Berger, Melissa; Schaffner, Sandra

Working Paper

A note on how to realize the full potential of the EU-SILC data

ZEW Discussion Papers, No. 15-005

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Berger, Melissa; Schaffner, Sandra (2015) : A note on how to realize the full potential of the EU-SILC data, ZEW Discussion Papers, No. 15-005, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-387483

This Version is available at:
https://hdl.handle.net/10419/106400

Discussion Paper No. 15-005

A Note on How to Realize the Full Potential of the EU-SILC Data

Melissa Berger and Sandra Schaffner

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 15-005

A Note on How to Realize the Full Potential of the EU-SILC Data

Melissa Berger and Sandra Schaffner

A Note on How to Realize the Full Potential of the EU-SILC Data[1]

Melissa Berger[2], Sandra Schaffner[3]

January 8, 2015

Abstract

The European Union Statistics on Income and Living Conditions (EU-SILC) is a rotational panel provided by Eurostat that covers variables with a high potential for comparative European labour market and social research. Unfortunately, its current availability limits its potential research applications. This research note describes these shortcomings of the current data provision. Furthermore, we make two contributions for a better exploitation of these data sets: First, we develop a method for combining the different waves in order to increase the number of usable observations; and second, we indicate how monthly data on income and hourly pay can be derived.

JEL Classification: C81; C83; D31

Keywords: EU-SILC, sampling weights, income, Europe, data quality, panel data

[1] *We are grateful to Alfredo Paloyo, Benjamin Bittschi and Friedrich Heinemann for their useful comments.*
[2] *Corresponding Author: ZEW Mannheim, L7,1, 68161 Mannheim, Tel: +49 (0)621 1235-169, Fax: +49 (0)621 1235-215; melissa.berger@zew.de*
[3] *RWI Essen; Schaffner@rwi-essen.de.*

1 Introduction

The European Union today faces serious challenges with respect to labour market performance, the quality of education systems, distributive outcomes and issues of social exclusion. An indispensable precondition for both any informed debate on these issues and substantive research are reliable and consistent databases which allow for meaningful cross-national comparisons. In particular, macroeconomic indicators are insufficient for a comprehensive analytical approach. For an in-depth scrutiny of the issues mentioned, reliable micro-data on wages, income and individual conditions in general are an absolute necessity.

Existing national data sets, like for example the German Socioeconomic Panel, with their broad coverage have the advantage of a rich set of variables and a rather large number of observations. In addition, their structure is well adapted to the idiosyncratic settings of the respective country. However, these national data bases hardly offer a high potential with respect to their comparability between countries and cross-country analyses. By contrast, the European Union Statistics on Income and Living Conditions (EU-SILC) aim to be comparable for all EU Member States while maintaining high-quality standards, featuring data accuracy, precision, timeliness, clarity and comparability between subgroups/regions. The data set is the successor of the European Community Household Panel (ECHP), which ran from 1994 to 2001 covering similar topics and countries as the EU-SILC does.

Although the EU-SILC data cover a wide variety of subjects, the number of existing studies using these data is quite rare. So far, the main applications relate to topics like poverty (e.g., Longford et al. 2012 and Whelan and Maitre, 2012), inequality (e.g. Giannetti et al, 2009), housing quality (Angel and Bittschi, 2014) and wage mobility (e.g. Aristei and Perugini, 2012 and Bachmann et al. 2014). This limited use might be due to the fact that there are still some shortcomings of the data set that discourage its extensive use in empirical social science research.

A number of papers on the quality of the EU-SILC data already exist. In these studies, the authors recommend strategies to improve the data design or to exploit the full potential of the existing data. Iacovou et al. (2012) give a comprehensive overview of strengths and weaknesses of the EU-SILC data regarding sampling and design, household dynamics, and incomes. Based on their findings, they recommend several changes regarding data collection and data provision. Frick and Krell (2011) show that there are differences in the measured inequality and poverty for Germany compared to values derived from the well-established German Socioeconomic Panel. Goedéme (2010) presents the necessary sample design to estimate reliable standard errors when using EU-SILC data.

As an extension to the existing papers on data quality, our study gives a brief overview on data problems and the quality of the EU-SILC data with a special focus on income and the rotational panel design. Compared to the European Labour Force Survey, EU-SILC's longitudinal structure and its information on income are a clear strength. However, the data sets provided by Eurostat do not cover all waves of the rotational panel in one data set. Each year, a different bundle of rotational groups is merged into one data set. Thus, the number of observations and years is smaller than it could to be. However, to efficiently estimate parameters in multivariate analyses, a large number of observations is necessary. Therefore, researchers are interested in capturing all information that is available in one data set.

We propose a strategy to limit this waste of information: The number of observations can be increased by first merging different data sets and then reweighting the observations. We describe this procedure for the years 2004 to 2011 and show that, consequently, we are able to use almost all available observations. Apart from the reduction of observations in the data sets delivered by Eurostat, the EU-SILC data suffers from the shortcoming that income information is only available on a yearly basis and that labour market status and additional variables are not captured for the same time period. This limits the possibilities to use the data set for analyses of the European labour markets. We present a strategy on how to

calculate monthly income as well as hourly wages based on the yearly income measure provided in the data. These measures correspond to the same observation period as the additional information on labour market status in the yearly interview. Based on our strategy, it is possible to use the data for a multitude of labour market studies.

The remainder is structured as follows: section 2 gives a brief overview of the EU-SILC data set and its characteristics. The merging of different waves of data is described in section 3. In section 4, we show how to derive a monthly data set. Section 5 describes the strategy to calculate monthly and hourly pay and, finally, section 6 concludes.

2 The EU-SILC Data

In this section, we describe the data design of the EU-SILC data and its consequences for data preparation. The EU-SILC data is made available through two different types of data sets: cross-sectional and longitudinal micro data sets. Both data sets are collected and published on a yearly basis however the longitudinal files contain more precise information.

Due to the advantage of panel data over cross-sectional data in econometric analyses, we concentrate on using the longitudinal files, containing, up to now, observations for the years 2004-2011. Except for a comparison of labour income the whole analysis concentrates on the longitudinal files. We use the *EUSILC LONGITUDINAL UDB 2011 – version 1 of August 2013 (L2011); EUSILC LONGITUDINAL UDB 2010 – version 3 of August 2013 (L2010); EUSILC LONGITUDINAL UDB 2009 – version-4 of March 2013 (L2009), EUSILC LONGITUDINAL UDB 2008 – version-4 of March 2012 (L2008), EUSILC LONGITUDINAL UDB 2007 – version-5 of August 2011 (L2007), EUSILC LONGITUDINAL UDB 2006 – version-2 of* February 2008 *(L2006) and EUSILC LONGITUDINAL UDB 2005 of* February 2008 *(L2005)*.

