EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Integration of income administrative data into the Portuguese household income distribution

A first national experience using employees' income tax data

Eduarda Góis
Head of the Living Conditions Statistics Unit
Department of Demographic and Social Statistics
Statistics Portugal

Estoril, Portugal June 5, 2024

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

INSTITUTO NACIONAL DE ESTATÍSTICA | STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

PT-SILC

- **Target population:** all persons living in the national territory during the reference period
- **Sampling frame:** dwellings
- **NUTS 2 stratified and multi-stage sampling**
  - 1st stage: area selection (INSPIRE grid cells)
  - 2nd stage: selection of dwellings by area
- **Sample characteristics:** probabilistic (includes stratification and selection of several units in several steps); cross-sectional, longitudinal (4-year rotational scheme)
- In PT-SILC, all households and all individuals who consider the selected dwelling to be their **main residence** are interviewed
- Direct data collection through **CAPI** (face-to-face computer-based interviewing) and **CATI** (computer-based telephone interviewing)
- The **income** reference period of SILC collected in a specific year is the previous calendar year

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Problem

Sensitivity of income information → lower data robustness in comparison to qualitative data

Strategy

Use of administrative tax data

Outcome

Better survey data quality

The conference is partly financed by the European Union

The problem

- The survey collects
  - **a wide range of variables** associated to objective dimensions (e.g. monetary income components), and to subjective dimensions (e.g. material and social deprivation, activity status, health status, housing conditions and social exclusion)
  - **longitudinal data along 4 consecutive years**
  - both at the **individual and household levels**
- There is an increased difficulty in **keeping respondents motivated** → significant non-response rates
- The number of **proxy answers** may be significant (around 40% in 2023) → lower response quality
- **Income** information is particularly sensitive
- Annual **change rates** on survey income data (in particular, for employees' income) were too high in comparison to other sources (e.g. national accounts and social security data)

The conference is partly financed by the European Union

The problem

- Complexity of employees' income data collection

  → the survey includes **alternative questions** to facilitate the answer

  - based on the **income tax return** or **payslip** (preferable way)
  - **without documentation** (several questions to collect the same data)

  → the interviewed show some difficulty in **distinguishing gross and net amounts**

- Household surveys may **fail to capture** incomes at the **top** of the distribution
  - sparseness at the top of the distribution (sampling errors) and underreporting of the richest households (non-sampling errors)
- Income tax data **may not cover the bottom half** well by comparison with household surveys (Carranza et al., 2021; Jenkins, 2022)
- Concern that both inequality levels and trends over time may be mis-measured

The conference is partly financed by the European Union

The strategy

Current orientation in European statistics is to **encourage** the use of administrative data (European Commission, 2022)

- considering the past experience in official data collection, the European Statistical System mentioned in the Wiesbaden Memorandum (No. 5, b) the use of administrative data as a **key factor for the development of European social studies**

Using administrative data to **produce official statistics**

- 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)
  - – **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
  - \+ **higher** frequency → potential lower lag between the time of reference of the information and that of dissemination

The conference is partly financed by the European Union

The strategy

Aiming at improving the income components of the survey, Statistics Portugal decided to integrate data on employees' income from the survey with the tax authority administrative files in 2022 (2021 income)

- 2023 PT-SILC (2022 income) benefited from some refinements in the data transmitted by the tax authority

but there are challenges

- conceptual difference between SILC's private household definition and the tax household definition

- the lag between the time of the survey data collection (2nd quarter of the year) and the income tax reference period (year n-1) → possible differences in household composition
  e.g. 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

- 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

The conference is partly financed by the European Union

The strategy

Data on single persons within 25-59 age group show the overall similarity between the two series

- **joint tax returns** add an additional complexity in comparing these two data sources

The conference is partly financed by the European Union

The strategy

Combining the survey data and tax administrative data is a key step in the **integration process**

