Integration of income administrative data into the Portuguese household income distribution: a first national experience using employees' income tax data

Eduarda Góis[1], Carlos Farinha Rodrigues[2], David Leite[3], Daniel Gomes[1], Maria Manuel Pinho[1]

[1]Statistics Portugal, Portugal

[2]ISEG - Lisboa School of Economics & Management, Portugal

[3]Paris School of Economics

Abstract

The Portuguese results on poverty and economic inequality are based on a 4-year longitudinal sampling survey of households and their members, carried out every year, which is part of the EU-SILC programme since 2004. The survey collects data on qualitative aspects (e.g. health, housing and material and social deprivation) where alternative sources are difficult to find, and on quantitative aspects related to monetary income where, on the contrary, alternative sources are known, namely tax and social security data. Until 2021, monetary income data, including employees' income, were obtained exclusively through direct collection from the selected households and individuals, with proxy responses being accepted in situations of individual temporary absence or in incapacity, and frequently without consulting information organised for tax purposes, even though the questionnaire includes the possibility of responding by transcribing data from the annual tax return: that, combined with sampling weighting, increases the possibility of deviations from nearly exhaustive administrative data. From 2022 onwards (2021 income), considering that the integration between personal income tax data (IRS) and survey data can be ensured for most incomes, even though not covering for taxpayers exempt from submitting the annual return, and taking advantage of previous studies regarding other data collections about wages and salaries, the survey started to integrate administrative data on employees' income collected by IRS Model 3, Annex A, in order to improve the consistency and quality of information before deduction of taxes and social contributions. Overall, compared to data relying exclusively on the original survey data, the new income distribution that includes imputation of administrative data is more homogeneous, with both a lower average employees' income and the corresponding standard deviation. The adjustment also has a significant impact on employees' income deciles, with a significant change in the first six deciles, and there is evidence that the incorporation of administrative data increases income inequality.

Keywords: monetary income, employees' income, administrative data, economic inequality

1. Introduction

The EU statistics on income and living conditions (EU-SILC) aims to collect timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion, and living conditions. EU-SILC a is collection of data on the living conditions and income distribution

of households and individuals, which is regulated, and as such harmonised, at the European level in terms of concepts, validation and methodology for compiling the results.

In Portugal, SILC has been carried out since 2004, targeting all individuals living in the Portuguese territory during the income reference period. The sample of the national SILC uses a complex design, including stratification by NUTS level 2 and multi-stage sampling i.e. the selection of units in several stages: in the first stage, areas are selected based on the INSPIRE grid cells, and, in the second stage, dwellings are selected by area. In order to allow for the estimation of both cross-sectional and longitudinal indicators (for example, the persistence of poverty as well as the at-risk-of-poverty rate), the PT-SILC sample is based on a 4-year rotational subsample scheme. Data are directly collected through CAPI (face-to-face computer-based interviewing) while CATI (computer-based telephone interviewing) is also available since the COVID-19 pandemic. To ensure the extrapolation of sample data to the populations under study, different types of weights, which include a compensation for non-response and calibration, are applied in all statistical outcomes, whether they are cross-sectional or longitudinal, and concern individuals or households. The income reference period of SILC for a specific year is the previous calendar year.

Our paper addresses the problem of the sensitivity of the income distribution when resulting from survey data. The main argument is that in comparison to sociodemographic qualitative data, the information on income is particularly sensitive, implying lower data robustness. The suggested strategy to tackle this problem is the appropriation of income administrative data provided by the tax authority, which brings an opportunity to enhance the quality of income official statistic but comes with some challenges.

This paper is structured as follows. The following section discusses the problem at hand and section 3 describes the strategy used to tackle the problem. Finally, the corresponding empirical outcome is presented and discussed.

2. The sensitivity of income data

SILC collects a wide range of variables associated to objective dimensions such as monetary income components, and to subjective dimensions such as material and social deprivation, labour status, health status, housing conditions and social exclusion. In Portugal, longitudinal data is collected along four consecutive years for each sub sample. Over the years, it has become clear that there is an increased difficulty in PT-SILC in keeping respondents motivated. At the same time, the number of proxy answers – personal interview with another member of the household – is significant (around 40% in 2023). There is therefore an

increasing concern about the risk of obtaining a reduced number of responses and of not keeping the desired quality standards.

