RUHR
ECONOMIC PAPERS

Melissa Engel
Sandra Schaffner

How to Use the EU-SILC Panel to Analyse Monthly and Hourly Wages

RWI #390

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #390

Responsible Editor: Christoph M. Schmidt

ISSN 1864-4872 (online) – ISBN 978-3-86788-445-7

Ruhr Economic Papers #390

Melissa Engel and Sandra Schaffner

How to Use the EU-SILC Panel to Analyse Monthly and Hourly Wages

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788445
ISSN 1864-4872 (online)
ISBN 978-3-86788-445-7

Melissa Engel and Sandra Schaffner[1]

How to Use the EU-SILC Panel to Analyse Monthly and Hourly Wages

Abstract

The European Union Statistics on Income and Living Conditions (EU-SILC) is a rotational panel provided by Eurostat that covers important variables over all EU Member States. Unfortunately, Eurostat provides separate data sets which do not cover all waves. Furthermore, information on monthly income and hourly wages are missing. In this paper, we make two contributions: first, we develop a method for combining the different waves in order to increase the number of observations; second, we derive monthly and hourly pay.

JEL Classification: C81, C83, D31

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

December 2012

1 Melissa Engel, University Erlangen-Nürnberg; Sandra Schaffner, RWI. – We are grateful to Alfredo Paloyo for useful comments. – All correspondence to Sandra Schaffner, RWI, Hohenzollernstr. 1-3, 45128 Essen, Germany, E-Mail: sandra.schaffner@rwi-essen.de.

1 Introduction

The European Union has seen an increase in the number of inhabitants as well as in the circumstances that are regulated on a European level. Therefore, analyses on the European level as well as between the different Member States have become more important. To this end, individual data for inhabitants of EU Member States are very important for economic research because it sheds light on issues which may be glossed over by relying on macroeconomic indicators alone.

National data sets have the advantage of a rather large number of observations. In addition, the structure is well adapted for the idiosyncratic settings of the respective country. However, the use of national data leads to problems regarding the comparability between countries and thus for cross-country analyses. Due to this concern, 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.

Although the EU-SILC data cover a wide variety of subjects, the number of existing studies using these data is quite rare. With very few exceptions, only studies on poverty (e.g., Longford et al. 2012 and Whelan/Maître 2012), inequality (e.g. Giannetti/Federici/Raitano 2009) and wage mobility (e.g. Aristei/Perugini 2012 and Bachmann/Bechara/Schaffner 2012) exist. This can be due to the fact that there are still some shortcomings of the data sets that discourage further analyses.

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 best analyse the existing data. Iacovou, Kaminska and Horacio (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 et al. (2010) show that there are differences in the measured inequality and poverty for Germany compared to values derived using the well-established German Socioeconomic Panel. Goedemé (2010) presents the necessary sample design to estimate reliable standard errors using EU-SILC data.

As an extension to the existing papers on data quality, the present study gives a brief overview on data problems and the quality of the data with special regard to income and the rotational panel design. Compared to the European Labour Force Survey, its longitudinal structure and the information on income are advantageous. 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 are smaller than what is possible. However, to precisely estimate parameters in multivariate analyses, a large number of observations is necessary. Therefore, it is of interest for researchers to capture all information available in one data set.

The number of observations can be increased by first merging different datasets and then reweighting the observations. In this paper, we describe this process for the years 2004 to 2009 and consequently 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 do not capture the same time period. This limits the possibilities to use the data set for important 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.

Table 1 Data Availability by Country
2004 to 2009

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

Source: EU-SILC.

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 set covers two different types of data: the cross-sectional and longitudinal micro data set. 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 2004-2009. Except for a comparison of labour income the whole analysis concentrates on the longitudinal files. We use the *EUSILC LONGITUDINAL UDB 2009 – version-2 of March 2012, EUSILC LONGITUDINAL UDB 2008 – version-4 of March 2012, EUSILC LONGITUDINAL UDB 2007 – version-5 of August 2011, EUSILC LONGITUDINAL UDB 2006 – version-2 of* February 2008 *and EUSILC LONGITUDINAL UDB 2005 of* February 2008.

