Tax microdata vs. surveys: which are better for analysing income mobility?

María Gil Izquierdo*

Universidad Autónoma de Madrid

Fidel Picos Sánchez**

GEN-Universidade de Vigo

Abstract

Income mobility can be measured using different sources of information. Panel surveys have been the main resource used to track mobility worldwide. However, administrative data (and specifically tax data) have become a powerful alternative to surveys in last decades in many countries, such as the US, Great Britain or the Nordic countries. In Spain this alternative has remained uncommon when measuring mobility. After examining the advantages and disadvantages of both kind of data, the paper analyses whether income mobility results are different when using a survey (specifically, the European Union Living Conditions Survey) or tax data (the Spanish PIT Return Panel). We also intend to establish if these potential differences are due to methodological decisions or to the nature of the databases. We may conclude that tax data report higher immobility than surveys and that the patterns of mobility reported in every case are different.

* Corresponding author. María Gil Izquierdo. Departamento de Economía Aplicada, Fac. CC. EE., Avda. Tomás y Valiente, 5, Cantoblanco. 28049 Madrid (Spain). Phone: +34914973524. E-mail: maria.gil@uam.es

** Fidel Picos Sánchez. GEN-Departamento de Economía Aplicada, Facultade de Ciencias Empresariais e Turismo, Universidade de Vigo. As Lagoas s/n, 32004 Ourense (Spain). Phone: +34988368771. E-mail: fidel@uvigo.es

1. Introduction

Income mobility analysis is nowadays a well-established research line, with both theoretical and empirical contributions that have underpinned the study of longitudinal trends in inequality. For the last decades, the availability of new and more complete longitudinal databases in many countries (especially Anglo-Saxon countries) has definitely boosted the production of results related to changes in income distribution and therefore, the conclusions we may obtain from these results. Focusing on income mobility, there are three main ways of obtaining information to run applied analysis: census data, surveys and administrative data. Surveys have been the most commonly used data source for this purpose from the seminal studies (Solon, 1992) in most countries, though administrative data have also become an interesting alternative source of information in last decades, even certainly common in some countries (Nordic countries, Canada and more recently the US). These administrative data, which comprise tax files and Social Security records, are being considered to be more extensively exploited in future years in many countries, in order to take advantage of the easiness of information retrieval and the large coverage of individuals over the years; see for example Grusky, Mitnik and Wimer (2001) for a proposal for the US to explore the potential data sources to measure intergenerational economic mobility with tax-return data.

In this sense, the situation in Spain is somehow different. Two situations have concurred to have less empirical evidence on income mobility than in other countries. On one hand, panel databases have not experienced the same level of development as cross-section databases, so the possibilities to empirically contrast the patterns of mobility in Spain have been limited; on the other hand, the use of administrative records for mobility analysis purposes is uncommon in Spain – we are aware of just one study using a tax data (Ayala and Onrubia, 2001) – although administrative data are nowadays available.

In this context, the main question we try to answer is whether tax data can be a complementary source to survey data in order to monitor income mobility, and in for which purposes. This is a relevant questions especially in countries where tax data are not commonly used and there are maybe some prejudices to use them in mobility studies. There are some questions that may arise when thinking of exploring this possibility: are results going to be different? Of course the target population is not the same (taxpayers vs. a sample of individuals and households) but, would we obtain the same patterns in terms of mobility? To

what extent can differences in results be due to different methodological decisions for every data source or due to the nature of data? In what aspects are tax data superior to survey data?

In order to answer these questions, we compare the characteristics of the European Union Living Conditions Survey (EU-SILC hereinafter) and the Spanish PIT Return Panel 1999-2009 (Panel hereinafter), an expanded panel representative of Spanish PIT tax filers. We also use the Panel to replicate several indices calculated by Bárcena and Moro (2013), the most recent work for Spain that uses EU-SILC. In the end, this work aspires to be a user's guide to anyone interested in methodological issues and their effects on income mobility results when using tax data instead of surveys.

The paper is organized as follows. After this introduction, we revise some previous works income mobility in relation to the data used and the methodological decisions taken. Then we illustrate the advantages and disadvantages of each type of data examining and comparing the characteristics of EU-SILC and the Panel. The fourth section offers the compares de mobility results obtained with the Panel with the ones obtained with EU-SILC by Bárcena and Moro (2013). We close summarizing the main conclusions.

2. Previous works: sources of information to measure mobility

Income mobility as well as intergenerational mobility studies[1] depend upon longitudinal databases referred to individuals or households. There are three main types of data that can be used: surveys, census and administrative records. Surveys are the most broadly used to carry out this kind of empirical analysis - under the format of pure or rotating panels[2]-, due to their advantage in terms of cost in relation to census data. Administrative registers can also provide plenty of reliable longitudinal income data, collected from tax or Social Security

[1] We will not be analysing the intergenerational income mobility, that is, the transmission of economic, social, educational, etc. status from parents to descendents. Pascual (2009), Cervini-Plà (2011) or Gil and De Pablos (2010) provide results in these terms for Spain.

[2] Some examples are the US Panel Study of Income Dynamics (PSID, 1968-nowadays) and the National Longitudinal Survey (several dates); the German Socio-Economic Panel (SOEP, 1984-nowadays); the British Household Panel Survey (BHPS, 1991-2008); the Household, Income and Labor Dynamics in Australia (HILDA, 2001-nowadays); the Canadian Survey of Labor and Income Dynamics (SLID, 1998-2011). For a set of countries, the Cross-National Equivalent File (CNEF) began in 1991 with harmonization of data from the US PSID and German SOEP, incorporated the BHPS and SLID in 1999 and HILDA in 2007; data for more countries have been added subsequently (Jänttia and Jenkins, 2013); the European Community Household Panel (ECHP), 1994–2001, an harmonized cross-nationally pure panel for many of the EU country members; and its substitute, the European Statistics on Income and Living Conditions (EU-SILC), from 2005-nowadays, being its main drawback for mobility studies the only four years-tracking of the same individuals.

records. However, the use of this source of information has remained rare in mobility analysis until recently in most countries, with the exception of Scandinavian countries, though becoming increasingly used in Canada and the US (Jänttia and Jenkins, 2013). Taking into account that panel surveys have some not despicable problems (in terms of costs and size of samples, especially), and due to the full digitalization of records nowadays, administrative data have become an extraordinary resource to monitor income mobility (Grusky et al., 2001). In this sense, mainly the Anglo-Saxon and some Nordic countries have made use of different tax returns records[3] in the last decades.