- the taxpayer number is not collected → **direct integration it is not possible**
- combining data requires that both sources **share a set of key variables**, 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 (and the integration may fail)

Personal income tax data (IRS Model 3, Annex A – **employees** and pensions income)
↳ selection of employees' income only for records with social contributions

↳ calculation of the share of each person's income in the survey for the distribution of labour income from tax data **(if joint tax return)**

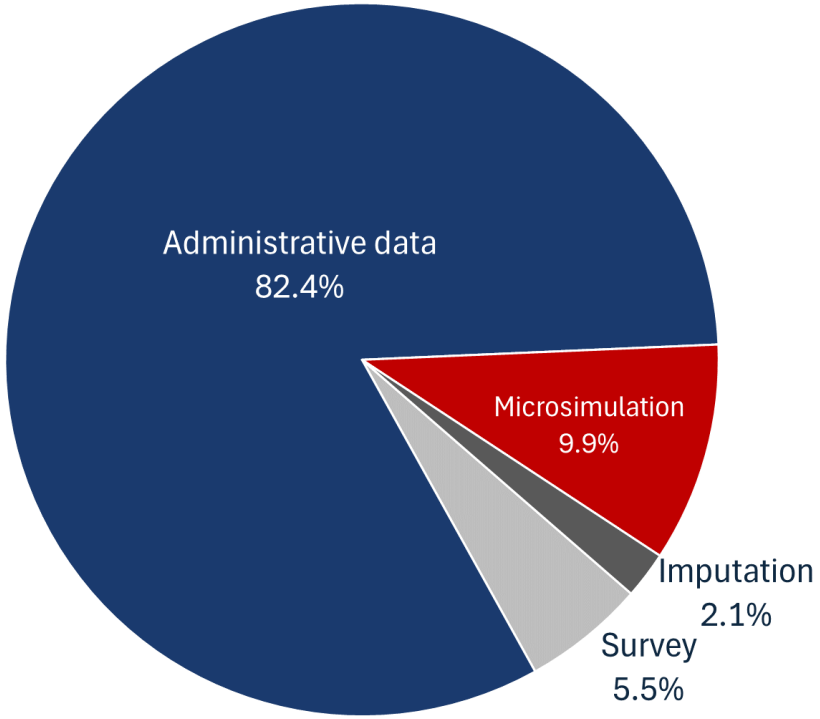
↳ **appropriation of employees' income** (tax authority) **only when the individual has indicated** employees' income in the survey

The conference is partly financed by the European Union

The outcome

The integration of tax administrative data (**PY010G**)