In addition to the structure of the data, the cross-sectional data and the longitudinal data also differ to some extent in the covered variables. 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 would concern 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

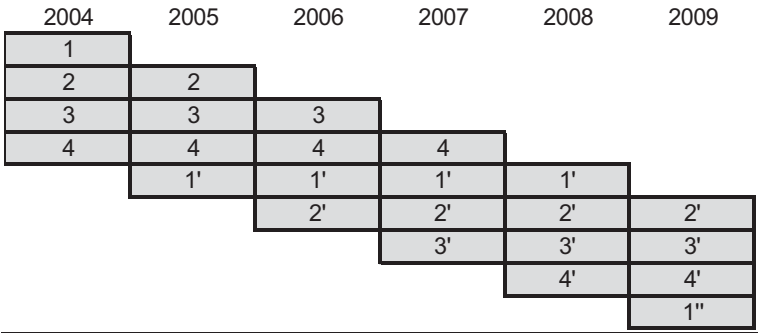
Besides the variables which are available in the cross-sectional data files, the exact month of the interview would be helpful but is not available. Especially in order to generate monthly income and monthly transition rates, it is important to compare the yearly interviews with the calendar data as presented in the next sections.

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, are only available for these selected respondents. Therefore, the number of observations decreases if variables affected by this selection process are used.

The data versions delivered by Eurostat contain the longitudinal files L2005-L2009, 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 2009. For most countries, information is available for a shorter period (see Table 1). Since the 2003

wave was a pilot survey, we exclude it from the analyses. Data for the whole period (2004-2009) is available for Austria, Belgium, Denmark, Estonia, Spain, Finland, France, Greece, Ireland, Iceland, Italy, Luxembourg, Norway, Portugal and Sweden. For all other countries, five years are available except for Bulgaria (2006-2009), Germany (2005-2006) [3], Romania (2007-2009) and Malta (2006-2009).

Figure 1 The integrated design of EU-SILC



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[4]) and each year one quarter of all respondents are replaced by new respondents. The integrated design consists in selecting four panels at the first wave. Each subsequent year, a panel is dropped and replaced by a new replication. 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 2 Structure of the 2009 longitudinal data file (L2009)