Analyzing the works based on surveys (to cite some, Bradbury and Katz, 2002a, 2002b; Acs and Zimmerman, 2008; Formbya, Smith and Zheng, 2003; Mazumder, 2005) and the ones using administrative data (basically, tax data, as in Auten and Gee, 2007, 2009; Splinter and Diamond, 2009; Carrol, Joulfaian and Rider, 2006; Osterberg, 2000; Corak and Heisz, 1999) in the countries with both kind of studies, one may find that they focus in those aspects of the data whose quality is comparatively better - for instance, in the case of tax records: covering long periods of time, accurate information about personal income and detailed information about the upper part of the income distribution-. This way, results provided by survey data panels and administrative records offer complementary points of view. In fact, in the case of the US or Canada, income mobility analysis with tax data constitute a line of research itself, which attempts to determine the mobility of taxpayers. In terms of methodology, the decisions taken obviously vary due to the nature of data: individuals living in households vs. taxpayers, definitions of income provided or availability of personal information. Anyway, methodological decisions tend to be as similar as possible within every kind of study (administrative data or survey data), in order to allow the comparison of results. Focusing on differences, income mobility with tax data is usually measured with transition matrices and the indices derived from them, while survey studies tend to provide also a full set of different absolute and relative indices. Another difference is that tax data research can be very accurate on the analysis of the better-off individuals (Piketty and Saez, 1998 and 2007; Auten, Gee and Turner, 2013) while survey studies can be related to poverty dynamics (Finnie and Sweetman, 2003).

[3] Just to cite some: in the US, the Statistics of Income 1999 Edited Panel, a sample of income tax returns 1999-2007; a 17 year panel of federal individual income tax returns (1979-1995); the studies of the US Treasury, covering the period 1979-1988 (1992a, 1992b); a 10-year panel of a sample of individual income tax returns for the years 1987 through 1996 or the tax returns data published annually by the Internal Revenue Service (IRS), from 1913. For Sweden, the Swedish Income Panel (SWIP, 1978-1992). For Canada, a panel of individuals who filed an income tax return at some point between 1982 and 1986.

Deepening in the aim of this paper, that is, comparing the differences in income mobility patterns due to the differences in the source of information, there is a dearth of evidence, both international and nationally speaking. To our knowledge, only the work by Dragoset and Fields (2008) has made an attempt to empirically test how much difference it makes to mobility estimates to use administrative-based earnings rather than survey-based earnings. They find that, qualitatively, the results are similar but not identical; quantitatively, they see that magnitudes are often very different, and that these are not systematical. In a similar way, but just comparing results of previous studies in terms of volatility, Splinter et al (2009) conclude that survey-based analyses show more volatility than those using Social Security data.

Finally, and regarding the Spanish case, mobility has been mainly measured using surveys (see for example Cantó, 2000; Ayala and Sastre, 2005; Gradín, Cantó and del Río, 2008; Bárcena and Moro, 2013). A special reference has to be made to the work by Ayala and Onrubia (2001), as it is, as far as we know, the only one that has used tax data for Spain. The authors use a first Spanish PIT Return Panel for the period from 1982 to 1994. In this pioneering study both inequality and mobility issues related to the Spanish income distribution for these years are analysed. At the same time, a significant effort is made in order to explain specific methodological decisions and their implications in interpretation when using tax data instead of survey data. Their results can be slightly different in comparison to previous works, due to the nature of data. The authors justify these results referring to the sample selection and the range of the considered period of analysis as the source of the low value of the indices.

3. Data and methodological issues

As an example of tax microdata we use the 1999-2010 Spanish PIT Return Panel (hereinafter Panel), which is an expanded panel representative of each year's tax filers. Table 1 shows its main specifications.[4]

[4] Detailed information about the original design and structure of the Panel can be found in Onrubia et al (2011). Information on subsequent expansions can be found in Onrubia et al. (2012) and Pérez et al. (2013 and 2014) [all in Spanish]. The Panel is disseminated by the Spanish Institute for Fiscal Studies (Instituto de Estudios Fiscales, IEF) and available free of charge for researchers at http://www.ief.es/recursos/estadisticas/fuentes_tributarias.aspx [retrieved 16th July 2013].

Table 1. Design specifications of the Spanish 1999-2010 PIT Return Panel

Period		1999-2010
Type of microdata		Expanded panel
Scope	Reference population	Personal Income Tax filers
	Geographic scope	All the Spanish Territory except the chartered regions (the Basque Country and Navarre)
Observation unit		Tax return (individual or joint)
Sampling	Type	Minimum variance stratification under Neyman allocation
	Sample income	Sum of gross labour income, net income from other sources and income imputations
	Stratification variables	• Income level (10 levels) • Autonomous Communities (17: 15 of the Common Tax Regime, plus the Autonomous Cities of Ceuta and Melilla) • Source of income (2: labour income represents more or less than half of the total income of the return)

Source: own elaboration

All the comparisons will be made against the Spanish EU-SILC. Its main characteristics are summarized in Table 2[5].

Table 2. Design specifications of the Spanish EU-SILC

Period		**2004-2013**
Type of microdata		Rotating panel (1/4th of renovation of the sample per year) and cross-sectional data
Scope	**Reference population**	Private households and their current members residing in the territory of the countries at the time of data collection
	Geographic scope	All the Spanish Territory
Observation unit		Household and adult individuals
Sampling	**Type**	Two-stage sampling with first stage unit stratification.
	Sample income	-
	Stratification variables	• Autonomous Communities (19) • Municipalities by size (6 levels)

Source: own elaboration

All the comparison and calculations carried out in the following sections exclude the Autonomous Communities of the Basque Country and Navarre, given that they are not included in the Panel because it only offers information of the so called Common Tax Regime (CTR hereinafter).