The cross-sectional data and the longitudinal data differ to some extent in the variables covered. There are some variables in the cross-sectional data file that are also of interest for the analysis of labour market transitions and mobility, but they are not included in the longitudinal data sets. This concerns the following variables in particular:

- Information on the use of child care (variables RL030-RL070);
- The reason for working less than 30 hours (part-time) (PL120).
- Firm characteristics: number of persons working at the local unit (PL130), industry (PL110)
- Indicators related to immigration, such as the country of birth (PB210) and citizenship (PB220A)
- The gross monthly earnings for employees (PY200G), which are only available for some years and countries

In addition to the problem that important variables of the cross-sectional data files are unavailable in the longitudinal sets, monthly information in the cross-sections is imprecise. The exact month of the interview is not available. Especially in order to generate monthly income and monthly transition rates, it is important to know the precise month of the yearly interviews.

In some countries, only one person, the “selected respondent”, answers the questionnaire for the entire household. This is true in all Scandinavian countries, as well as Ireland, Iceland, the Netherlands and Slovenia. Although most information is available for all household members, some indicators, especially the calendar data which contains the date of the interview, are only available for these selected respondents. Therefore, the number of observations decreases if variables affected by this selection process are used.

Data versions delivered by Eurostat contain the longitudinal files L2005-L2011, each including information for the corresponding year as well as up to the three preceding years.

The first observations are from 2003 and thus the observation period is 2003 to 2011. For most countries, information is available for a shorter period (see Table A1 in the Appendix). Since the 2003 wave was a pilot survey, we exclude it from the analyses. Data for the whole period (2004-2011) is available for Austria, Belgium, Denmark, Estonia, Spain, Greece, Iceland, Italy, Luxembourg, Norway, and Portugal.

For all other countries, six or seven years are available except for Croatia (2010-2011), Germany (2005-2006) [4] and Romania (2007-2011).

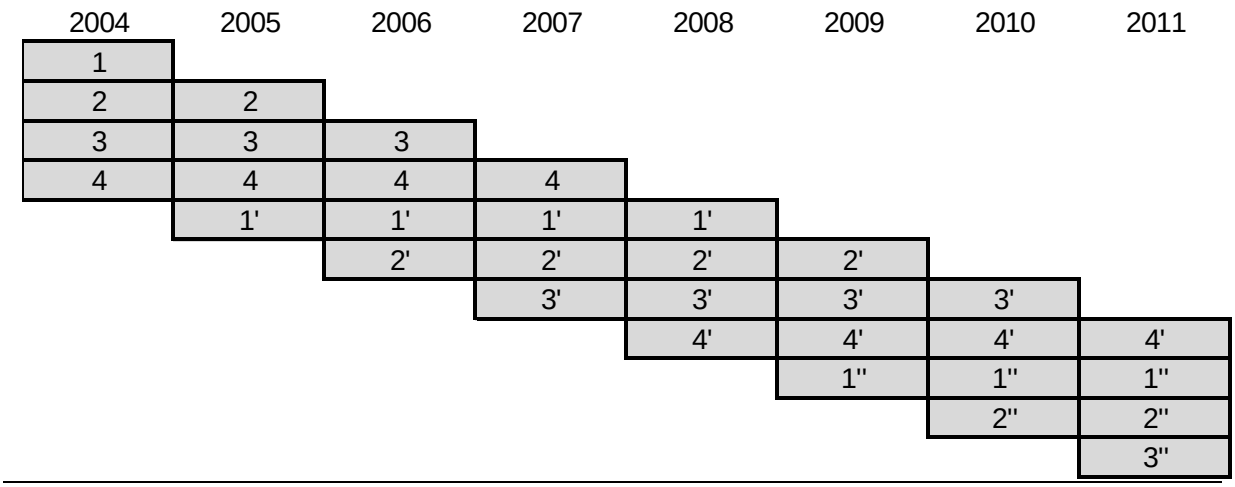
The EU-SILC panel is a rotational panel (except for Luxembourg) which is comparable in its structure to the Current Population Survey (CPS). In a rotational panel, the same persons are interviewed for a certain time period (in this case four years[5]) and each year one quarter of all respondents is replaced by new respondents. The integrated design consists in selecting four panels at the first wave. Each subsequent year, one panel is dropped and replaced by a new group of respondents. This enables us to follow persons over two, three or four consecutive years. From the fourth wave on, all respondents can be observed for four years. Therefore, each person is interviewed up to four times (if they do not refuse to participate), while the number of persons stays almost stable over all periods.

Figure 1 shows the panel structure of the EU-SILC data for a country that first starts in 2004. Of the individuals interviewed in 2004, three quarters are also interviewed in 2005 while the first group is replaced by a new subsample (1'). In the following year another quarter of individuals (group 2) are replaced by a new group (2'), and so on. Therefore, in 2007 only 25 per cent of the original sample interviewed in 2004 is still being interviewed. This fraction decreases to zero in the 2008 wave and its consecutive waves. Group 4 is the first group that is interviewed over a four year period. Therefore, for countries with data availability from 2004 to 2011, five rotational groups (group 4, group 1', group 2', group 3' and group 4') are

[4] *For disclosure control reasons, the data set does not include longitudinal 2007 and 2008 data for Germany.*
[5] *Exceptions are France with 9 years and Norway with 8 years as well as Luxembourg without any rotational scheme.*

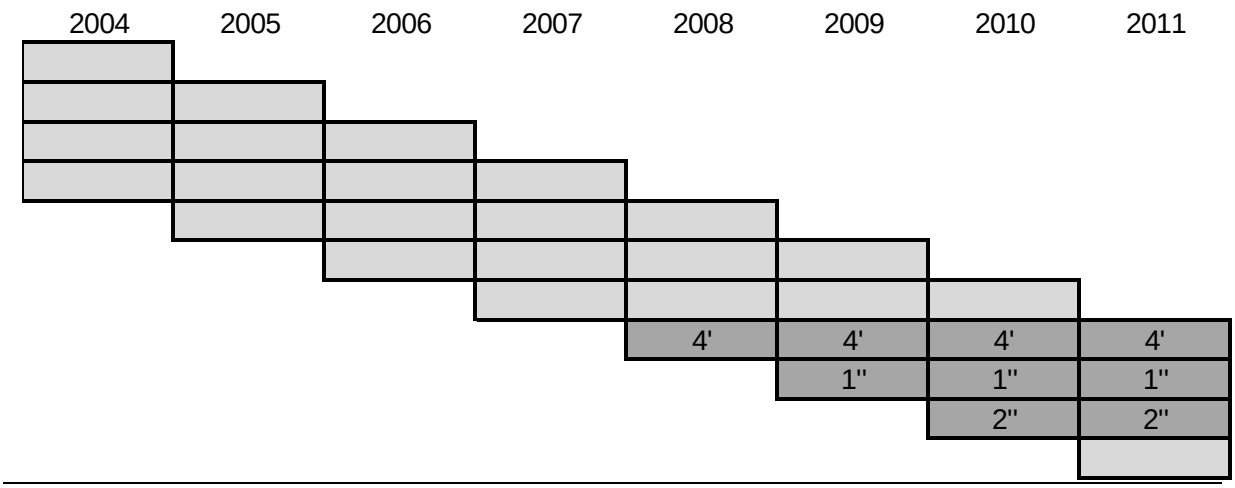
interviewed four times in the whole period. However, the data sets distributed by Eurostat do not cover all of these rotational groups in one data set.