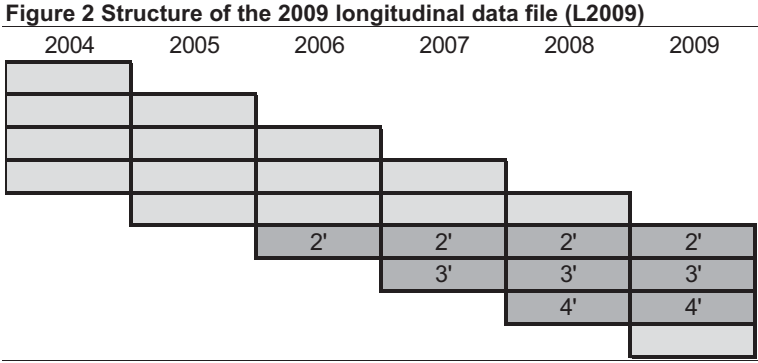
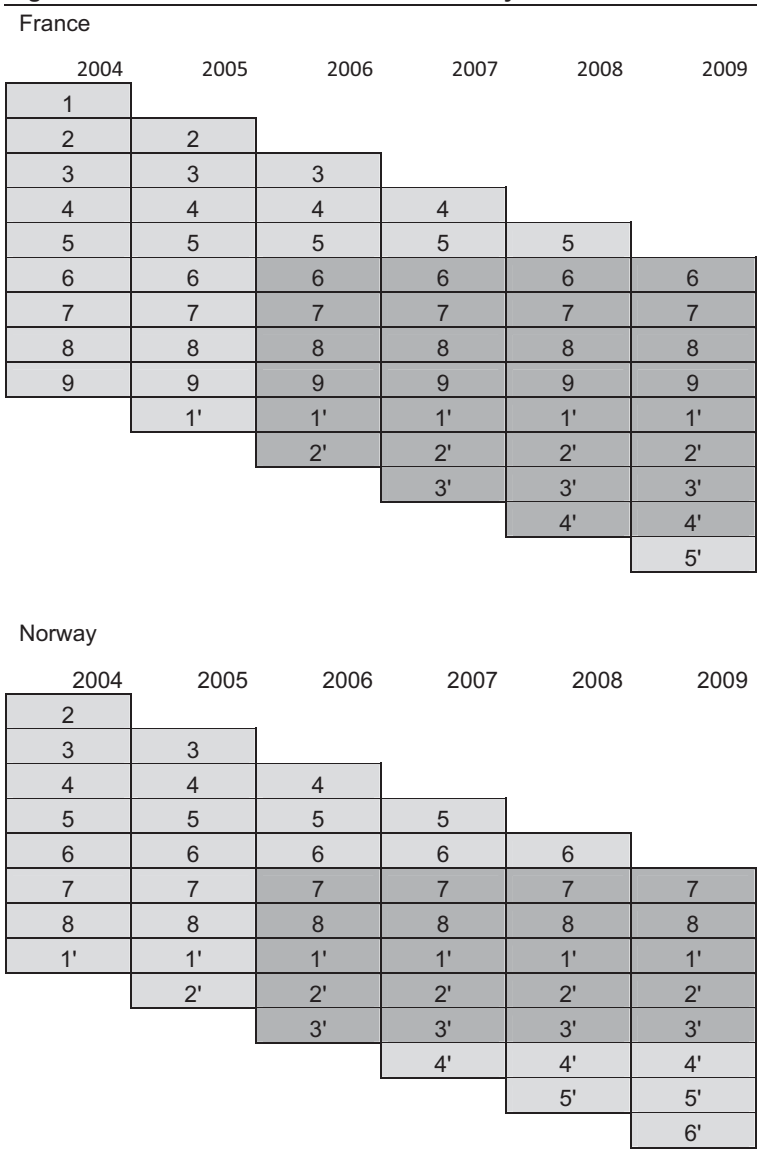


Figure 1 shows the panel structure of the EU-SILC data for a country that first starts in 2004. Of the individuals interviewed in 2004, three quarters are also interviewed in 2005 while the first group is replaced by a new subsample (1'). In the following year another quarter of individuals (group 2) 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 interviewed. This fraction decreases to zero in the 2008 and 2009 waves. Group 4 is the first group that is

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

interviewed over a four year period. Therefore, for countries with data availability from 2004 to 2009, three rotational groups (group 4, group 1' and group 2') are interviewed four times in the whole period. However, the data sets that are distributed by Eurostat do not cover all of these rotational groups in one data set.

Figure 3 Panel Structure in France and Norway



For a given year, the respective longitudinal file available from Eurostat (e.g. L2009) only contains those respondents that were interviewed both in the respective year and in the preceding year. This means that in the 2009 longitudinal wave (L2009), information is only included for those individuals who were interviewed at least in 2009 and 2008. Individuals,

who were interviewed in 2004, 2005, 2006, 2007 and/or 2008 but not in 2009, are not included in the 2009 longitudinal wave. Figure 2 illustrates the panel groups that are included in the 2009 longitudinal file (dark grey). This figure shows that only 25 per cent of all interviews conducted in 2006 are reported in the 2009 longitudinal file, and there are no observations for 2004 and 2005 at all. Of all 2007 interviews only one half and of the 2008 and 2009 observations three quarters are reported. Therefore, this way of constructing the longitudinal data set leads to an important loss of observations, and, 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 by time.

For France, Norway and Luxembourg, the panel structure is different. 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 observation are replaced. The same is true for Norway where the panel consists of eight rotational groups. Each year 1/8 of observations are substituted with new persons. While the French L2009 version uses the groups 6 to 4' for the years 2006 to 2009, the panel of Norway is smaller: the groups 6 to 3' are reported. In contrast to the other countries, a balanced panel is constructed in Norway. For France, 66 to 89 per cent of all observations for the years 2006 to 2009 are in the data file. For Norway, only 61 to 88 per cent are reported. Overall, it can be seen that the share reported 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 36 to 53 per cent are included in the L2009 file.