[5] Detailed information about the Spanish EU-SILC can be found in Encuesta de Condiciones de Vida, Metodología, 2013 [in Spanish]. It is disseminated by the Spanish National Statistics Institute (Instutito Nacional de Estadística, INE) and available free of charge for researchers at http://www.ine.es/daco/daco42/condivi/ecv_metodo.pdf [retrieved 10th September 2014].

3.1. Cross-sectional representativity

Each cross-section of the Panel is representative of the tax filers of the corresponding year. Being or not a tax filer depends on the legislation of each country, but in general low-income earners are not required to file the tax. In Spain filing the tax is compulsory for every person except those who do comply with any of the two rules in Table 3[6].

Table 3. Rules for not being required to file in the Spanish Panel

Rule	Income type	Limit (EUR)	Other conditions
1	Labour income	22,000	One employer
		11,200	Two or more employers
	Capital income Capital gains	1,600	Withheld
	Imputed rental income	1,000	-
	Other	0	
2	Labour income Capital income Rental income Self-employment income Capital gains	1,000	-
	Capital losses	500	-
	Other	0	

Source: own elaboration based on Agencia Tributaria (2010)

This means (rule 1) that low and middle income workers will not have to file if the only additional income they have is small income from capital or secondary dwellings (imputed rental income is exempt for the main dwelling). The rest of the population (notably self-employed and lessors) will have to file unless they obtain insignificant amounts of income (rule 2).

But these restrictions are not as important as they seem for research purposes, because although these individuals are not required to file the tax, they may do it if they want, usually to get a tax refund. Table 4 shows the number of income filers represented in the Panel classifying them according to the rules in Table 3.

[6] These are 2009 rules, but they have been almost constant from 1999 until 2014.

Table 4. Number of taxpayers fulfilling each rule

Rule	Number
Rule 1, labour income ≤11,200	632,073
Rule 1, labour income >11.200 and ≤22,000*	1,139,960
Rule 2	388,173
Other (required to file)	17,152,890
Total	**19,313,096**

Source: own elaboration

* We have no information on the number of employers; therefore some of these taxpayers may have been required to pay if they had two or more employers in 2009.

Although it is clear that the Panel contains all kinds of observations, the problem is that we do not have a direct comparable administrative source to check the representativity of each group. We know that all taxpayers required to file will be correctly represented except for tax fraud and elusion, but for taxpayers not required to file we have only the information of those who voluntarily filed.

A second issue regarding cross-sectional representativity is the observation unit. As it happens in general with tax data, the Panel gives limited information on some individuals and some types of households. The reason is that tax returns in Spain may correspond to individuals (unmarried persons, and married persons who decide to file individually) or married couples (when they choose to file jointly, which is optional). In the former case the Panel provides the sampled tax return and, if the tax filer is married, the tax return of his/her spouse (even if it had not been originally sampled); in the latter case the Panel provides only one tax return without separate information for the two spouses, since all incomes are summed up. As a consequence the only possibility to homogenize observations is to also sum up the incomes of married couples that file individually. The Panel does this by creating a new observation unit called “tax household”, which consist of a married couple or an unmarried individual.

Additionally, the Panel provides some information on some persons not included in the tax household: descendants of the taxpayer under 25 years old and ascendants over 65 years old, provided that in the corresponding year they do not obtain income above 8,000 euro (and therefore they depend economically on the taxpayer). The Panel does not directly provide these data, but it contains information on personal and family allowances that in most cases can be used to know the number of dependent persons.

But even with this additional information, the concept of tax household does not correspond to the usual concept of economic household, since it is limited to an individual or a married couple and some of their dependent descendants and ascendants. In many cases this will match the definition of economic household, but it will not in other cases as when descendants or ascendants do not fulfil the aforementioned characteristics. In these cases several tax households will live together in an economic household, but we do not have information to identify them, so we cannot rebuild the economic household. This does may pose a problem if we are interested in analysing equivalent income using equivalence scales, since tax households are smaller than economic household. This would presumably lead to lower equivalent incomes in tax data in relation to surveys.

In order to analyse the extent of these differences, Table 5 compares the number of households by size in the Panel (tax households) and EU-SILC (economic households) for 2009.[7]

Table 5. Comparison of household sizes (2009)

Household size	Panel	EU-SILC	Panel/EU-SILC (%)
1	6,149,309	3,709,292	165.78
2	4,872,586	4,791,138	101.70
3	2,529,366	3,545,325	71.34
4	1,964,623	3,167,991	62.01
5	345,737	878,907	39.34
6 or more	70,192	350,515	20.03
Total households	**15,931,813**	**16,443,167**	**96.89**
Total individuals	33,515,830	42,719,069	78.46
Average household size	2.10	2.60	80.97

Source: own elaboration

Table 5 clearly shows that tax households are smaller than economic households, as expected (2.10 vs. 2.60 on average). Furthermore, the 33.5 million individuals represented in the Panel are 75.8% of the CTR population (44,205,768 inhabitants[8]), while EU-SILC represents almost the whole CTR population (96.6%).

But to what extent are these limitations of tax data relevant? On one hand, the lack of population representativity may be a problem, although it may be possible to reweight the observations to obtain full representativity if an accurate benchmark is available. On the other

[7] See Appendix for full data on household sizes.

[8] This figure corresponds to the 1st January 2010. We use it because the information on PIT returns is referred to 31st December 2009.

hand, the different concept of household may pose a problem when analysing wellbeing, but not necessary when analysing mobility. Income mobility analysis usually split household income among the members of the economic household. This is useful if we want to analyse wellbeing (as we interpret that all members benefit from all income), but it can be misleading if we want to understand income mobility (as we may be more interested in understanding how individual persons change their economic possibilities, independently of the fact that they live on their own or with other people). If we take this latter point of view, tax households (which relies on the concept of economic independence) may be more useful than the concept of economic household.