Regarding income data, annual change rates on survey income data have been concluded to be too high (specially for employees' income) in comparison to other sources (for example, national accounts and social policy data). On top of this, income data collection is particularly complex, with the survey including alternative questions to facilitate the answer: one possibility, the preferable way, is to base the answers on the employees' income tax form or on the withholding certificate; alternatively, the answer can be provided without documentation, but this implies going over several questions to collect the same data. Another problem respects the presumed difficulty of individuals in distinguishing gross and net income amounts.

3. On the linkage between data and tax data

Current orientation in European statistics is to encourage the use of administrative data. In fact, considering the experience in official data collection, the European Statistical System mentioned in the Wiesbaden Memorandum (2011, No. 5, b) the use of administrative data as a key factor for the development of European social studies. Also, specifically for distributive income assessments (DIA), Member States are encouraged 'to combine survey data and administrative data when doing DIAs' (European Commission, 2022: 7).

Using administrative data to produce official statistics has gained relevance over the years (European Commission, 2021). In spite of the administrative data being collected for purposes other than statistical production, they make it possible to reconcile the growing and increasingly refined demand for statistical information with the pressure on statistical authorities to increase the process efficiency (Eurostat, 2013). Using administrative registers in the production of official statistics implies lower costs (surveys and censuses are expensive and labour-intensive), less burden on the respondent (the same information is not required for different purposes), better coverage (more comprehensiveness, no sampling errors, and less non-response) and higher frequency (potential lower lag between the time of reference of the information and that of dissemination).

The use of administrative income tax data has become more frequent as the reliance on income data from household surveys has been questioned. The problem of survey under-coverage of top incomes was the first to be addressed. There was an increasing awareness that household surveys may fail to capture incomes at the top of the distribution – there are issues of sparseness at the top of the distribution (sampling errors) and underreporting and lower survey participation of the richest households (non-sampling errors). Consequently, both inequality levels and trends over time may be mis-measured. Two main approaches have been

used to tackle the under-coverage of top incomes: replacement methods and reweighting methods (Carranza et al., 2021). UK official statistics on income distribution have incorporated top-income adjustments to household survey data since 1992 (Jenkins, 2022; Webber et al., 2020) and exercises on top-income adjustments have also been applied to Portugal: Carranza et al. (2021), using a reweighting method, conclude for an increase in inequality but a decrease in the average income for the 2006-2017 period and Hlasny and Verme, for 2018, also using a reweighting method, conclude for an increase in inequality with the correction across EU MS being positively associated with the mean income and the non-response rate, but, using a replacing method, the authors obtain unclear results.

So, why not use tax administrative data more extensively (and not only for top incomes)?

Over the implementation of EU-SILC, an increasing number of countries have been combining survey data with some administrative data, with the extent and nature of the use of administrative data varying widely. Few countries, 'old register countries', already rely completely on administrative data (the leading examples are the Nordic countries with a long history of using comprehensive registers). Other countries have registers but greater concerns about privacy and national identity numbers, in a way that linked registers are not widely used, and that is also the belief that income tax data may not cover the bottom half well by comparison with household surveys. In fact, the literature suggests that administrative data tend to perform better for top-level income and that survey data tend to perform better for bottom-level income. Many countries, especially poorer ones, do not have suitable registers on income distribution, which means that the use of household surveys is inevitable (Carranza et al., 2021; Jenkins, 2022).

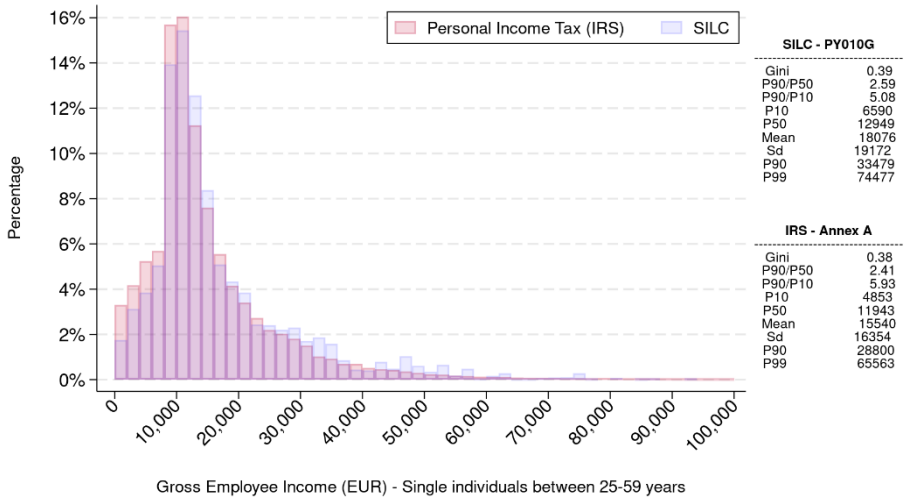
Instead of simply using income tax data, an alternative strategy is to link administrative records to survey respondents and replace survey income responses with the administrative ones, assuming that the linked data are of better quality than the survey responses. However, combining information from surveys and tax data is challenging in that the two sources mostly employ different income concepts and income recipient units.

