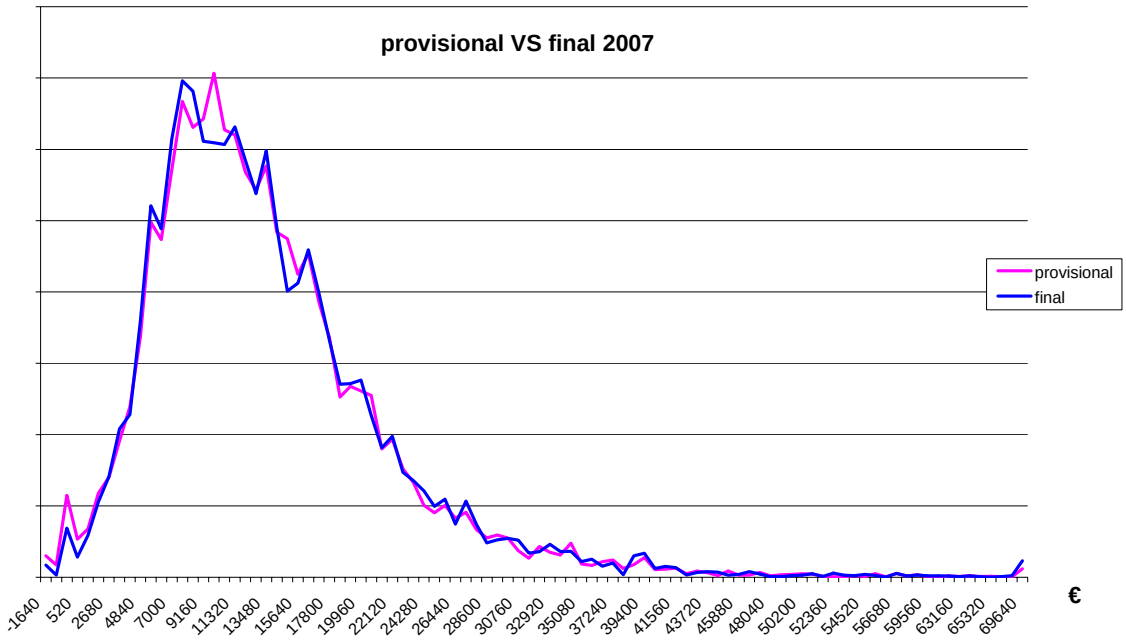
In the next table we compare an indicator based on the level of income, the average annual net income per person. In the same way, there aren't important differences:

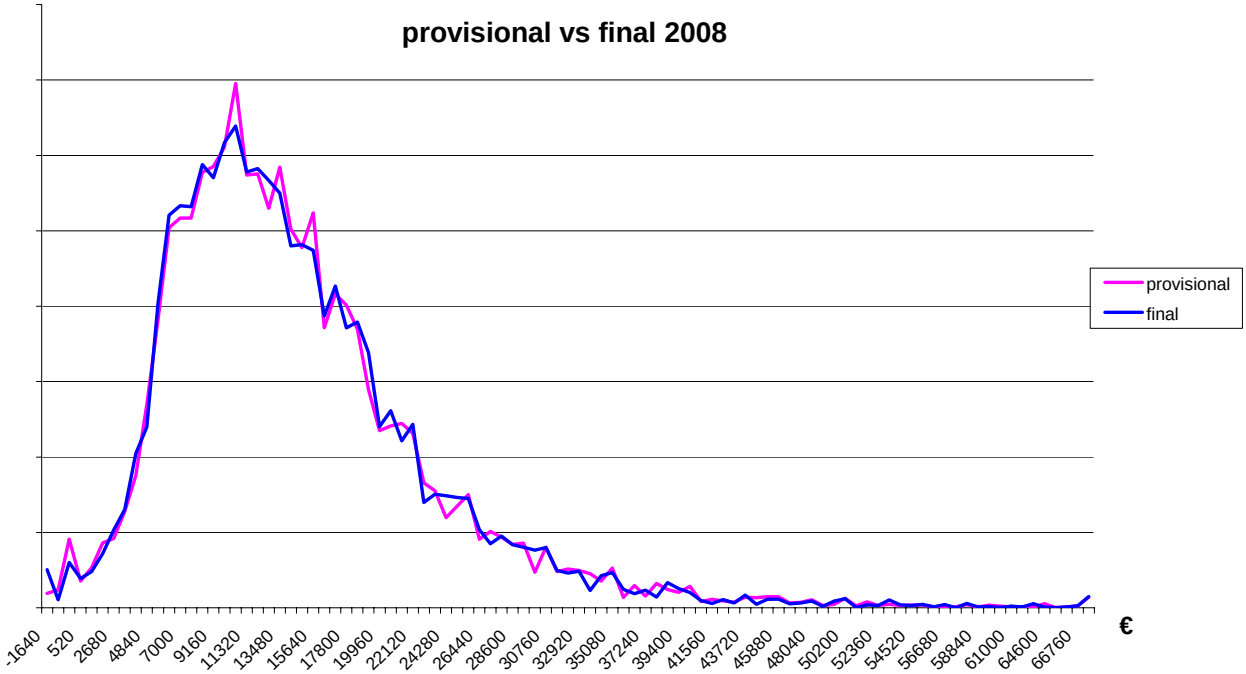
Average income per capita in € (selected regions)