Figure 1: The integrated design of EU-SILC



For a given year, the respective longitudinal file available from Eurostat (e.g. L2011) only contains those respondents that were interviewed both in the respective year and in the preceding year. This means that in the 2011 longitudinal wave (L2011), information is only included for those individuals who were interviewed at least in 2011 and 2010. Individuals, who were interviewed in 2004, 2005, 2006, 2007, 2008, 2009 and/or 2010 but not in 2011, are not included in the 2011 longitudinal wave. Figure 2 illustrates the panel groups that are included in the 2011 longitudinal file (dark grey). This figure shows that only 25 per cent of all interviews conducted in 2008 are reported in the 2011 longitudinal file, and there are no observations for 2004, 2005, 2006 and 2007 at all. Of all interviews in 2009 only one half is reported and of the 2010 and 2011 observations three quarters are implemented. Therefore, this way of constructing the longitudinal data set leads to an important loss of observations. As a consequence, the number of observations becomes relatively small. This aspect is of particular importance when analysing small countries, where the original sample is small to start with and it shortens the possibilities to analyse the development over time (implying that there are no other events influencing the variables of interest).

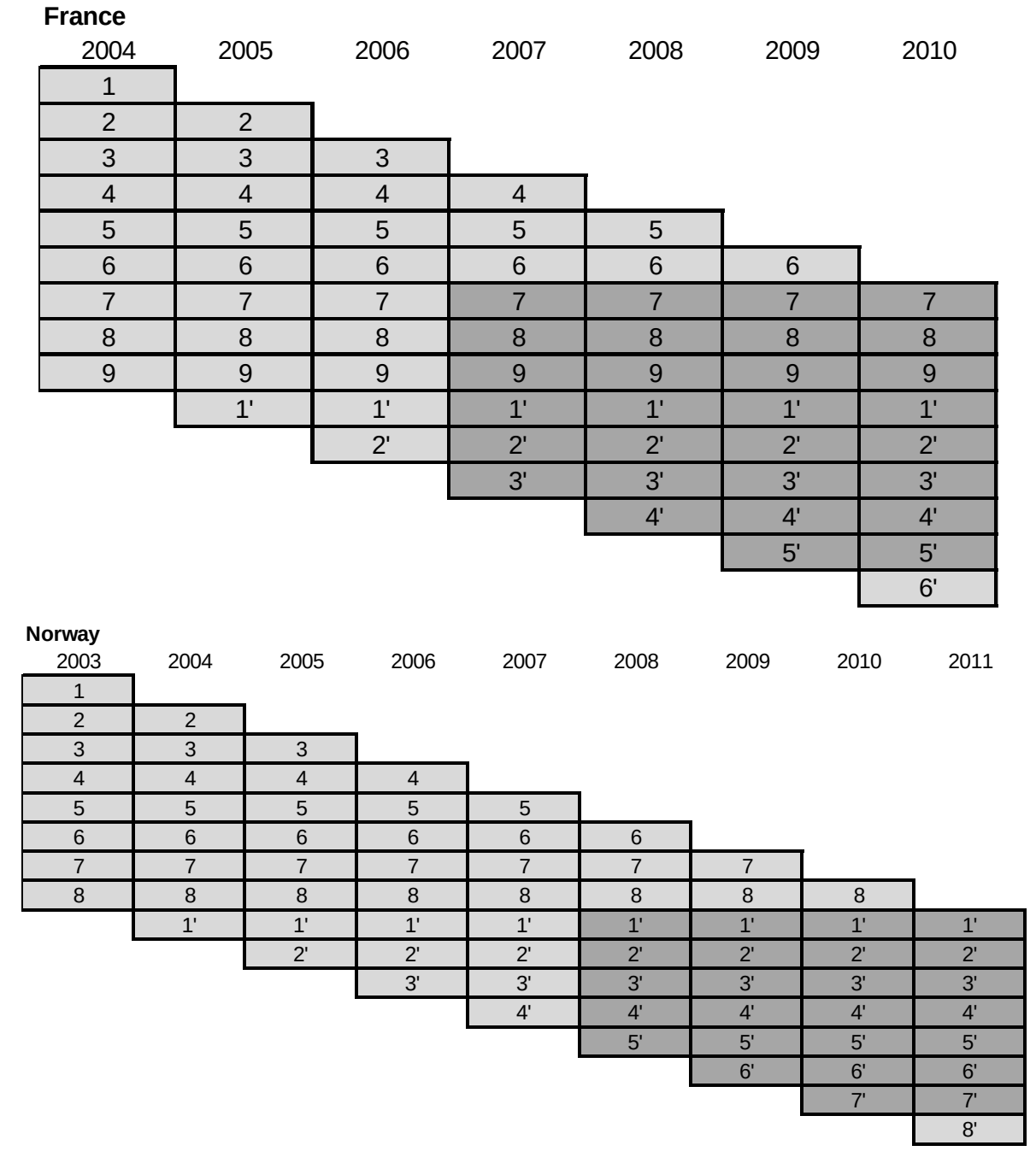
Figure 2: Structure of the 2011 longitudinal data file (L2011)



For France, Norway and Luxembourg, the panel structure is different to the one described above. In contrast to the standard structure with four rotational groups, France and Norway chose to use nine and eight groups respectively. As illustrated in Figure 3, the panel of France includes nine groups in each year to cover about 95,000 observations. As in the standard panel, each group is replaced by a new one in the following year which leads to the result that always 1/9 of all observations is replaced. The same is true for Norway where the panel consists of eight rotational groups. Each year 1/8 of observations is substituted with new persons. While the French L2010 version uses the groups 7 to 5' for the years 2007 to 2010, the panel of Norway is smaller: the groups 1' to 7' are reported. For France, 66 to 89 per cent of all observations for the years 2007 to 2010 are in the most current data file. For Norway, only 62.5 to 88 per cent are reported. Overall, it becomes apparent that the reported share is higher than in countries with four groups only. Summed up, except for Luxembourg, a loss of observations not only in the first but also the most recent years can be observed. Of all observations available, only 10 to 60 per cent are included in the L2011 file.[6]

[6] *Without the first rotation group that is not available as they are interviewed only once.*

Figure 3: Panel Structure in France and Norway



3 Construction of a "full" data set

As described in the previous section, not all available observations are included in the data files that are distributed by Eurostat. However, it can be of interest for researchers to increase the overall number of observations and the number of observations that cover four periods in particular. To construct a data set with as many observations as possible, we combine the longitudinal files for 2005, 2006, 2007, 2008, 2009, 2010 and 2011 (L2005-L2011). Each of them can be derived from Eurostat. Due to the integrated design most observations are reported in several longitudinal files. For those observations that are included in several longitudinal files, we keep the observation of the most recent panel version. For Denmark, Lithuania, Portugal and Slovakia some special features have to be considered when merging the different files. In Lithuania and Slovakia the same IDs are assigned to different individuals in different waves. This is true for rotation group 2 that starts in 2007 and is first published in L2008, rotation group 3 that starts in 2008, rotation group 4 that starts in 2009 and rotation group 1 that starts in 2010.