Aiming at improving the income components of the survey, Statistics Portugal decided to begin crossing the data on employees' income from the survey with the tax authority administrative files in 2022 (2021 income). In 2023 (2022 income), PT-SILC already benefited from some refinements in the data transmitted by the tax authority and this articulation is expected to improve in the upcoming years.

Figure 1 shows the distribution of the 2022 employees' income collected both from the survey and from tax administrative data. The picture constrains data to single persons and to the 25-59 age group with the purpose of showing the overall similarity between the two series.

Considering this context, the figure shows that survey data tend to underestimate the lowest income levels. As explained below, if couples decide to file a joint tax return, there is, at this point, an additional complexity in comparing these two data sources, and for that we decided not to include that data in the picture.

Figure 1: Gross employees' income 2022 – survey and administrative data comparison



Yet, using administrative data comes with challenges. We emphasize two relevant ones. First, there is a conceptual difference between SILC's private household definition and the tax household definition. Second, the lag between the time of the survey data collection (second quarter of the year) and the income tax reference period (year n-1) may lead to differences in the household composition (for example, in cases where individuals leave the household during the first part of the year, their income will be excluded from the survey, but included in the tax authority's database). Also, an extreme value in the administrative records has a different meaning in the sample as it is extrapolated to the population, so that the appropriation must be done with cautious.

Matching the survey data and tax administrative data is a key step in the integration process. Although Portuguese residents have a unique tax identification number, before 2024 this was not collected in the survey, which means that direct linkage is not possible. Also, matching data requires that both sources share a set of key characteristics (for example, location, age, type of income), which can be used to associate a sample unit to the external source (longitudinal record data can further increase the possibilities of identification) through several iterations. If, in the end of the process, the two sources still do not have identical values for the key variables, the integration fails.

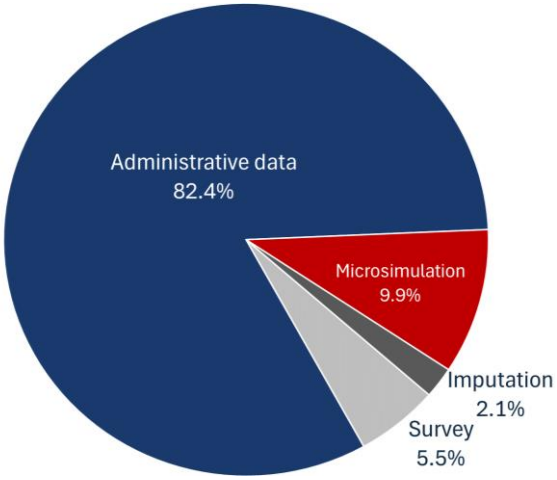
Our strategy tackles, for the time being, only employees' income and comprises three main steps. The first one implies selecting employees' income only for records with social

contributions from the personal income tax data (IRS Model 3 – Annex A). Up to 2022, data received from the tax authority did not distinguish between employees' income and pensions as these income sources are jointly displayed in IRS Model 3 – Annex A. From April 2024 onwards (for income tax declarations of 2022), it is already possible to separate those employees' and pensions income, with which we expect to increase the quality of the outcome of this appropriation process. If couples decided to file a joint tax return, the separation between employees' and pensions income is based on the survey's structure. The second step consists in calculating the share of each person's income in PT-SILC for the distribution of labour income from tax data by household members, also expected to be improved in the near future once we have access to tax data split by household member. Finally, employees' income is appropriated from the tax authority data only when the individual has reported employees' income in the survey.

4. The impact on PT-SILC income data

As mentioned before, the empirical strategy described above was applied to both 2021 and 2022 SILC employees' income, but the results reported here respect 2022 income data (2023 SILC). Prior to the appropriation of the tax income data, records were subject to an outlier detection and trimming process. The integration of tax administrative data – impacting PT-SILC PY010G variable – changed 82.4% of the 2022 income survey data (Figure 2). The remaining records were either subject to microsimulation for net-gross conversion (9.9%) or to imputation from the previous year or from a donor (2.1%). The other 5.5% kept the original survey data.