3.2. Longitudinal representativity

The sample for the base year of the Panel (2003) was made following the method explained in section 3.3. For subsequent years the observations of each stratum were extracted in two phases. In the first one, the returns corresponding to taxpayer(s) who had been selected in the previous year were also selected, independently of the strata they were located in. In the second phase, additional returns were extracted randomly among the new tax filers of that year until the pre-fixed sample size was achieved. The process for the years prior to 2003 is symmetric. This procedure led to the possibility of following a large number of individuals for long periods: almost 400,000 of the originally sampled taxpayers (64%) are followed from 1999 to 2010 (12 years).

In turn, EU-SILC is a rotating panel. The sample is divided into four sub-samples, called "rotation groups". Every year, one of the subsamples is replaced (25% of renovation rate), so every subsample remains in the Panel for four years and then it is replaced by another one. It means that we can only observe the same households (and their members) for four years: 25,000 households approximately and 50,000 individuals. If our aim is to compare long periods of time (more than 4 years), the main implication of this scheme is that we can only establish comparisons in terms of aggregated results by characteristics of households or individuals. This is an important difference in relation to the Panel, where the same individual can be tracked from the first to the last year (1999-2010). The obvious consequence is that the Panel is a much richer source of information if we are interested in analysing income mobility, since it makes possible to follow a larger number of observations in much longer periods of time.

As important as the quantitative aspects of the attrition are the qualitative ones, so we have to analyse the reasons behind the dropouts of the Panel. One obvious reason is the death of the taxpayer; this is documented in the Panel for both the taxpayer and his/her spouse. Other possible reason is the income level, as explained in section 3.1. This means that for people whose income is always above the thresholds we will have information for all the years of the Panel; we will also have information on people below the thresholds but who always file. But for people that sometimes is above and sometimes below and file when it is compulsory, we will have intermittent information. This means that sometimes we may not be able to analyse the mobility of some individuals between adjacent years, but will be able to do it between non-adjacent years.

The reasons behind the dropouts in EU-SILC are completely different to the ones of the Panel: unknown or untraceable address, death or inability of all members of a household, household that has transferred their residence abroad or to a collective household, or absence or refusal to answer. In addition to dropouts, non-response problems are also an issue in EU-SILC. Non-response can happen both in the household and in the individual questionnaires, mainly due to absence, incapacity or refusal to answer. Besides, the non-response is a cumulative phenomenon along all periods of collaboration, which follows an increasing trend that has to be taken into account when using longitudinal data.

3.3. Income representativity

The Panel offers accurate measures of seven kinds of income: labour income, self-employment income, capital income, rental income, capital gains, income imputations (the main one being imputed income for second dwellings) and income allocation (income from some companies is allocated to their owners). For each kind there is a second level of disaggregation, therefore total income can be split in several tenths of different types. Individual returns for unmarried persons reflect all income obtained directly by them, while individual returns for married persons offer individual income from labour and self-employment and (usually) half of all other sources of income received by any of the spouses. Joint returns, as explained, do not allow to distinguish between spouses.

Surveys usually offer less detailed information, although EU-SILC offers a great level of disaggregation for labour income, pensions and social benefits; yet it does not show great detail for other types of income. Another difference is that labour income, pensions and most

social benefits are individualized, while all other income figures are available only for the household as a whole.

Table 6 compares the level of disaggregation in the two databases for the income concepts that are common to both and that can be grouped in four different concepts. To allow data comparability, we use gross concepts (i.e. before deducting income tax and social security contributions).

Table 6. Disaggregation by income sources

Income type		**Panel**	**EU-SILC***
Labour income and social benefits	Labour income	c1	PY010G
	Pensions		PY100G PY080G
	Unemployment benefits		PY090G
	Other social benefits		PY110G PY120G PY130G PY140G HY060G HY070G HY050G
Income from capital	Capital income	c22 c23 c24 c25 c26 c27 c28	HY090G
	Capital gains	c457	
Rental income		c70	HY040G
Self-employment income		c140 c170 c195	PY050G

Source: own elaboration

* PY refers to personal variables and HY to household variables.

However, the level of disaggregation is useful only if the data collected for each level is reliable. Table 7 compares income for each disaggregation level.

Table 7. Panel and EU-SILC income comparisons, households and household members

		Households with income≠0		Members that belong to households with income≠0	
	Total (EUR)	**Number**	**Average (EUR)**	**Number**	**Average (EUR)**
Panel					
Labour income and social benefits	395,958,887,101	14,579,148	27,159.26	30,918,192	12,806.66
Capital income and capital gains	50,028,603,875	14,148,084	3,536.07	29,982,493	1,668.59
Rental income	17,263,195,895	1,567,190	11,015.38	3,482,354	4,957.34
Self-employment income	29,613,200,361	3,995,306	7,412.00	9,487,975	3,121.13
Total	**492,863,887,235**	**15,925,997**	**30,947.13**	**33,506,485**	**14,709.51**
EU-SILC					
Labour income and social benefits	427,058,722,355	15,677,745	27,239.81	41,451,961	10,302.50
Capital income and capital gains	4,181,344,424	3,602,570	1,160.66	9,294,406	449.88
Rental income	7,180,324,059	994,381	7,220.90	2,622,929	2,737.52
Self-employment income	28,411,731,018	2,099,703	13,531.31	6,614,784	4,295.19
Total	**466,832,121,983**	**16,186,036**	**28,841.66**	**42,719,069**	**10,927.96**

Source: own elaboration

Although the Panel represents fewer households and individuals, the total amount of income is higher (492,864 million euro vs. 466,732). If we observe the different types of income, labour income and social benefits are the best represented types of income in EU-SILC[9], while rental income and especially capital income and capital gains are to a great extent underreported. This may be related not only to underreporting by surveyed households, but also to the fact that most capital income is earned by the richest households, which usually are not surveyed because income is not used as a sampling variable (see section 3.4). Finally, self-employment figures are very similar, but only by chance, since in most cases the figures of the Panel derive from an objective assessment scheme.