- changed 82.4% of the 2022 income survey data



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

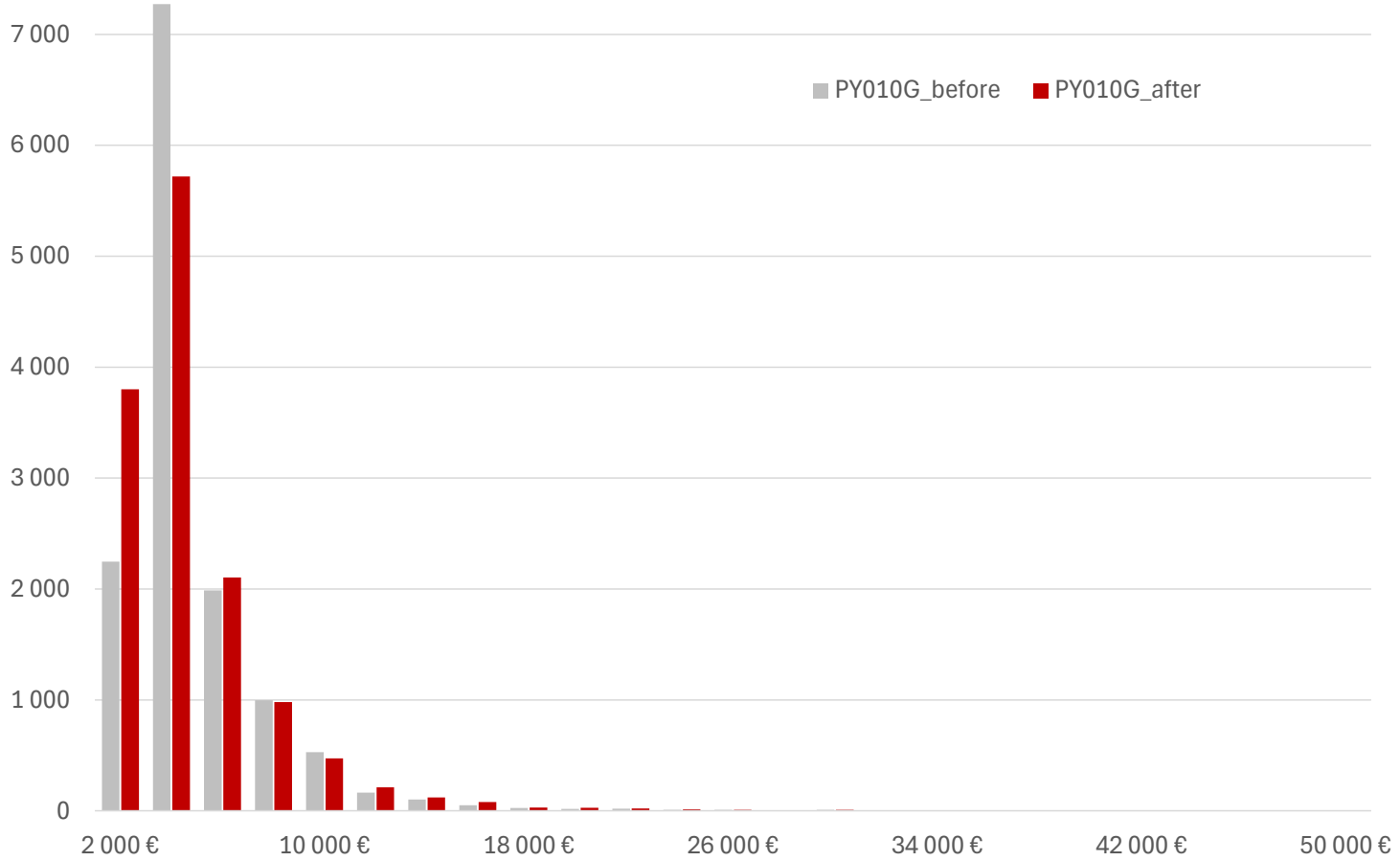
The conference is partly financed by the European Union