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 number of observations overall 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 and 2009 (L2005-L2009) that each can be derived by 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. Figure 4 presents the composition of our resulting data set for the countries that are observed for the entire time period 2004-2009. It can be seen that all observations are included except for the observations of group 1 and the 2009 observations of group 1''.[5] Therefore, in 2004 and 2009 three quarters of all observations are included in our resulting data set, while all the observations for the years 2005 to 2008 are included. More generally, 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 will also be true if additional panel waves (L2010, L2011,...) are added. However, with

[5] *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.*

L2005-L2009 being available, a full sample of all observations can be reconstructed for the years 2005 to 2008.

Table 2 Number of observations
2004 to 2009

Country	2004	2005	2006	2007	2008	2009
Austria	7,930	9,079	10,144	11,185	9,075	5,975
Belgium	6,672	8,525	9,342	10,054	9,733	6,692
Bulgaria	-	-	4,309	5,878	7,738	6,088
Cyprus	-	5,816	7,593	7,276	6,878	4,765
Czech Republic	-	7,170	12,274	15,852	18,284	11,779
Germany	-	16,964	15,653	-	-	-
Denmark	4,941	6,775	8,257	8,496	8,538	5,682
Estonia	8,353	8,541	11,370	10,193	9,105	7,152
Spain	23,356	25,937	23,796	23,500	24,395	17,642
Finland	11,038	13,910	12,809	12,107	11,684	7,972
France	14,595	16,130	19,240	17,153	16,781	13,906
Greece	8,563	9,833	9,877	9,575	10,885	7,699
Hungary	-	9,816	13,570	15,261	15,375	10,935
Ireland	5,152	7,973	7,807	7,024	7,114	3,833
Iceland	4,229	5,698	5,239	5,061	5,144	3,419
Italy	31,851	38,043	36,425	35,110	34,642	24,006
Lithuania	-	6,352	8,385	8,664	8,140	6,222
Luxembourg	6,786	6,750	7,012	7,068	6,860	5,523
Latvia	-	6,119	7,369	7,405	8,739	6,367
Malta	-	-	2,306	4,293	5,470	4,713
Netherlands	-	14,571	15,688	17,665	17,349	10,864
Norway	9,665	11,041	10,254	8,529	6,755	5,163
Poland	-	25,760	31,835	29,966	28,765	20,074
Portugal	4,790	5,035	5,454	5,471	6,110	5,728
Romania	-	-	-	10,224	12,982	9,556
Sweden	9,416	12,063	10,735	10,789	10,876	7,431
Slovenia	-	20,925	23,812	21,343	21,517	14,242
Slovakia	-	8,746	11,185	10,683	11,242	8,719
United Kingdom	-	13,561	15,192	14,093	13,519	8,007
EU-SILC	157,337	321,133	356,932	349,918	353,695	250,154

Source: EU-SILC, own calculations.