Once we have compared the two databases, we check to what extent their results are close to numbers from the official statistics. We can do it only for two variables for which we have aggregate information: wages (from the National Accounts) and all kinds of social benefits (from the Social Security). The Panel does not distinguish between wages and social benefits, so we only compare the aggregate figure. Table 8 shows the results.

[9] However it has to be noted that some imputations and adjustments are done by INE so that aggregated figures are more accurate, dealing for instance with the non-response issue.

Table 8. Income comparisons between the microdata and official statistics

	Official Statistics[a]	Panel		EU-SILC	
	Total (EUR)	Total (EUR)	% of official statistics	Total (EUR)	% of official statistics
Wages[b]	372,632,478,060	n.a.	-	311,292,826,999	83.5
Social benefits[c]	242,233,265,175	n.a.	-	114,011,322,255	47.1
Total[d]	614,865,743,235	395,958,887,101	64.4	425,304,149,254	69.2

Source: own elaboration

a. The Basque Country and Navarre have been excluded based on the proportion of earners that live in each territory, since we do not have detailed information for Autonomous Communities.

b. Source for Official Statistics: National and Regional Accounts, National Statistics Institute (INE): http://www.ine.es/daco/daco42/cre00/b2008/dacocre_base2008.htm

c. Source for official Statistics: Social Protection Statistics, National Statistics Institute (INE): http://www.ine.es/daco/daco42/cre00/b2008/dacocre_base2008.htm

d. Official Statistics do not include private pensions, so they have been excluded from EU-SILC. They have not been excluded from the Panel because it is not possible to distinguish them.

Even though aggregated total wages are higher in EU-SILC than in PIT (as we saw in table 7), we see that they are still far away from the official statistics. Social benefits are much lower in terms of comparability with official statistics, representing less than 50%. In global terms, EU-SILC represents 69.2% of this type of incomes, while the Panel represents 64.4%.

3.4. Representativity by groups and geographical location

The Panel was sampled using minimum variance stratification under Neyman allocation. The first step under this methodology is to choose stratification levels, which are three: income level (10 brackets), geographical location (Autonomous Communities), and the main source of income (labour income or other). The number of final strata is 340 (17x2x10), and the use of stratification means that each dataset represents correctly the income of the population within each of those strata, with a similar level of error. This is attained by sampling more (less) observations when the income variance of the corresponding stratum is higher (lower).

As we said before, the EU-SILC sampling is performed as a two-stage sampling. The first stage units are the census sections, while those of the second stage are main family dwellings. Stratification sampling is done in the first stage, under a geographical criteria (Autonomous Communities and municipalities). It implies that there are 140 strata (19 Autonomous Communities and 6 types of municipalities regarding their size). Finally, 16.000 dwellings distributed in 2,000 census sections are sampled.

While the main source of income and geographical aspects may only be relevant when studying mobility within those strata, the stratification by income level is crucial to get a good picture of all income levels. This is not a problem for average income levels, but it is for high income earners: due to the small number of this type of individuals and the high variance of their incomes, a sampling method that does not stratify by income will pose problems on analysing mobility to/from/within the top income quantiles. The stratification of the Panel by income levels is especially useful because three of the ten levels are above 60,000 euro, thus giving a good representation to observations that otherwise would have probably been kept out of the sample. In surveys like EU-SILC it is not possible to take income into account when sampling, because the samples are designed using only geographical and census information, and not income information.

4. Results

In this section we try to replicate the calculations carried out by Bárcena and Moro (2013) using the Spanish EU-SILC for 2004-2009. We have chosen some of the mobility indices they define (transition matrices and the Shorrocks index for relative mobility; Fields and Ok index for absolute mobility; and the correlation coefficient, the regression slope and the Hart index for income correlation) and have calculated them using the Panel. We have also reproduced, as possible, the methodological decisions they took. This means basically that we have analysed income mobility for pairs of adjacent years, we have followed the individuals living in the same household for those pairs of years, and we have calculated the inflation-adjusted equivalent disposable household income and allocated it to each taxpayer. As Bárcena and Moro, disposable household income has been measured as the sum of all sources of income minus taxes and social security contributions using the available variables in the Panel[10]. For equivalising incomes we have also used the modified OECD equivalence scale, but assuming that all descendants are above 14 years old (we do not know their age in the Panel). Of course conceptual differences of the databases remain, like the population represented and the household concept, as explained in section 3.

Table 9 shows our results for all the pairs of adjacent years, while Table 10 offers the analogous results for EU-SILC taken from Bárcena and Moro (2013)

[10] The specific income variables used for each year can be found in Table 17 (first column) in the Appendix.

Table 9. Mobility indices (Panel)

	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009
Correlation coefficient	0.8491	0.8140	0.8079	0.8474	0.8747
Regression slope	0.8971	0.9135	0.7926	0.8561	0.8711
MHart	0.1803	0.1894	0.2143	0.2218	0.2286
F-Ok	0.2201	0.2372	0.2470	0.2548	0.2518
Ms	0.5068	0.5183	0.5326	0.5336	0.5069
Stays in the same decile (%)	54.39	53.35	52.07	51.97	54.38
One decile up (%)	14.12	14.78	17.02	19.71	15.13
One decile down (%)	15.79	15.44	13.71	12.43	15.40
Two deciles up (%)	4.42	4.81	5.35	4.91	3.74
Two deciles down (%)	4.28	4.20	4.01	3.99	4.90
More than two deciles up (%)	3.70	4.29	4.41	3.58	2.93
More than two deciles down (%)	3.3055	3.1342	3.4323	3.4157	3.5166

Source: own elaboration

Table 10. Mobility indices (EU-SILC)

	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009
Correlation coefficient	0.768	0.74	0.747	0.733	0.743
Regression slope	0.708	0.686	0.671	0.754	0.615
MHart	0.282	0.306	0.307	0.397	0.397
F-Ok	0.287	0.294	0.285	0.304	0.322
Ms	0.752	0.753	0.729	0.73	0.727
Stays in the same decile (%)	31.7	31.7	34.1	34.4	34.6
One decile up (%)	17.8	17	18.8	15.9	16
One decile down (%)	15.5	15.6	13.5	16.8	17.3
Two deciles up (%)	9.4	8.9	9.8	7.5	7.3
Two deciles down (%)	7.5	7.6	6	7.8	7.8
More than two deciles up (%)	10.5	9.8	11.3	8.7	7.8
More than two deciles down (%)	7.6	9.4	6.6	8.8	9.2

Source: Bárcena and Moro (2013)

The main conclusion we get from comparing the two tables is that the indices calculated with the Panel always report higher immobility than the ones calculated with EU-SILC. This may be due to the fact that the individuals with more fluctuating incomes may file PIT intermittently, therefore they become excluded from the Panel when they do not file. Besides, we observe that the patterns of evolution are also different, in terms of studying when mobility rises or decreases; the graphs 1 to 10 in the Appendix show a quick overview of this result.