Figure 2: Gross employees' income (PY010G) 2022 – population impact indicators

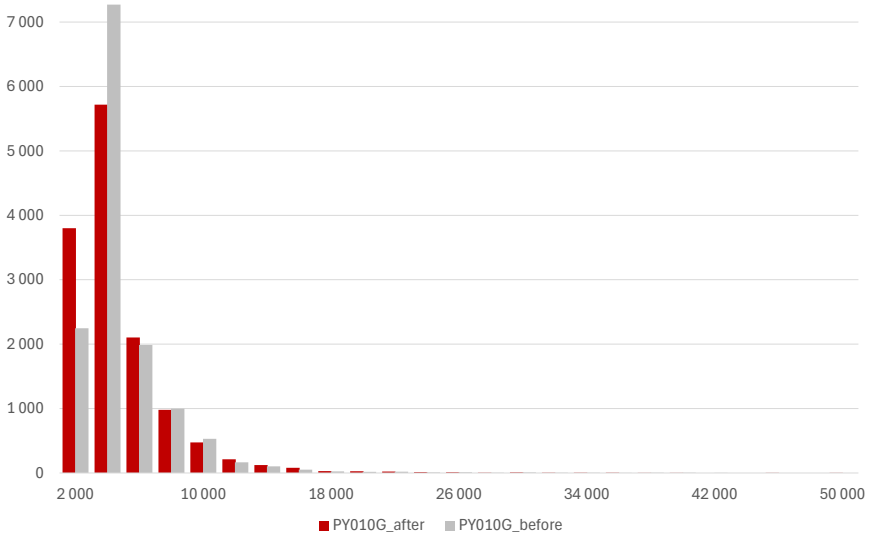


Change rate		% of individuals
Decrease	> 10%	51.4
	≤ 10%	16.9
Increase	≤ 10%	10.7
	> 10%	21.0

As for the income distribution, the appropriation led to an increase in the homogeneity with lower employees' income mean and standard deviation (Figure 3 and Figure 4). This impact of the appropriation is particularly relevant for low-income classes – up to 2,000 euros, there is an increase in the number of individuals, with the opposite occurring in the following class (from 2,000 to 4,000 euros), suggesting that survey respondents tend to underestimate very low income levels.

The analysis of the impact of the administrative data appropriation must take into account that, prior to that integration, a microsimulation technique was applied to raw survey data aiming at converting reported net amounts into gross amounts.

Figure 3: Gross employees' income (PY010G) class distribution 2022



Furthermore, the integration of administrative data significantly impacted employees' income deciles, particularly in the first six deciles, and the inequality of the distribution. Comparing the original employees' income deciles with the final ones:

- 39.8% of the individuals remained in the decile;
- 29.6% of the individuals were assigned to a lower decile;
- 30.6% of the individuals were assigned to a higher decile.

Figure 4: Gross employees' income (PY010G) percentile distribution 2022

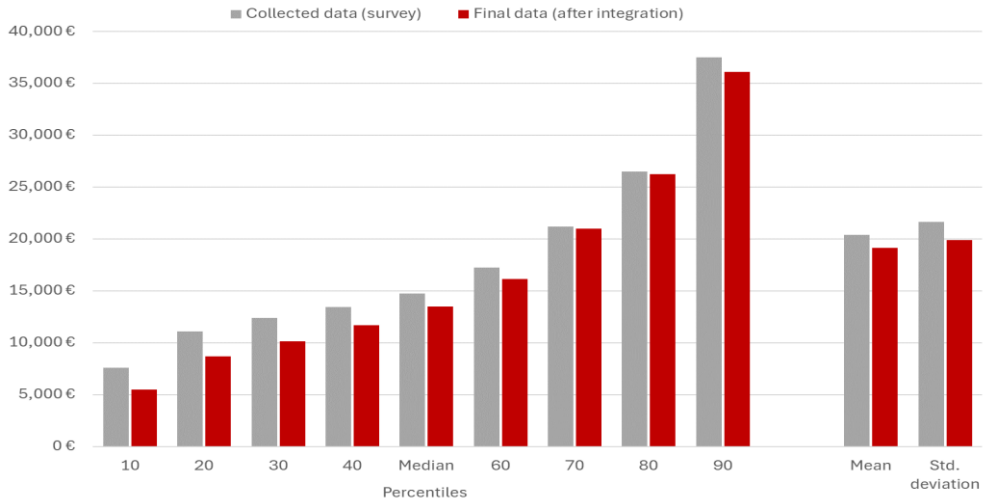


Table 1 displays the statistics on the distribution before and after the appropriation of administrative data, with both unweighted and weighted data. There is evidence that the incorporation of administrative data increases income inequality based on the results of the Gini index and the S80/S20 and S90/S10 ratios. The distribution becomes slightly more concentrated on lower income levels with lower mean and standard deviation.