In L2010 and L2011 some zeros in the IDs for Denmark are deleted in comparison to the preceding waves.[7] In the Portuguese data new IDs are generated for each wave from L2008 on. Therefore, it is not possible to combine the different observations. We therefore take the whole L2011, combine it with rotation group 3 of L2010, rotation group 2 of L2009 and rotation group 1 of L2008.

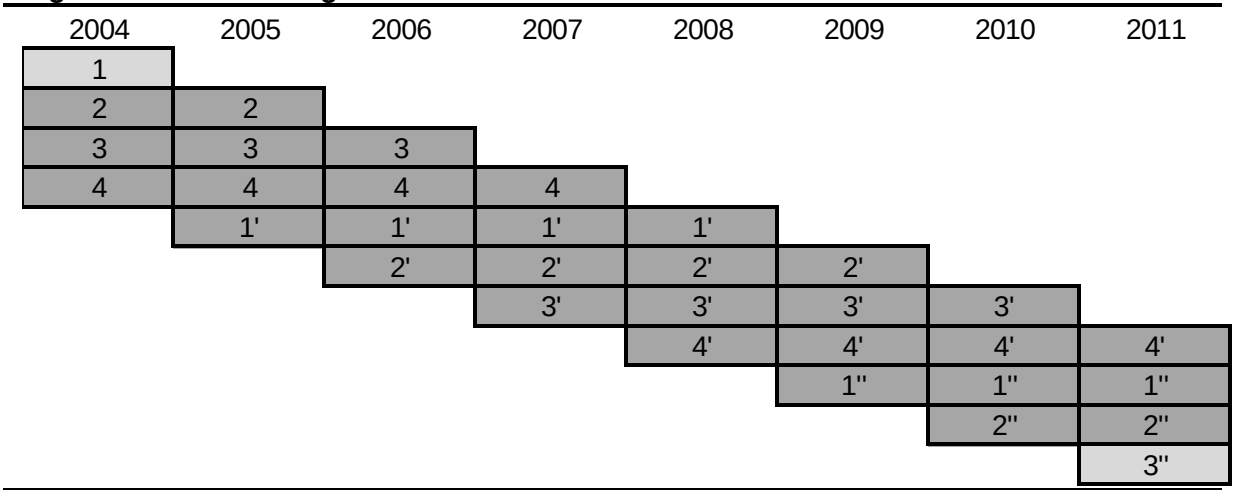
Figure 4 presents the composition of our resulting data set for the countries that are observed for the entire time period 2004-2011. It appears that all observations are included except for the observations of group 1 and the 2011 observations of group 3''.[8] Therefore, in 2004 and 2011 three quarters of all observations are included in our resulting data set, while all of the observations for the years 2005 to 2010 are included. More generally,

[7] *The person ID can be adjusted to the L2009 version by: id_p =(id_p-id_p2)*100+id_p2*

[8] *The reason for not including these two groups is that we only observe them for one year, which means that we cannot use them for many analyses concerning labour market transitions.*

independently of the respective observations period, only one quarter of observations of the first year and of the last year are missing when using our proposed procedure of constructing the data set. This is also true if additional panel waves (L2012, L2013, ...) are added. However, with L2005-L2011 being available, a full sample of all observations can be reconstructed for the years 2005 to 2010.

Figure 4: The resulting estimation data set



Finally, when combining L2005 – L2011, we have about 3.7 million observations in the data set. It can be seen in the overall distribution of observations that in the first and in the last year during which we observe a country, the smallest number of observations is recorded (see Table A2 in the Appendix). Therefore, more than 520,000 individuals are observed in the years 2006 to 2010, whereas we only observe roughly 231,000 in 2004, 477,000 in 2005 and 328,000 in 2009.

As described in the previous section, in some countries only the “selected respondent” answers all questions. For these countries and the respective variables the number of observations decreases. Especially for Iceland and Denmark (see Table 1), the number of observations becomes very small.

Table 1: Number of "selected respondents"
2004 to 2011

Country	2004	2005	2006	2007	2008	2009	2010	2011
Denmark	2,775	3,778	4,090	3,962	4,209	4,549	4,483	3,107
Finland	5,873	7,344	6,975	6,740	6,572	6,317	7,852	5,738
Iceland	2,024	2,646	2,476	2,424	2,462	2,451	2,573	1,766
Netherlands	-	8,399	8,986	10,219	10,337	9,728	10,134	6,772
Norway	5,148	5,799	5,575	5,795	5,308	5,138	4,848	3,613
Sweden	5,404	7,287	6,076	6,271	6,337	6,424	4,500	0,0
Slovenia	-	8,287	9,462	8,635	8,999	9,279	9,364	6,195

Source: EU-SILC, own calculations.

In survey data, weights are used if the survey is not representative for the total population. Weights inform about the number of individuals of the total population which are represented by a single individual. Therefore, those groups that are underrepresented in the data have a higher weight since as they represent more people in the whole population. To account for this new data structure, the weights delivered by Eurostat have to be adopted as they are made for the design with fewer observations. Therefore, it is necessary to weigh the individual years in the longitudinal version differently. The aim is to design the weights in such a way that the number of observations represents the whole population.

In the data set reported, i.e. the integrated design, the weights are, for example, adjusted to the fact that the number of observations for 2006 and the one for 2009 are different in the L2009 version. In the data provided by Eurostat, longitudinal weights and the so-called base weights are reported. "The base weights are the back spine for the computation of both cross-sectional weights and longitudinal weights. They are computed and updated for a single panel…" (Eurostat, 2010, p.35). Longitudinal weights take the time period for which a transition is computed into account. Therefore, the two-year longitudinal weight is necessary for transitions between t-1 and t, while the three-year longitudinal weight is used for transitions from t-2 to t. The weights are only available for observations in t and not for earlier observations.