In order to determine to what extent these differences are driven by the underlying differences between the databases, we run a double sensitivity analysis for the income concept used (since we have seen that the aggregate income figures in the tax data are quite

different to those in surveys) and the equivalence scales (as we can only approximate the OECD modified scale when using tax data). This leads to twelve ways of calculating individual income combining three income concepts[11] and four equivalences scales, as shown in Table 11 (the variable we used in tables 9 and 10 is Income14).

Table 11. Definitions of individual income depending on the income concept and the equivalence scale

Income definitions / **Equivalence scales**	All sources of income	All sources of income, except capital income	All sources of income, except capital and rental income
Equally distributed among taxpayers	Incom_ 11	Incom_ 21	Incom_ 31
OECD modified scale, assigning 0.3 to all members younger than 25 years old (assuming they are under 14)	Incom_12	Incom_22	Incom 32
OECD modified scale, assigning 0.4 to all members younger than 25 years old (intermediate option)	Incom_13	Incom_23	Incom_33
OECD modified scale, assigning 0.5 to all members younger than 25 years old (assuming they are above 14)	Incom_14	Incom_24	Incom_34

Given that the results for all the variables and all the pairs of years would be quite extensive to reproduce, we will show only the results for the first pair of years (2004/2005), comparing the results for different concepts of income for a fixed equivalence scale, and then the results for different equivalence scales for a fixed income concept. We have checked that the results of the sensitive analysis are very similar for all pairs of years, all income concepts and all equivalence scales.

Table 12 shows the results of comparing the impact in terms of mobility of using different definitions of income, keeping the equivalence scale fixed. We see that some indices increase while other decrease, but they all go in the same direction: removing capital income, or capital plus rental income, increases mobility. This can be explained because capital gains, which are only considered in the first definition of income, introduce false mobility because of their volatility, as they are only declared when assets are sold.

[11] The specific income variables used for each year can be found in Table 17 in the Appendix.

Table 12. Mobility indices for three definitions of income (Panel, 2004-2005)

	Incom_14	Δ Incom_24-Incom_14	Δ Incom_34-Incom_14
Correlation coefficient	0.8491	7.2%	7.1%
Regression slope	0.8971	3.8%	3.5%
MHart	0.1803	-12.5%	-11.2%
F-Ok	0.2201	-14.3%	-14.7%
Ms	0.5068	-6.7%	-8.5%
Stays in the same decile (%)	54.39	5.6%	7.1%
One decile up (%)	14.11	2.4%	-0.9%
One decile down (%)	15.79	-9.0%	-11.0%
Two deciles up (%)	4.41	-6.6%	-8.1%
Two deciles down (%)	4.27	-5.8%	-5.9%
More than two deciles up (%)	3.69	-22.3%	-22.5%
More than two deciles down (%)	3.30	-18.6%	-16.9%

Source: own elaboration

In Table 13 we check the sensibility to the introduction of different equivalence scales. Referred to the first column (where actually no equivalence scale is used, as income is equally distributed only among taxpayers), we observe an increase in mobility when equivalence scales are introduced. We also see that the highest the weight for members of the household older than 25 years, the highest the mobility turns out to be. The reasons behind these results are not straightforward, as there are many ways that scales can affect mobility results.

Table 13. Mobility indices for four definitions of equivalence scales (Panel, 2004-2005)

	Incom_11	Δ Incom_12-Incom_11	Δ Incom_13-Incom_11	Δ Incom_14-Incom_11
Correlation coefficient	0.8620	-0.9%	-1.2%	-1.5%
Regression slope	0.9063	-0.5%	-0.8%	-1.0%
MHart	0.1767	2.3%	2.2%	2.0%
F-Ok	0.2066	4.0%	5.3%	6.6%
Ms	0.4738	5.3%	6.5%	7.0%
Stays in the same decile (%)	57.35	-3.9%	-4.8%	-5.2%
One decile up (%)	13.26	7.2%	7.7%	6.4%
One decile down (%)	15.05	4.5%	5.2%	4.9%
Two deciles up (%)	3.89	7.4%	9.4%	13.5%
Two deciles down (%)	3.65	9.1%	13.8%	17.1%
More than two deciles up (%)	3.60	0.8%	1.9%	2.7%
More than two deciles down (%)	3.18	-0.6%	1.2%	3.9%

Source: own elaboration

Summarizing, the closest results to EU-SILC are the ones calculated with all the sources of income and the OECD modified scale assuming all members younger than 25 years old to be

older than 14 (Incom_14). Results with all the other definitions of income show higher mobility, but the differences are small and tax data always show higher immobility than survey data.

Finally, Table 14 summarizes the differences in mobility found between the Panel and EU-SILC.