Figure 4 The resulting estimation data set

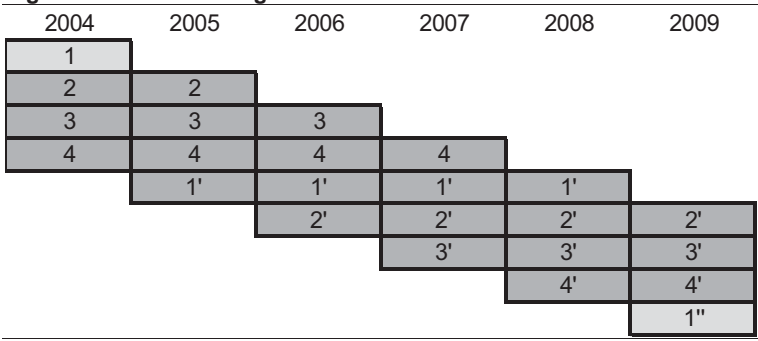


Table 3 Number of "selected" respondents
2004 to 2009

Country	2004	2005	2006	2007	2008	2009
Denmark	2,325	3,175	3,400	3,241	3,400	2,602
Finland	5,062	6,282	5,926	5,705	5,485	3,788
Iceland	1,723	2,239	2,088	2,034	2,083	1,415
Netherlands	-	7,167	7,578	8,606	8,557	5,404
Norway	4,457	4,980	4,779	4,020	3,270	2,523
Sweden	4,314	5,312	4,896	4,984	4,983	3,519
Slovenia	-	6,737	7,728	7,037	7,254	4,803

Source: EU-SILC, own calculations.

Finally, when combining L2005 – L2009, we have about 1.8 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 for which we observe a country, the smallest number of observations is recorded (see Table 2). Therefore, more than 320,000 individuals are observed in the years 2005 to 2008, while we only observe roughly 158,000 in 2004 and 250,000 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 3), the number of observations becomes very small.

In survey data, weights are used when the survey is not representative for the whole population. Weights cover the information how many individuals in the whole population are represented by a single individual. Therefore, those groups that are underrepresented in the data have a higher weight since they represent more people in the whole population. To account for this new data structure, the delivered weights have to be adopted since they are made for the design with fewer observations used by Eurostat and therefore a different weighting of the different years in the longitudinal versions. The aim is that the weights are designed in such a way that the number of observations represents the whole population.

In the dataset reported, i.e. the integrated design, the weights are, for example, adjusted for 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 the observations in t and not for the earlier observations.

Table 4 Construction of cross-sectional weights (using base weights)

2004 to 2009

	2004	2005	2006	2007	2008	2009
Austria, Finland, Sweden	p2005	p2005*3/4	p2006/4	p2007/4	p2008/4	p2009/3
		p2006/4	p2007/4	p2008/4	p2009/4	
Bulgaria, Romania			(p2008/2)	p2008/3	p2008/4	p2009/3
					p2009/4	
Belgium, Cyprus, Denmark, Hungary, Portugal, Slovenia, Slovakia, United Kingdom	(p2006/2)	p2006/3	p2006/4	p2007/4	p2008/4	p2009/3
	(p2005/4)	(p2005/4)	p2007/4	p2008/4	p2009/4	
Czech Republic		p2008	p2008/2	p2008/3	p2008/4	p2009/3
					p2009/4	
Germany		p2006/3	p2006/4			p2009/3
Estonia	p2006/4	p2006/5	p2006/5	p2007/4	p2008/3	p2009/3
Spain	p2005*4/3	p2005/4	p2006/4	p2007/4	p2008/4	p2009/3
		p2006/4	p2007/4	p2008/4	p2009/4	
France	p2006*4/3*8/9	p2006*4/3*8/9	p2006*4/3*8/9	p2007*4/3*8/9	p2008*4/3*8/9	p2009*4/3
	p2005*8	p2005*8	p2007*4/3*8/9	p2008*4/3*8/9	p2009*4/3*8/9	
Greece	p2005	p2005*3/4	p2007/3	p2007/4	p2008/4	p2009/3
		p2007/4		p2008/4	p2009/4	
Ireland	p2006/2	p2006/4	p2006/4	p2007/4	p2008/4	p2009/3
		p2005*3/4	p2007/4	p2008/4	p2009/4	
Iceland, Italy	p2005	p2005*3/4	p2006/4	p2007/4	p2008/4	p2009/3
		p2006/4	p2007/4	p2008/4	p2009/4	
Lithuania	p2005/3	p2006/4	p2007/4	p2007/4	p2008/4	p2009/3
		p2007/4	p2008/4	p2008/4	p2009/4	
Luxembourg	p2007	p2007	p2007	p2009	p2009	p2009
Latvia, Netherlands		p2005/3	p2006/4	p2007/4	p2008/4	p2009/3
			p2007/4	p2008/4	p2009/4	
Malta			p2009	p2009/2	p2009/2	p2009/3
Norway	p2005* 42,205,170,857	p2005*347	p2006/8	p2007/7	p2008/6	p2009/3
	p2007/7	p2006/8	p2007/8	p2008/7	p2009/6	p2009/5
		p2007/8	p2008/8	p2009/7		
			p2009/8			
Poland			p2006/3	p2007/4	p2008/4	p2009/3
			p2007/4	p2008/4	p2009/4	

Source: EU-SILC, own calculations.