The procedure to build new weights for our merged data set can be described as follows: The longitudinal weights are taken from the different longitudinal data files (L2005, L2006, L2007, L2008 and L2009) provided by Eurostat. We take the weights of L2011 for the year 2011, the weights of L2010 for the year 2010 and so on. In 2005, the base weights correspond to the two-year longitudinal weights. For those observations that are not included in the respective data file, we take the weights of the subsequent file. Summed up, for one-year and two-year transitions, the two-year and three-year longitudinal weights are taken from Eurostat. Due to the merging process of the data sets, we find more observations than in the original files. Particularly, it has to be taken into account that in the first and last year only three of four rotational groups are included in the data set. Therefore, to adjust the weights in such a way that the whole population of each country is always represented by the observations included in the data, we multiply the weights by 4/3 in the first and last year.

Although using panel data it is also important for researchers to have cross-sectional weights for cross-sectional analyses. To construct cross-sectional weights, we rely on the so-called base weights provided by Eurostat. If available, we take the base weights for 2005 from the 2005 file, the weights for 2006 from the 2006 file and so on. However, one rotational group (see Figure 2) is not included in each of the different longitudinal files. We therefore take the base weight for this group from the subsequent longitudinal file. Furthermore, we have to reweight the first and last year of each country by 4/3 because we only observe three quarters of observations. For most countries, these weights are the cross-sectional weights. However, in some countries the overall sum of the weights in 2004 does not correspond to the number of inhabitants. In these cases, we reweight all weights with the same country-specific factor[9] to derive the population. The factors as well as the country-specific calculation methods are shown in Table A3. It illustrates how to compute proper weights for each year by using weights from previous, current or following years adjusted by certain factors. As can be seen in the table, the same strategy is used for most of the countries and

[9] *The factors are derived from the population numbers provided by Eurostat.*

depends on the year of the first and the last observations. For example for Austria, Finland and Sweden the weights for 2004, as well as for the year 2005, can be taken from the L2005 file. Since they only represent three quarters of the population they are weighted with this factor. The observations that are first illustrated in the L2006 file receive the weight of the L2006 file.

Besides the weighting scheme, the personal identifier is of importance for constructing the data set. The personal identifier (RB030) in the longitudinal files allows for the opportunity to observe one person over several years. After merging the different data, we find changes in the data concerning some persons' gender and/or their date of birth.[10] In our sample of persons aged between 15 and 65, there are changes regarding the gender of 104 persons (Table 2) and the year of birth of 189 persons over time (Table 3). A simultaneous change in age and sex can be observed for 88 individuals (see Table 4). It is possible that some identifiers (IDs) are assigned to different individuals. However, the number is negligible and can also be due to measurement errors as these are survey data. Based on these findings we assume that our way of merging data sets works when the described adaptions are made.

Table 2: Number of persons with changing sex

Country	2004	2005	2006	2007	2008	2009	2010	Total
Belgium	1	0	0	0	0	0	0	1
Cyprus	0	0	0	0	0	1	1	2
Finland	1	0	0	0	0	0	0	1
France	4	2	48	4	1	0	0	59
Greece	0	0	0	2	0	0	0	2
Luxembourg	1	0	0	0	0	0	0	1
Norway	16	12	4	0	2	0	0	34
Romania	0	0	0	3	0	0	0	3
UK	0	0	1	0	0	0	0	1
Total	23	14	53	9	3	1	1	104

Source: EU-SILC, own calculations.

[10] *In the L2005 to L2008 files the age of persons aged 80 and older is censored. Therefore, differences between the different longitudinal versions can occur.*

Table 3: Number of persons with changing year of birth

Country	2004	2005	2006	2007	2008	2009	2010	Total
Cyprus	0	0	0	0	1	1	1	3
Denmark	0	0	0	2	0	0	0	2
Finland	1	0	0	0	0	0	0	1
France	6	4	68	4	2	0	0	84
Greece	0	0	0	2	0	0	0	2
Luxembourg	2	0	0	0	0	0	0	2
Norway	39	23	7	2	4	1	0	76
Romania	0	0	0	15	0	0	0	15
Slovenia	0	0	3	0	0	0	0	3
UK	0	0	1	0	0	0	0	1
Total	48	27	79	25	7	2	1	189

Source: EU-SILC, own calculations.

Table 4: Number of persons changing sex and year of birth

Country	2004	2005	2006	2007	2008	2009	2010	Total
Cyprus	0	0	0	0	0	1	1	2
Finland	1	0	0	0	0	0	0	1
France	0	2	45	4	1	0	0	52
Greece	0	0	0	2	0	0	0	2
Norway	16	9	3	0	0	0	0	28
Romania	0	0	0	3	0	0	0	3
Total	17	11	48	9	1	1	1	88

Source: EU-SILC, own calculations.

4 Construction of a monthly data set

One of the main advantages of the EU-SILC data set is that it covers a set of variables (PL210A-PL210L and PL211A-PL211L, respectively) regarding monthly information and therefore transitions of the preceding year. In this calendar data, respondents declare their main activity in each of the twelve months. In addition to the yearly data which allows us to observe labour market transitions from one year to the next, monthly transitions and employment statuses based on the calendar information can be generated. This variable covers four employment statuses as well as education, retirement, military service and inactivity. Based on this information, it is possible to generate a monthly data set regarding the employment status. The calendar data refers to the income reference period of the respective interview while most of the characteristics of the respondents refer to the date of the interview. The income reference period is defined by Eurostat as follows: “The income

reference period shall be a twelve-month period. This may be a fixed twelve-month period (such as the previous calendar or tax year) or a moving twelve-month period (such as the twelve months preceding the interview).” (Eurostat, 2010).

Except for Ireland and the UK, the reference period is always the preceding calendar year. That means that the calendar data in the 2011 survey cover January to December 2010. These twelve months are those immediately preceding the date of the interview in Ireland and those of the current year in the United Kingdom.

Our aim is to generate a monthly data set to cover labour market dynamics within a year. We expand the yearly data set by twelve (since there are twelve months in a year) and generate a monthly data set for the years 2003 to 2010 and for 2004 to 2011 for the UK, respectively. Unfortunately, we can observe other characteristics (e.g. marital status, health, household size etc.) only for the date of the interview. As a result, there is a time lag between the different kinds of information. In the first year, there only is information available for the next interview. Afterwards the monthly information can be combined with the yearly interview of the same year.

The comparison between the calendar data and the yearly interviews can give some hints about the quality of the retrospective calendar data. As can be seen in Table A4, one year later the majority of individuals report the same labour market status in the retrospective monthly version as during the interview. However, there are some differences. These differences can be a result of recall errors that can also be observed in other data sets (e.g. Jürges, 2007 and Mathiowetz et al., 1988). Furthermore, the different definitions of labour market status in the two questions can lead to differences. In the yearly interview, individuals are asked about their actual labour market status. That might be only one day of unemployment, for example, while it is the main activity of the month in the retrospective data. Therefore, some differences can be expected as can be observed in Bachmann and Schaffner (2009).