Table 14. Summary of results of income mobility: Panel vs. EU-SILC

Index		Diff. (Panel - EU-SILC)	Range of differences	Closest result	Years with smaller differences	Years with larger differences
Correlation	Correlation coefficient	+	11%/20%	Incom_14	06/07	08/09
	Regression slope	+	27%/33%	Incom_14	06/07 07/08	08/09
	MHart	-	-0.36%/-0.46%	Incom_14	06/07	07/08 08/09
Absolute	F-Ok	-	-0.23%/-0.39%	Incom_14	06/07	04/05 08/09
Relative	Ms	-	-0.33%/-0.43%	Incom_14	06/07	04/05 08/09
	Stays in the same decile	+	25%/29%	Incom_14	04/05	08/09
	Moves to other deciles	-	5%	Incom_14	04/05	08/09

Source: own elaboration

As said before, all the indices report less mobility with the Panel than with EU-SILC. The correlation coefficient and the regression slope are higher when using tax data, in a range that varies from 11% to 20% and from 27% to 33%, respectively, depending on the definition of income and equivalence scale used. On the other hand, the Hart, Fields and Ok and Shorrocks indices show a lower value using tax data than using a survey, although the range of differences is much smaller. Finally the deciles show in tax data a much higher share of individuals staying in the same decile.

5. Conclusions

Throughout this paper we have carried out a comparison between survey data and administrative tax data in order to identify relevant differences for mobility analysis. Our main conclusion is that both kinds of microdata present merits and shortcomings, but those are different in each case, making each data source suitable for different purposes. Taking the Spanish data as an example, we find that while both types of microdata offer a good

representation of middle income households, surveys underrepresent the upper part of the income tail while tax data underrepresent the lower part. Regarding income accuracy, surveys offer more detailed information on wages, pensions and social benefits (though tax data is also accurate in aggregate terms), while capital and rental income is much better in tax data (this kind of data in surveys is almost unusable). Another relevant difference is the unit of analysis: while in surveys we have information on households and its members, in tax data we have only information of tax units (basically married couples) and persons depending on them. We think that those differences do not lead to choose one data source over the other, but each of them will be more adequate for different research purposes.

After analysing the characteristics of both sources we have compared the differences in mobility results. We have compared our results with the recent work by Bárcena and Moro (2013) after homogenizing as much as possible the methodologies, running some sensitivity analysis for some relevant variables. What we find is that the differences between the results are systematic, showing the tax data higher immobility for all the indices calculated. We also find that the results are only slightly sensitive to the choice of the income concept and equivalence scale. Our conclusion is that although both results measure mobility, they are not directly comparable, meaning that results of income mobility analysis should only be compared to other results using the same type of data.

References

Acs and Zimmerman, 2008; Acs, G. and Zimmerman, S. (2008) "Like Watching Grass Grow? Assessing changes in U.S. Intragenerational Economic Mobility Over the Past Two Decades." Washington, D.C.: Pew Charitable Trusts, Pew Foundation Economic Mobility Project, 2008.

Agencia Tributaria (2010), *Manual Renta 2009*, Agencia Estatal de Admnistración Tributaria.

Auten, G.E.; Gee, G. (2007). "Income mobility in the U.S.: evidence from income tax returns for 1987 and 1996", U.S. Department of the Treasury OTA Papers, 99.

Auten, G.E.; Gee, G. (2009). "Income Mobility in the United States: New Evidence from Income Tax Data", National Tax Journal, LXII(2):301-328

Auten, G.E.; Gee, G. and Turner, N. (2013) "Income Inequality, Mobility, and Turnover at the Top in the US, 1987-2010". The American Economic Review 103.3 (May 2013): 168-172.

Ayala, L.; Onrubia, J. (2001). "La distribución de la renta en España según datos fiscales", Papeles de Economía Española, 88:89-112.

Ayala. L.; Sastre, M. (2005). "La movilidad de ingresos en España", Revista de Economía Aplicada, 38(XIII), 123-158.

Bárcena-Martín, E.; Moro Egido, A.I. (2013). "Movilidad de ingresos en España: el efecto de la crisis", Papeles de Economía Española, 135.

Bradbury, K.; Katz, J. (2002a) "Are Lifetime Incomes Growing More Unequal? Looking at New Evidence on Family Income Mobility." Federal Reserve Bank of Boston Regional Review 12 No. 4 (September, 2002): 3–5.

Bradbury, K.; Katz, J. (2002b) "Women's Labor Market Involvement and Family Income Mobility When Marriages End." New England Economic Review No. 4: 41–7.

Cantó, O. (2000). "Income mobility in Spain: how much is there?", Review of Income and Wealth, 46:85-102.

Carroll, R.; Joulfaian, D.; Rider, M. (2006). "Income Mobility: The Recent American Experience", Andrew Young School of Policy Studies Research Paper Series 07-18.

Cervini-Plá, M. (2011) Intergenerational earnings and income mobility in Spain. Munich Personal RePEc Archive, nº 34942

Corak, M., Heisz, A., 1999. The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data. Journal of Human Resources 34 (3), 504–556.

Dragoset, L.M.; Fields, G.S. (2008) "U.S. Earnings Mobility: Comparing Survey-Based and Administrative-Based Estimates" Unpublished paper. Available at http://digitalcommons.ilr.cornell.edu/workingpapers/88/

Finnie, R.; Sweetman, A. (2003) "Poverty Dynamics: Empirical Evidence for Canada," Canadian Journal of Economics, 36(2), 291–325, 2003.

Formby, J. P.; Smith, W. J.; Zheng, B. (2004) "Mobility measurement, transition matrices and statistical inference". Journal of Econometrics 120, 181–205.

Gil, M., De Pablos, L., Martínez, M. (2010) **"**Los determinantes socioeconómicos de la demanda de educación superior en España y la movilidad educativa intergeneracional", *Hacienda Pública Española/ Revista de Economía Pública,* nº 193 (2/2010) pp. 75-108.

Gradín, C., Cantó, O., Del Río, C. (2008) "Inequality, poverty and mobility". Investigaciones económicas, vol xxxii (2)

Grusky, D.B.; Mitnik, P.A., Wimer, C. (2001) ‘‘Measuring Intergenerational Economic Mobility with Tax-Return Data Towards an IRS Platform. A Proposal to the Economic Mobility Project of the Pew Charitable Trusts". Stanford Center for the Study of Poverty and Inequality.

Jäntti, M.; Jenkins, S. P. (2013) "Income mobility", SOEP papers on Multidisciplinary Panel Data Research, No. 607.

Mazumder, B. (2005) "Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data". Review of Economics and Statistics 87 (2), 235–255.