	2007			2008		
	Provisional	final	Deviation (prov-final) / final	Provisional	final	Deviation (prov-final) / final
SPAIN	**8.766**	**8.916**	**1,71%**	**9.605**	**9.560**	**0,47%**
Andalucía	**7.180**	**7.231**	**0,71%**	**7.870**	**7.743**	**1,61%**
Cataluña	**9.857**	**10.107**	**2,54%**	**10.774**	**10.755**	**0,18%**
Comunidad Valenciana	**8.767**	**8.827**	**0,68%**	**9.497**	**9.423**	**0,78%**
Extremadura	**6.730**	**6.668**	**0,92%**	**7.068**	**7.008**	**0,85%**
Galicia	**8.097**	**8.186**	**1,10%**	**8.652**	**8.711**	**0,68%**
Madrid	**10.334**	**10.726**	**3,79%**	**11.287**	**11.431**	**1,28%**
Navarra	**11.988**	**11.871**	**0,98%**	**12.254**	**12.079**	**1,43%**
País Vasco	**10.159**	**10.493**	**3,29%**	**11.609**	**11.526**	**0,71%**
La Rioja	**8.765**	**8.911**	**1,67%**	**9.522**	**9.493**	**0,30%**

differences higher than 3% are coloured in red

The following graphs also illustrate the very similar distribution of provisional and final data for both years.





4. Provisional indicators issued[1]

After a detailed study of the main variables it was eventually decided to select for the release of provisional results only those with the smallest potential deviation and only at the national level. The selected indicators are:

1) At risk of poverty rate by age and sex group (national level)
2) Average income per household, person and consumption unit
3) Ability to make ends meet
4) Capacity to afford some aspects of living standards
5) Housing related arrears

In order to arrange an intermediate calendar, the data for 2009 were published in March 2010 and those for 2010 are scheduled for November 2010.

5. Production of provisional indicators. Automatic treatments

As it has been said before, in order to produce the provisional results in the test of the 2007 and 2008 operations only some automatic procedures are applied to raw data. These procedures applied are as follows:

[1] See press release issued on march 17th 2010: http://www.ine.es/en/prensa/np589_en.pdf

5.1 Weighting

The raising factors used for the provisional results are also provisional. It is essential to have a good control of the raw information collected during the interview in fieldwork, in particular for the structural controls. We can achieve a good follow-up of units, households and persons, if we include controls in the data entry program. In the same way it is important to check the basic demographic information (sex and age) using the information available. This will allow obtaining weights of good quality at this stage.