Besides the combination with additional variables, it is also necessary to generate longitudinal weights for monthly transitions in this new data set. We generate two-months (from t-1 to t) longitudinal weights only. Longitudinal weights take panel attrition into account. However, between the months of January to December no panel attrition occurs, because the calendar information for one entire year is given retrospectively by the survey respondents. Therefore, panel attrition and the new composition of respondents have to be taken into account only between December and January. This means that cross-sectional weights are sufficient for the transitions between all months with the exception of the transition between December and January. However, cross-sectional weights are not provided in the longitudinal data set. We therefore define the new weights following the base weights of the longitudinal data sets. By applying this method, we aim at reproducing the procedure used by Eurostat. For this approach, rotational structures in the different countries and years have to be taken into account. This procedure is the same as the one described above to generate longitudinal weights in the yearly data set.

5 Calculation of monthly and hourly pay

One important dependent and explanatory variable in labour market analyses is the wage rate. The EU-SILC data, in comparison to other data (ESS, EU-LFS), covers income information which is another advantage of this data set. In this section, we provide a procedure to calculate pay and income variables that correspond to the observable labour market states.

EU-SILC covers information on labour income as well as other sources of income. In the longitudinal files, income gained from employment is covered by the variable “Employee cash or near cash income (gross/net)”. Cash income, non-cash income, unemployment benefits, old-age benefits, sickness benefits, and taxes are also measured. These variables cover the income gained in the income reference period which covers twelve months.

Additionally, information on the current economic situation of the individuals in the data set is available for the time of the interviews; the economic status is also included in the monthly information for the previous calendar year (see previous section). For example, the labour income in the 2011 interview covers the calendar year 2010. Therefore, it is possible that it does not correspond to the current job as described in the 2011 interview. If someone has been interviewed before, the interview of 2010 and the income information of 2011 overlap. However, the problem that the labour market status is only a snap-shot and that income information covers twelve months still exists. Therefore, it seems obvious that the income divided by twelve could be different to the monthly income of the job at the date of the interview of the previous or the current year. Especially for workers with unstable careers (job changes, unemployment interruptions etc.) different time periods can result in large biases.

To derive monthly earnings and benefits or even hourly wages, a strategy for computation is necessary. First, in order to measure labour income, we use the (gross) employee cash income, the calendar data and the number of hours usually worked per week in the main job. Information on the number of hours usually worked and the calendar data are combined in order to compute the number of hours supplied by the worker. Together with the cash income, this is used to calculate monthly income and hourly wages.

As mentioned before, yearly income measures cannot be used as a proxy for monthly income measures, as the yearly income may accrue in only a few months of employment. Therefore, the duration spent in the different statuses during the year has to be taken into account. The retrospective main economic status (calendar data) provides us with some information that can help to divide the income into monthly parts. Furthermore, differences in the income/benefit levels between different employment/unemployment spells have to be considered. However, the calendar data only covers information on the employment status without any additional information (e.g. on direct job changes, occupation, hours worked, wage level etc.).

In the data, only 8 per cent of all individuals who report that they were employed or unemployed during the previous calendar year had at least two different labour market statuses (full-time or part-time employment, or unemployment). Therefore, calculating income and wages should be straightforward for the majority of observations. However, we cannot distinguish between two different full-time (part-time) jobs. Therefore, we only observe a weighted mean of the income in two different jobs, if persons changed their job or experienced a wage increase. For those with only one labour market status during the whole period we apply the first step to derive monthly earnings and benefits:

1. *For those workers who are either full-time employed, part-time employed, self-employed or unemployed in all twelve months, the labour income or the unemployment benefits are divided by 12 to obtain the monthly labour income or unemployment benefits, respectively.*

Additionally to those who have one of the three labour market statuses during the whole year, there are also workers that have only one continuous employment or unemployment spell per year. For example, someone is employed until March, unemployed between April and September and employed afterwards. In this example there is one continuous unemployment spell of five months. For this spell we assume that the monthly unemployment benefits are stable and divide the whole unemployment benefits by five. However, the employment spell is not continuous. If someone is employed in the first half of the year and unemployed in the second half of the year, both, unemployment and employment are continuous and earnings as well as unemployment benefits are divided by six to derive monthly income. This procedure can be described as follows:

2. *For those workers who have only one employment and/or unemployment spell (of several months), labour income/unemployment benefits are divided by the number of months of this spell.*

By now, we only take into account spells that are within one year. However, there can also be continuous spells that are within two different calendar years and for one part of the spell we cannot calculate the respective income with the first two steps. In this case, we extrapolate the income into the next year or the preceding year, respectively. In our first example, it is possible that in the year before, there is an employment spell of 12 months and we can calculate the monthly labour income. We now assume that this income is the same until the end of employment (March). We only adjust it with an inflation indicator.

3. *The derived monthly income is extrapolated to the following months of the next year or to the previous months of the preceding year as long as the labour market status and the full-time/part-time status (in the case of employment) do not change. For example, the income of a worker who is employed full-time in December 2004 is extrapolated to January and February 2005 if the worker is still full-time employed in January and February 2005, but becomes part-time employed, inactive or unemployed in March.*
4. *If there is only one employment spell left in a calendar year with no monthly income derived in step 3, the yearly income is reduced by the income that is assigned to all other employment spells in the respective year (from the extrapolation in step 3) and then divided by the number of months of the remaining employment spell.*

Other benefit variables, such as housing as well as family and children allowances, can play an important role in the income situation of an unemployed or low income person/household. In most of the countries, they are not directly dependent on the employment status but on the income situation and family/household characteristics. We therefore assume these values are uniformly distributed over the year.[11]

In the cross-sectional data provided by Eurostat, monthly labour income information is available for Austria, Bulgaria, Spain, Greece, Hungary, Ireland, Iceland, Italy, Poland,

[11] *This may lead to problems as allowances can only be assigned for a certain time or in a certain situation and these regulations may differ across countries. The data, however, do not allow for distinguishing between different cases.*

Portugal and the UK. However, the IDs are different to the ones in the longitudinal file and it is not possible to directly compare the numbers.

This monthly labour income can be the basis for calculating hourly wages. However, working hours are only measured at the date of the interview. That means that the timing of the information is different to the income period. Therefore, hourly wages can be derived only in the combination of monthly and yearly data at the date of the interview. For those workers with different employment statuses in the calendar data and at the current interview it is not possible. Otherwise hourly wages can be calculated. Due to the different time period, no income information is available for the last interview.

6 Conclusion

Comparative studies are indispensable to contribute to current European policy debates on labour markets and other social issues. Besides the European Labour Force Survey (EU-LFS) and the European Community Household Panel (ECHP), the EU-SILC is an important data set for these analyses. Unfortunately, the data provided by Eurostat are split into different files and this way of provision reduces the number of observations. Furthermore, information on income cannot be related to the economic status as the calendar information lacks preciseness which reduces the value of the data set for labour market analyses.

In this paper, we describe these and other shortcomings of the data set in detail and propose a strategy to increase the number of observations by merging different data sets with appropriate weights. As our description shows, we can increase the number of observations by a large extent, especially regarding those observations which can be observed for four years.