Notes: p2005, p2006, ..., p2009 represent the weights of the L2005, L2006, ..., L2009 longitudinal files. Denmark: The first file is L2006 that also covers data for 2003, 2004 and 2005; Ireland: All observations from the L2005 file are also included in 2006. Therefore only L2006, L2007 and L2008 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 and 2005.

Table 5
Number of persons with changing sex

Country	2004	2005	2006	2007	2008	Total
Belgium	1	0	0	0	0	1
Finland	1	0	0	0	0	1
France	2	1	35	4	1	43
Greece	0	0	0	3	0	3
Lithuania	0	417	5	1	0	423
Luxembourg	1	0	0	0	0	1
Norway	6	6	1	0	0	13
Portugal	0	591	17	2	0	610
Slovakia	0	460	9	1	0	470
UK	0	0	1	0	0	1
Total	11	1.475	68	11	1	1566

Source: EU-SILC, own calculations.

The procedure to build new weights in 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 L2009 for the year 2009, the weights of L2008 for the year 2008 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 observe 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, we adjust the weights in such a way that the whole population of each country is always represented by the observations included in the data by multiplying the weights by 1.333 in the first and last year.

Although using panel data it is of interest for researchers, having cross-sectional weights for cross-sectional analyses is so as well. 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 and 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 1.333 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[6] to derive the population. The factors as well as the country-specific calculation methods are shown in Table 4. 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 depends on the year of the first and the last observations. For example for Austria, Finland and Sweden the weights for 2004 can be used from the L2005 file. For the year 2005 one uses the weights of the L2005 file, too. Since they only represent three

[6] *The factors are derived with the population numbers provided by Eurostat.*

quarter of the population they are weighted with this factor. Those observations that are first illustrated in the L2006 file get the weight of the L2006 file.

Table 6
Number of persons with changing year of birth

Country	2003	2004	2005	2006	2007	2008	Total
Denmark	0	0	0	0	2	0	2
Finland	0	1	0	0	0	0	1
France	0	5	1	48	4	2	60
Greece	0	0	0	0	4	0	4
Lithuania	0	0	889	10	2	0	901
Luxembourg	1	1	0	0	0	0	2
Norway	27	7	14	1	0	0	49
Portugal	0	0	1704	51	6	0	1761
Slovenia	0	0	0	3	0	0	3
Slovakia	0	0	1230	14	2	0	1246
UK	0	0	0	1	0	0	1
Total	28	14	3838	128	20	2	4030

Source: EU-SILC, own calculations.

Table 7
Number of persons changing sex and year of birth

Country	2003	2004	2005	2006	2007	2008	Total
Finland	0	1	0	0	0	0	1
France	0	0	1	34	4	1	40
Greece	0	0	0	0	3	0	3
Lithuania	0	0	412	5	1	0	418
Norway	11	2	5	1	0	0	19
Portugal	0	0	582	17	2	0	601
Slovakia	0	0	446	9	1	0	456
UK	0	0	0	1	0	0	1
Total	11	3	1446	67	11	1	1539

Source: EU-SILC, own calculations.

The person identifier (RB030) in the longitudinal files gives the opportunity to observe one person over several years. However, after merging the different data some persons who change their gender and/or their date of birth can be observed in the data. Therefore some identifiers (IDs) seem to be assigned to different individuals. In our sample of persons aged between 15 and 65, there are 1566 persons who change their gender (Table 5) and 4030 persons who change their year of birth over time (Table 6). A simultaneous change in age and sex can be observed for 1539 individuals. (see Table 7). The majority of these cases occur due to the merging of the different files. In Lithuania, Portugal and Slovakia different individuals in different data versions have the same ID. These observations have to be separated into two new IDs before and after the change. However, some occurrences appear to be due to different persons with same IDs. Is the same ID assigned to different individuals at different time periods? Or are several IDs interchanged and therefore the same person is interviewed with different IDs?