The outcome

The 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**

Gross employees' income class distribution, 2022



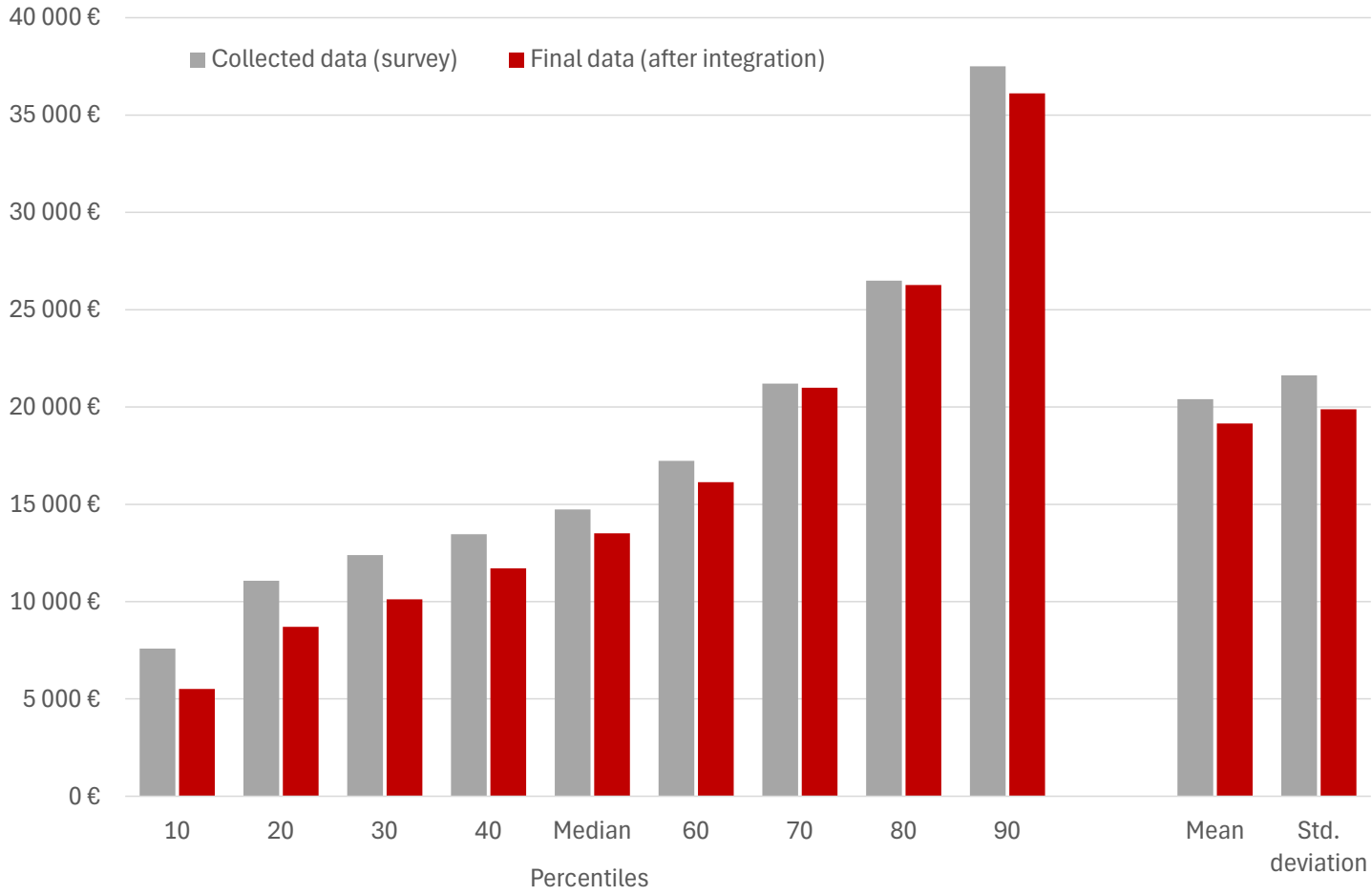
The conference is partly financed by the European Union

The outcome

The integration of tax administrative data (**PY010G**)

- impacted employees' income deciles, particularly in the first six deciles
  - 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
- lower mean and standard deviation

Gross employees' income decile distribution, 2022



The conference is partly financed by the European Union

Main findings

There is evidence that the incorporation of administrative data **increases overall inequality for employees' income**

	Weighted			
	PY010G_BEFORE		PY010G_AFTER	
	N	Value	N	Value
GINI	4,422,843	0.38	4,524,698	0.42
S80S20	4,422,843	6.82	4,524,698	8.81
S90S10	4,422,843	14.92	4,524,698	18.64
Mean	20,401.32€		19,155.68 €	
Standard deviation	21,627.82 €		19,878.09 €	

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Integration of income administrative data into the Portuguese household income distribution

Eduarda Góis

This is a joint work with
Carlos Farinha Rodrigues
David Leite
Daniel Gomes
Maria Manuel Pinho

Living Conditions Statistics Unit
Statistics Portugal

Estoril June 5, 2024

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union