Table 1: Gross employees' income (PY010G) distribution statistics before and after the appropriation of administrative data 2022

	Unweighted				Weighted			
	PY010G_before		PY010G_after		PY010G_before		PY010G_after	
	N	Value	N	Value	N	Value	N	Value
GINI index					4 422 843	0,38	4 524 698	0,42
S80/S20	13 467	6,17	13 623	8,06	4 422 843	6,82	4 524 698	8,81
S90/S10	13 467	13,41	13 623	16,79	4 422 843	14,92	4 524 698	18,64
Decile 1	1 346	7 547,17 €	1 362	5 662,92 €	442 266	7 584,27 €	448 761	5 505,62 €
Decile 2	1 347	10 980,39 €	1 362	8 764,05 €	440 584	11 070,56 €	455 147	8 703,18 €
Decile 3	1 347	12 337,82 €	1 362	10 246,15 €	443 701	12 387,88 €	453 291	10 122,00 €
Decile 4	1 346	13 257,86 €	1 363	11 676,43 €	442 533	13 462,03 €	450 596	11 703,88 €
Decile 5	1 347	14 313,97 €	1 362	13 357,72 €	442 246	14 738,89 €	445 056	13 513,51 €
Decile 6	1 347	16 402,12 €	1 362	15 840,03 €	441 688	17 232,38 €	459 705	16 133,61 €
Decile 7	1 346	19 755,77 €	1 363	20 077,22 €	442 768	21 212,12 €	453 770	20 981,83 €
Decile 8	1 347	25 470,33 €	1 362	26 109,05 €	442 135	26 486,49 €	449 696	26 264,09 €
Decile 9	1 347	34 658,89 €	1 362	35 086,75 €	442 370	37 517,63 €	454 905	36 084,46 €
Mean		19 076,51 €		18 421,94 €		20 401,32 €		19 155,68 €
Standard deviation		17 811,23 €		16 986,24 €		21 627,82 €		19 878,09 €

5. Concluding remarks

Following the European trend and recommendations, Statistics Portugal is in an initial stage of a process of integrating tax administrative data into SILC's income distribution. The first

experiences were focussed on employees' income, but we expect to extend the analysis to other income sources. At the same time, the tax authority information made available to Statistics Portugal is becoming more refined and increasingly meeting official statistics needs. There is evidence that the incorporation of administrative data increases overall inequality for employees' income.

References

Carranza, R., Morgan, M., and Nolan, B. (2021). Top income adjustments and inequality: an investigation of the EUSILC. *INET Oxford Working Paper*, no. 2021-16. Oxford: Institute for New Economic Thinking at the Oxford Martin School, University of Oxford.

European Commission. (2021). *Support for the Final Evaluation of the European Statistical Programme 2013-2020*. Final Report (D6). Tetra Tech International Development. https://ec.europa.eu/eurostat/documents/10186/13705908/2021-ESP-evaluation_contractor-report.pdf

European Commission (2022). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Better assessing the distributional impact of Member States' policies*, 494 final. https://ec.europa.eu/social/BlobServlet?docId=26123&langId=en

Eurostat. (2013). *The use of registers in the context of EU–SILC: challenges and opportunities*. Edited by Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier. https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

Hlasny, V. and Verme, P. (2018). Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data. *Econometrics*, *6*(2): 1–21.

Jenkins, S. P. (2022). Top-income adjustments and official statistics on income distribution: the case of the UK. *The Journal of Economic Inequality*, *20*: 151–168.

Webber, D., Tonkin, R., and Shine, M. (2020). Using tax data to better capture top incomes in official UK income inequality statistics. In: Chetty, R., Friedman, J. N., Gornick, J. C., Johnson, B., and Kennickell, A. (Eds.) *Measuring and Understanding the Distribution and Intra/Inter-Generational Mobility of Income and Wealth*. National Bureau of economic research, forthcoming.

Wiesbaden Memorandum (2011), as adopted by the ESSC on 28th September 2011. https://ec.europa.eu/eurostat/documents/13019146/13237859/Wiesbaden+Memorandum+2011.pdf