Additionally, we suggest a strategy for deriving data on monthly labour income and benefits received. Based on these calculations, it is possible to relate employment characteristics to

earnings. However, limitations remain and for some workers, we cannot calculate monthly income. This is particularly true for those workers that have relatively unstable labour market histories characterized by job changes and interruptions (e.g., unemployment and inactivity). Therefore, the resulting data are based on a selected sample. Thus, it would be highly desirable that a more comprehensive provision of the EU-SILC data with their high potential for research could be achieved in the future.

Literature

Aristei, D. and C. Perugini , “The Drivers of Income Mobility in Europe”, ECINEQ Working Paper 262*, Society for the Study of Economic Inequality*, 2012.

Angel, S. and B Bittschi, “Housing and Health“, *ZEW Discussion Paper No. 14-079*, Mannheim, 2014.

Bachmann, R., P. Bechara and S. Schaffner, “Wage Inequality and Wage Mobility in Europe”, *Review of Income and Wealth*, forthcoming, 2014.

Bachmann, R. and S. Schaffner , “Biases in the Measurement of Labor Market Dynamics”, *SFB 475 Technical Report #12*, 2009.

Eurostat, “DESCRIPTION OF TARGET VARIABLES: Cross-sectional and Longitudinal 2008 operation (Version January 2010)”, 2010.

Frick, J. R. and K. Krell, “Einkommensmessungen in Haushaltspanelstudien für Deutschland: Ein Vergleich von EU-SILC und SOEP“, *AStA Wirtschafts- und Sozialstatistisches Archiv,* 5(3), 221-248, 2011.

Giannetti, M., D. Federici and M. Raitano, “Migrant remittances and inequality in Central-Eastern Europe”, *International Review of Applied Economics*, 23(3), 289-307, 2009.

Goedemé, T., “The standard error of estimates based on EU-SILC. An exploration through the Europe 2020 poverty indicators”, *Herman Deleeck Centre for Social Policy Working Paper 1009*, University of Antwerp, 2010.

Iacovou, M., O. Kaminska and H. Levy, “Using EU-SILC data for cross-national analysis: strengths, problems and recommendations”, *ISER Working Paper Series 2012-03*, Institute for Social and Economic Research, 2012.

Jürges, H., “Unemployment, life satisfaction and retrospective error”, *Journal of the Royal Statistical Society,* 170(1), 43-61, 2007.

Longford, N. T., M. G. Pittau, R. Zelli and R. Massari, “Poverty and inequality in European regions”, *Journal of Applied Statistics*, 39(7), 1557-1576, 2012.

Mathiowetz, N. A. and G. J. Duncan, "Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment," *Journal of Business & Economic Statistics*, American Statistical Association, 6(2), 221-229. 1988.

StataCorp. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP. 2011.

Whelan, C. T. and B. Maitre, “Identifying Childhood Deprivation: How Well Do National Indicators of Poverty and Social Exclusion in Ireland Perform?”, *The Economic and Social Review, Economic and Social Studies,* 43(2), 251–272, 2012.

Appendix

Table A1: Data Availability by Country
2004 to 2011

Country	2004	2005	2006	2007	2008	2009	2010	2011
Austria	x	x	x	x	x	x	x	x
Belgium	x	x	x	x	x	x	x	x
Bulgaria			x	x	x	x	x	x
Cyprus		x	x	x	x	x	x	x
Czech Republic		x	x	x	x	x	x	x
Germany		x	x					
Denmark	x	x	x	x	x	x	x	x
Estonia	x	x	x	x	x	x	x	x
Spain	x	x	x	x	x	x	x	x
Finland	x	x	x	x	x	x	x	
France	x	x	x	x	x	x	x	
Greece	x	x	x	x	x	x	x	x
Croatia							x	x
Hungary		x	x	x	x	x	x	x
Ireland	x	x	x	x	x	x		
Iceland	x	x	x	x	x	x	x	x
Italy	x	x	x	x	x	x	x	x
Lithuania		x	x	x	x	x	x	x
Luxembourg	x	x	x	x	x	x	x	x
Latvia		x	x	x	x	x	x	x
Malta			x	x	x	x	x	x
Netherlands		x	x	x	x	x	x	x
Norway	x	x	x	x	x	x	x	x
Poland		x	x	x	x	x	x	x
Portugal	x	x	x	x	x	x	x	x
Romania				x	x	x	x	x
Sweden	x	x	x	x	x	x	x	
Slovenia		x	x	x	x	x	x	x
Slovakia		x	x	x	x	x	x	
United Kingdom		x	x	x	x	x	x	x

Source: EU-SILC.

Table A2: Number of observations
2004 to 2011

Country	2004	2005	2006	2007	2008	2009	2010	2011
Austria	11,550	13,264	15,071	16,939	13,853	13,803	14,307	10,020
Belgium	9,945	12,753	13,924	15,002	14,510	14,466	14,647	9,929
Bulgaria			6,335	8,765	11,839	14,337	15,831	12,993
Cyprus		8,506	11,204	10,756	10,139	9,415	11,241	8,440
Czech Republic		10,333	18,018	23,231	27,142	23,494	21,588	15,279
Germany		24,999	23,212					
Denmark	7,112	9,985	12,323	12,825	12,608	11,656	11,214	7,738
Estonia	11,665	12,115	16,153	14,622	13,227	13,843	13,687	9,622
Spain	34,232	38,271	35,307	35,252	36,621	37,433	37,670	25,808
Finland	15,474	19,741	18,499	17,639	17,062	16,267	19,755	14,568
France	22,144	24,463	29,335	26,289	25,878	25,988	22,244	
Greece	12,887	15,161	15,443	15,025	17,118	18,263	13,154	
Croatia							9,863	8,013
Hungary		14,567	20,194	22,471	22,626	24,663	24,807	17,557
Ireland	8,300	12,959	12,934	11,844	11,948	6,410		
Iceland	6,134	8,429	7,839	7,510	7,546	7,452	7,937	5,488
Italy	46,809	56,753	55,033	53,279	53,036	51,775	47,957	32,522
Lithuania		9,100	12,392	13,055	12,350	13,129	13,460	9,704
Luxembourg	9,780	9,806	10,313	10,341	10,144	11,526	13,510	11,685
Latvia		9,018	11,212	11,442	13,438	14,766	15,785	11,580
Malta		0,0	3,376	6,210	8,010	10,271	10,488	7,773
Netherlands		21,634	23,371	26,202	25,739	23,973	24,916	16,743
Norway	14,142	16,244	15,305	15,360	14,276	13,616	12,749	9,301
Poland		36,525	45,856	43,458	41,885	39,250	37,960	27,035
Portugal	7,092	13,227	18,274	17,339	15,426	20,691	20,445	10,526
Romania				14,902	19,272	18,829	18,424	13,530
Sweden	13,734	19,143	16,245	16,360	16,364	16,221	11,666	
Slovenia		27,679	31,903	28,885	29,511	30,179	30,127	19,899
Slovakia		11,779	15,080	14,329	14,992	16,017	11,684	
United Kingdom		20,816	29,095	23,886	21,386	19,664	18,989	11,800
EU-SILC	231,000	477,270	543,246	533,218	537,946	537,397	526,105	327,553

Source: Source: EU-SILC, own calculations.