To detect those IDs that are assigned twice and thus assumed to display more than one person, we try to identify the cases where the difference in the years of birth is greater than 5 or where there is a change of two or more characteristics. We assume the quarter and year of birth, the person's sex and its rotational panel group not to change over time. As Table 8

displays, there are 4842 persons who change more than one characteristic) over time or have a difference in their years of birth of at least 5 years. Persons changing only one indicator or having less than 5 years difference in their years of birth are not assumed to represent two persons. It is assumed that data has been falsely reported. We construct new IDs that are not existent in the data yet. Therefore all new IDs are designed to display a 9 after the country code which is a one or two-digit number.

Table 8
Number of IDs belonging to more than one person

Country	0	1	Total
Danmark	0	2	2
Spain	12	4	16
Finland	0	1	1
France	73	58	131
Greece	3	4	7
Lithuania	170	880	1050
Luxembourg	0	5	5
Norway	68	54	122
Portugal	310	1733	2043
Slovakia	236	1224	1460
Total	874	3968	4842

Source: EU-SILC, own calculations. One indicates that there are two or more changes in characteristics or years of birth differ by five or more years. Zero indicates those observations with only one change in characteristics.

4 Construction of a monthly data set

One of the main advantages of the EU-SILC data set is that they cover 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 states based on the calendar information can be generated. This variable covers four employment states 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 covers the income reference period of the respective interview while most of the characteristics of the respondents cover 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 of Ireland and the UK, the reference period is always the preceding calendar year. That means that the calendar data in the 2009 survey cover January to December 2008. 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 dataset to cover within year labour market dynamics. We expand the yearly dataset by twelve and generate a monthly data set for the years 2003 to 2008 and for 2004 to 2009 for the UK and Ireland, respectively. Unfortunately, we can observe other characteristics (e.g. marital status, health, household size etc.) only at the date

of the interview. As a result, there is a time lag between the different kinds of information. In the first year only information of the next interview is available. Afterwards the monthly information can be combined with the yearly interview of the same year.

Table 9
Share of consistence labour market states in monthly and yearly data by country
in per cent

Country	2004	2005	2006	2007	2008	2009
Austria	82.53	89.18	89.43	85.76	85.14	-
Belgium	86.72	85.43	87.87	88.00	85.92	-
Bulgaria	-	-	78.84	70.10	88.08	-
Cyprus	-	91.79	97.03	97.16	96.41	-
Czech Republic	-	87.76	89.93	89.85	93.15	-
Germany	-	85.51	-	-	-	-
Denmark	80.74	85.03	78.80	59.04	40.68	-
Estonie	92.18	93.44	97.58	97.47	97.11	-
Spain	78.08	79.28	80.33	82.28	82.16	-
Finland	68.83	64.12	64.80	61.81	41.13	-
France	95.44	95.26	95.79	95.98	95.84	-
Greece	84.06	87.79	87.87	87.62	89.72	-
Hungary	-	76.31	76.66	76.71	76.75	-
Ireland	79.16	81.57	82.14	82.11	81.37	69.39
Iceland	4.26	3.07	66.32	68.23	74.75	-
Italy	88.57	82.62	83.28	81.41	81.14	-
Lithuania		86.93	92.26	89.78	95.78	-
Luxembourg	85.52	93.59	94.47	94.39	94.66	-
Latvia	-	78.45	79.89	80.10	79.66	-
Malta	-	-	85.06	87.84	97.28	-
Netherlands	-	76.45	83.42	80.71	-	-
Norway	60.46	64.19	61.60	60.01	77.22	-
Poland	-	83.24	85.13	85.55	86.50	-
Portugal	89.39	82.71	81.86	80.84	86.20	-
Romania	-	-	-	83.34	64.75	-
Sweden	-	72.95	72.72	73.70	73.52	73.97
Slovenia	-	88.04	87.53	86.51	90.52	-
Slovakia	-	89.00	89.96	88.04	91.23	-
United Kingdom	-	-	77.85	77.66	77.45	69.93

Source: EU-SILC, own calculations.