Once the data are received in the Headquarters we can calculate the provisional weights. We have simulated the process with data of the SILC-2008 and the provisional weights obtained are very similar to the final ones. The only distortions were generated due to a few cases of structural errors.

5.2 Income variables

The following step is to check these variables in an automatic way. We will focus only on the income variables because the monetary poverty indicators are based on these variables and also these variables are the most relevant ones in the checking process. On the other hand the values of the subjective or material deprivation variables normally don't change during the cleaning process and are mainly kept as they are collected in fieldwork.

For the income variables some automatic treatments are implemented. These treatments are related to the out-of-range values (to improve the robustness of the indicators) and to the coverage of the income sources (lack of income sources in the original data), relating the detailed activity information to the income variables:

- *Out-of-range controls (income variables*). The limits of the outliers have been narrowed. If an original amount is out of the interval then "missing" is considered and the value of the amount will be imputed afterwards.

- *Employee income*. If the respondent states to be working (as employee) in the activity calendar, but answered 'No' in the filter of the employee income (and also 'No' in the self-employment income), then the filter of the employee income is replaced by 'Yes'. The value of the amount will be imputed afterwards.
- *Self-employment income.* If the respondent states to be working (as self-employment) in the activity calendar, but answered 'No' in the filter of the self-employment income (and also 'No' in the employee income) and there isn't anyone else in the household receiving self-employment income, then the filter of the self-employment income is replaced by 'Yes'. The value of the amount will be imputed afterwards.

- *Negative self-employment income*. If the respondent declares losses from the self-employment but at the same time answered that the household doesn't have difficulties to make ends meet, then the losses are replaced by profits. The value of the amount will be imputed afterwards.

- *Old-age benefits.* If the respondent answered in the activity calendar to be retired, to have worked at least 15 years (minimum legal limit) and is aged 65 or more, but he or she has not any social benefit, then the filter of being a recipient of an old-age benefit is replaced. The value of the amount will be imputed afterwards. Moreover, if the respondent has old-age benefits and the amount is lower than the previous year amount, then this amount is taken, updated according to the average increase of the pensions.

After this automatic checking, the standard imputation procedure is carried out obtaining the values of the income amounts. In the production of income variables, first the income amounts are calculated when there is enough information. Then, for the missing values, the amounts of the previous wave are used when available. Finally if the amount cannot be calculated at this stage then it is imputed. In the Spanish SILC, the statistical imputation software used is IVE-ware. The IVE-ware approach consists of a multivariate model involving a multiple regression sequence. The restriction of an interval can be included in the imputation procedure. After a logarithmic transformation the imputation is carried out jointly with others components collected at the same level (household or individual). All records with missing values, for income components, are imputed.

6. Conclusions and next steps

Although EU-SILC survey is considered somehow "structural", there is a strong need for timely information that is reinforced in a period of economic crisis like the one we are still facing today.

The results of the tests in Spanish have demonstrated that we can obtain very quick provisional results at the national level and for some aggregates (sex and large age groups) with a reasonable margin of error, in particular in relation to subjective and material deprivation variables. The key is to be able to obtain good quality provisional weights and to have procedures to improve automatically the income information. The identification of adequate automatic income treatments that work correctly either in a situation of economic crisis or in a more stable economic environment is still a challenge and it will take some more years to refine the whole process.

7. References

[1] European Parliament and Council of the European Union (2003). Regulation of the European Parliament and of the Council of 16 June 2003 concerning Community Statistics on Income and Living Conditions (EU-SILC) – (EC) NO. 1177/2003. Official Journal of the European Union L (Legislation), Vol. 46, no. 165 (3 July 2003), pp. 1-9.

[2] INE. Living Conditions Survey. Methodology. www.ine.es.

[3] ARGÜESO, Antonio; MENDEZ, José-María and VEGA, Pilar. *EU-SILC provisional results available two months after collection, a dream come true?*
Paper presented at European Conference on Quality in Official Statistics. Helsinki, may 2010. (q2010.stat.fi).