Onrubia, J.; Picos, F.; Pérez, C. (2011). Panel de Declarantes de IRPF 1999-2007: diseño, metodología y guía de utilización, Instituto de Estudios Fiscales.

Onrubia, J.; Picos, F.; Pérez, C.; Gallego, C.; González, M.C.; Huete, S. (2012). "Panel de declarantes de IRPF 1999-2008: Metodología, estructura y variables", Documentos de Trabajo del Instituto de Estudios Fiscales, 12/2012.

Osterberg, T. (2000) "Intergenerational income mobility in Sweden: What do tax-data show?" Review of Income and Wealth Series 46, Number 4, December 2000.

Pascual, M. (2009) "Intergenerational income mobility: The transmission of socio-economic status in Spain". Journal of Policy Modeling 31, pp. 835–846.

Pérez, C.; Gallego, C.; Huete, S.; Pradell, E. (2013). "Panel de Declarantes de IRPF 1999-2009: metodología, estructura y variables", Documentos de Trabajo del Instituto de Estudios Fiscales, 13/2013.

Pérez, C.; Villanueva, J.; Burgos, M.J.; Pradell, E.; Gallego, C. (2014). "Panel de Declarantes de IRPF 1999-2010: metodología, estructura y variables", Documentos de Trabajo del Instituto de Estudios Fiscales, 9/2014.

Piketty, T.; Saez, E. (1998) "Income inequality in the United States, 1913–1998" The Quarterly Journal of Economics. Vol. CXVIII February 2003 Issue 1

Piketty, T.; Saez E. (2007) "How progressive is the US Federal Tax System? A historical and international perspective". Journal of Economic Perspectives, 21 (1): 3-24.

Solon, G. (1992) "Intergenerational income mobility in the United States". American Economic Review 82 (3), 393–408.

Splinter, D.; Diamond, J.; Bryant, V. (2009). "Income Volatility and Mobility: U.S. Income Tax Data, 1999-2007" Proceedings of the 102nd Annual Conference of the National Tax Association.

US Department of the Treasury, Office of Tax Analysis (1992a). "Household Income Changes over Time: Some Basic Questions and Facts", Tax Notes, 56:1065–1074.

US Department of the Treasury, Office of Tax Analysis (1992b). "Household Income Mobility During the 1980s: A Statistical Assessment Based on Tax Return Data", Tax Notes, 55 (special supplement)

APPENDIX

Table 15. Number of tax households by composition (population figures)

		Number of tax filers >25 + dependent persons over 65						
		0	1	2	3	4	5 or more	Total
Number of tax filers <25 and dependent persons <25	0	-	5,673,127	3,818,802	78,520	6,610	93	9,577,151
	1	476,183	1,024,257	1,911,431	34,095	4,350	127	3,450,442
	2	29,528	532,842	1,843,594	27,077	4,852	113	2,438,005
	3	6,574	79,029	304,147	4,298	605	43	394,696
	4	1,295	9,938	42,609	425	157	0	54,423
	5 or more	188	2,635	14,155	112	8	0	17,097
	Total	513,767	7,808,395	7,961,577	144,795	16,671	375	15,931,813

Source: own elaboration

Table 16. Number of economic households by composition (population figures), EU-SILC

		Number of adults > 25						
		0	1	2	3	4	5 or more	Total
Number of members of the household < 25	0	-	3,612,412	4,375,207	1,237,897	405,056	79,856	9,710,428
	1	96,880	392,336	2,058,900	544,467	123,528	30,555	3,246,665
	2	23,596	232,588	2,171,817	226,587	72,456	10,176	2,737,220
	3	15,940	44,646	448,498	62,809	21,639	4,579	598,111
	4	2,005	8,606	66,096	13,650	13,036	7,516	110,910
	5 or more	1,317	5,620	27,938	3,804	467	688	27,439
	Total	139,738	4,296,208	9,148,456	2,089,214	636,182	107,355	16,443,167

Source: own elaboration

Table 17. Income concept use in the calculations of the mobility indices[a]

Year	Incom_1*	Incom_2*	Incom_3*
2004	c1 - c9 + c28 + c470 + c477 + c80 + c140 + c170 + c199 - c737	c1 - c9 + c80 + c140 + c170 + c199 - - taxratio2*c737	c1 - c9 + c140 + c170 + c199 - taxratio3* - c737
2005	c1 - c9 + c28 + c470 + c477 + c80 + c140 + c170 + c199 - c737	c1 - c9 + c80 + c140 + c170 + c199 - - taxratio2*c737	c1 - c9 + c140 + c170 + c199—taxratio3*c737
2006	c1 - c9 + c28 + c470 + c477 + c80 + c140 + c170 + c199 - c737	c1 - c9 + c80 + c140 + c170 + c199 - - taxratio2*c737	c1 - c9 + c140 + c170 + c199—taxratio3*c737
2007	c1 - c10 + c69 + c29 + c457 + c140 + c170 + c195 - c741	c1 - c10 + c69 + c140 + c170 + c195 - taxratio2*c741	c1 - c10 + c140 + c170 + c195 - taxratio3*c741
2008	c1 + c70 + c29 + c457 + c140 + c170 + c195 - c741 - c10	c1 - c10 + c70 + c140 + c170 + c195 - taxratio2*c741	c1 - c10 + c140 + c170 + c195 - taxratio3*c741
2009	c1 + c70 + c29 + c457 + c140 + c170 + c197 - c741 - c10	c1 - c10 + c70 + c140 + c170 + c197 - taxratio2*c741	c1 - c10 + c140 + c170 + c197 - taxratio3*c741

a. taxratio2 represents the ratio between the income variables included in Incom_2* relative to the income variables included in Incom_1*. taxratio3 does the same for Incom_3*. This is a way to allocate a share of the tax liability when we only take part of the income into account.

Graphs 1 to 10. Comparison of the evolution of the mobility indices (Panel Incom_14 vs. EU-SILC)

Panel Incom_14 (own calculations) | EU-SILC (Bárcena and Moro, 2013)