The comparison between the calendar data and the yearly interviews can give some hints on the quality of the retrospective calendar data. As can be seen in Table 9, the majority of individuals report one year later 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 remembering errors but also of the different definitions of labour market status in the two questions. In the yearly interview, individuals are asked about their actual labour market status, that can 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.

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 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 on the base weights of the longitudinal datasets. In this procedure, 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

Another important advantage of the EU-SILC data in comparison to the EU-LFS data is that it also covers income information. The wage rate is an important dependent and explanatory variable in labour market analyses. 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, the current economic states of the individuals in the data set are known at the time of the interviews; the economic status is also contained in the monthly information for the previous calendar year (see previous section). For example, the labour income in the 2009 interview covers the calendar year 2008. Therefore, it is possible that it does not correspond to the current job as described in the 2009 interview. If someone has been interviewed before, the interview of 2008 and the income information of 2009 overlap. However, the problems that the labour market status is only a snap-shot and that income information only cover twelve months still exist. 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.) the different time period 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, since the yearly income may accrue in only a few months of employment. Therefore, the duration spent in the different states during the year has to be taken into account. The retrospective main economic status (calendar data) gives us 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 cover only 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 were at least in two different labour market states (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 someone changed his 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 get the monthly labour income or unemployment benefits, respectively.*

Additionally to those who are in one of the three labour market states 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 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 working status but on the income situation and family/household characteristics. We therefore assume these values are uniformly distributed over the year.[7]

In the cross-sectional data provided by Eurostat, monthly labour income information is available for Austria, Bulgaria, Spain, Greece, Hungary, Ireland, Iceland, Italy, Poland, Portugal and the UK. However, the IDs are different to 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. Therefore, the timing of the information is different to the income period. Therefore, hourly wages can be derived in the combination of monthly data and yearly data at the date of the interview. For those workers with a different employment status in the calendar data and the current employment status in the 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

Due to the increasing political and economic integration of the European Union, it is important to conduct analyses on the European level as well as comparative studies between the different Member States. 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 separated into different files which reduce the number of observations, with negative implications for estimating statistically meaningful results. Furthermore, information on income cannot be related to the economic status and this reduces the value of the data set for labour market analyses.

In this paper, we describe some of the shortcomings of the data set 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 to a large extent, especially regarding those observations which can be observed for four years.

[7] *This may lead to problems since allowances can be assigned only for a certain time or situation and these regulations may differ over the countries. However, the data do not allow us to distinguish between different cases.*

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

Literature

Aristei, David and Cristiano Perugini (2012), The Drivers of Income Mobility in Europe, ECINEQ Working Paper 262, Society for the Study of Economic Inequality.

Bachmann, Ronald, Peggy Bechara and Sandra Schaffner (2012), Wage Inequality and Wage Mobility in Europe. Ruhr Economic Papers #386. RWI. DOI: 10.4419/86788441

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

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

Giannetti, Marilena, Daniela Federici and Michele Raitano (2009), Migrant remittances and inequality in Central-Eastern Europe, International Review of Applied Economics 23(3), 289-307.

Goedemé, Tim (2010), 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.

Iacovou, Maria, Olena Kaminska and Horacio Levy (2012), Using EU-SILC data for cross-national analysis: strengths, problems and recommendations, ISER Working Paper Series 2012-03, Institute for Social and Economic Research.

Longford, Nicholas T., Maria Grazia Pittau, Roberto Zelli and Riccardo Massari (2012), Poverty and inequality in European regions, Journal of Applied Statistics 39(7), 1557-1576.

Whelan Christopher T. and Bertrand Maitre (2012), 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.