Table A3: Construction of cross-sectional weights (using base weights)
2004 to 2011

	2004	2005	2006	2007	2008	2009	2010	2011
Austria, Finland	2005	2005*3/4	2006/4	2007/4	2008/4	2009/4	2010/4	2011/3
		2006/4	2007/4	2008/4	2009/4	2010/4	2011/4	
Bulgaria, Romania			(2008/2)	2008/3	2008/4	2009/4	2010/4	2011/3
					2009/4	2010/4	2011/4	
Belgium, Cyprus, Denmark,	(2006/2)	2006/3	2006/4	2007/4	2008/4	2009/4	2010/4	2011/3
Hungary, Portugal, Slovenia,	(2005/4)	(2005/4)	2007/4	2008/4	2009/4	2010/4	2011/4	
United Kingdom								
Czech Republic		2008	2008/2	2008/3	2008/4	2009/4	2010/4	2011/3
					2009/4	2010/4	2011/4	
Germany		2006/3	2006/4					
Estonia	2006/4	2006/5	2006/5	2007/4	2008/4	2009/4	2010/4	2011/3
			2007/5	2008/4	2009/4	2010/4	2011/4	
Spain	2005*4/3	2005/4	2006/4	2007/4	2008/4	2009/4	2010/4	2011/3
		2006/4	2007/4	2008/4	2009/4	2010/4	2011/4	
France	2006*4/3*8/9	2006*4/3*8/9	2006*4/3*8/9	2007*4/3*8/9	2008*4/3*8/9	2009*4/3*8/90	2010*4/30	
	2005*8	2005*8	2007*4/3*8/9	2008*4/3*8/9	2009*4/3*8/9	2010*4/3*8/90		
Greece	2005	2005*3/4	2007/3	2007/4	2008/4	2009/4	2010/3	
		2007/4		2008/4	2009/4	2010/4		
Croatia							2011/2	2011/2
Ireland	2006/2	2006/4	2006/4	2007/4	2008/4	2009/3		
		2005*3/4	2007/4	2008/4	2009/4			
Iceland, Italy	2005	2005*3/4	2006/4	2007/4	2008/4	2009/4	2010/4	2011/3
		2006/4	2007/4	2008/4	2009/4	2010/4	2011/4	
Lithuania	2005/3	2006/4	2007/4	2007/4	2008/4	2009/4	2010/4	2011/3
		2007/4	2008/4	2008/4	2009/4	2010/4	2011/4	
Luxembourg	2007	2007	2007	2009	2009	2010/2	2010/2	2011/3
Latvia, Netherlands		2005/3	2006/4	2007/4	2008/4	2009/4	2010/4	2011/3
			2007/4	2008/4	2009/4	2010/4	2011/4	
Malta			2009	2009/2	2009/2	2009/4	2010/4	2011/3
						2010/4	2011/4	
Norway	2005* 42.2	2005*347	2006/8	2007/7	2008/6	2009/5	2010/7	2011/6
	billions	2006/8	2007/8	2008/7	2009/6	2010/5	2011/7	
	2007/7	2007/8	2008/8	2009/7				
			2009/8					
Poland			2006/3	2007/4	2008/4	2009/4	2010/4	2011/3
			2007/4	2008/4	2009/4	2010/4	2011/4	
Sweden	2005	2005*3/4	2006/4	2007/4	2008/4	2009/4	2010/3	
		2006/4	2007/4	2008/4	2009/4	2010/4		
Slovakia	(2006/2)	2006/3	2006/4	2007/4	2008/4	2009/4	2010/3	
	(2005/4)	(2005/4)	2007/4	2008/4	2009/4	2010/4		

Source: EU-SILC, own calculations.

Notes: p2005, p2006, ..., p2011 represent the weights of the L2005, L2006, ..., L2011 longitudinal files. Denmark: The first file is L2006 that also covers data for 2003 and 2004; Ireland: All observations from the L2005 file are also included in L2006. Therefore only L2006, L2007, L2008 and L2009 are used; France: There are nine rotational groups instead of four. Norway: Eight rotational groups; Luxembourg: It is no rotational panel. Portugal: The first file is L2006 but also covers data for 2004.

Table A4: Share of consistent labor market states in monthly and yearly data by country (in per cent)

Country	2004	2005	2006	2007	2008	2009	2010
Austria	85.62	91.61	92.59	89.08	88.45	88.89	89.88
Belgium	90.11	88.25	89.92	90.69	88.39	89.13	89.08
Bulgaria			85.09	79.05	91.07	92.95	91.31
Cyprus		92.87	98.36	98.45	98.52	98.56	98.92
Czech Republic		90.72	91.83	92.22	95.07	97.09	97.03
Germany		89.04					
Denmark	82.40	86.48	85.81	63.51	87.64	86.15	87.57
Estonia	95.54	96.66	99.83	99.93	99.99	99.71	99.70
Spain	82.83	83.99	83.80	85.56	86.23	87.27	84.73
Finland	85.06	85.96	85.25	85.36	87.92	88.93	88.48
France	97.83	98.20	98.17	98.32	98.46	98.44	
Greece	88.38	91.21	90.74	91.16	92.57	91.95	
Croatia							88.85
Hungary		81.08	81.36	80.64	80.87	86.34	86.52
Iceland			70.45	73.25	79.82	76.61	77.00
Italy	89.98	83.73	83.99	82.97	82.27	84.75	80.33
Lithuania		91.23	98.46	97.09	98.69		
Luxembourg	88.87	96.37	96.67	96.76	98.39	98.96	99.01
Latvia		84.85	86.46	86.45	85.75	85.75	87.46
Malta			86.56	90.89	97.15	99.13	99.60
Netherlands		74.92	80.59	79.57	78.82	79.78	75.83
Norway	78.79	78.56	78.61	80.24	83.86	83.34	82.26
Poland		87.38	90.74	90.50	91.55	94.93	95.43
Portugal	92.19	92.45	93.26	87.49	86.86	90.97	90.62
Romania				86.11	51.68	94.20	51.29
Sweden	87.48	81.81	86.61	86.57	85.57	83.18	
Slovenia		90.98	90.67	90.32	93.79	93.36	93.24
Slovakia		91.20	92.55	93.71	93.89		
United Kingdom			96.45	98.24	97.55	97.65	98.26
EU-SILC	89.00	88.26	89.45	88.49	88.00	91.03	87.74

Source: EU-SILC, own calculations.