Moving from experimental to official statistics: increasing the scope of statistics on earnings based on new administrative data sources

Daniela Ramos[1], Célio Oliveira[2], Ricardo Cotrim[3]

[1] Statistics Portugal, daniela.ramos@ine.pt

[2] Statistics Portugal, celio.oliveira@ine.pt

[3] Statistics Portugal, ricardo.cotrim@ine.pt

Abstract:

In the context of statistical production, there are well-known advantages in the use of administrative data to provide statistics that are relevant, timely and cost-effective. One of the main advantages relies on the fact that administrative data can broaden the scope of statistics. Specifically, more detailed and disaggregated data can be provided, and the creation of linked datasets can enable not only a more efficient data infrastructure, but also shed light on new and emergent phenomena. Administrative data can also replace and/or supplement sample surveys, which are often costly, increase coverage and data reliability and reduce response burden on respondents.

Against this background, Statistics Portugal has increased and broadened the scope of the statistics published on earnings to the extent that the dissemination of consistent and regular data on earnings is an important dimension of analysis of the labour market and its evolution.

The aim of this paper is to present the various statistical production processes implemented, which have made it possible to guarantee the quality of the information released quarterly on earnings, and to show the wide variety of statistics released so far.

Initially part of Statistics in Development (StatsLab), as of September 2021, Statistics Portugal releases quarterly official data on gross monthly earnings per employee (per job), calculated based on information at enterprise level received from the Monthly Statement of Earnings from Social Security, and the Contributory Relation of *Caixa Geral de Aposentações*. These statistics include various breakdowns by earnings components and enterprise characteristics (economic activity, company size, institutional sector). An important component of the process of moving from experimental to official statistics has been the data flow and data treatment development, including of non-responses, which combines specific pre-defined criterion and a supervised machine learning algorithm.

The consolidation of this information also allowed to discontinue the survey data collection on wages for the Labour Cost Index (LCI), with the main advantage being that now the wage component of the LCI takes into account the universe of enterprises.

More recently, Statistics Portugal has started to receive similar information from the Tax Authority, but at the individual (worker) level, which, within the framework of the StatsLab, has made it possible to extend the scope of analysis through data linkages with other databases available as part of Statistics Portugal's wider National Data Infrastructure, thus allowing additional information to be provided on several sociodemographic characteristics (until now, sex, age and education level).

Keywords: earnings, labour market, administrative data, experimental statistics

1. Introduction

Society in general and groups of experts in particular, like research centres, business and professional associations, trade unions, public policy makers, among others, understandably demand more statistical information to deepen the knowledge on their field of study and better understand emergent phenomena. At the same time, those who are data producers often feel overloaded with requests for new, more disaggregated and timely data. In this context, the use of administrative registers for statistical purposes has become a priority for statistical systems.

With the aim of replacing the traditional ten-yearly population and housing censuses with other based on administrative data, continuously updated and of annual frequency, Statistics Portugal (INE) has been developing a Resident Population Database integrated into a National Data Infrastructure (NDI), which aims to make more intensive and integrated use of administrative data and substantially extend the domains covered, freeing up resources and making room for greater innovation.[1] However, as the project is still under development, studies that rely on and cross-reference data within the NDI, but which provide useful information for economic and social analysis, are included in the dedicated area on experimental statistics of the Statistics Portugal website - StatsLab.

In this context, INE has signed protocols with entities that hold relevant administrative data for statistical purposes, with the aim of reducing the burden on respondents and creating new statistical products. This article presents the products resulting from the signing of three of these protocols, relating to data on earnings, two of which have contributed to the dissemination of official statistics on earnings and labour costs.

2. From experimental to official statistics: gross monthly earnings per employee based on administrative data at enterprise level

In 2018, as part of the Simplex+ Programme[1], INE began to receive monthly administrative information on earnings based on the Monthly Pay Statement submitted to Social Security (DMR-SS) by all companies[2] with employees registered in that social protection system.

This information has been analysed and integrated into numerous statistical operations with the aim of replacing that collected via business surveys, such as those that support the sectoral Indexes on Turnover, Employment, Wages and Salaries and Hours worked and the Labour Cost Index (LCI). In the latter, the use of administrative data made it possible to move from a sample survey of 3,8 thousand entities to a universe of circa 386 thousand[3], thus

[1] Government administrative modernisation programme.

[2] To simplify language, the term ‘company’ is used together with ‘enterprise’, although in addition to companies/enterprises, other organisations such as foundations, institutes and other bodies of a public, private or social sector nature are included in the data.

[3] The size of this universe results from the integration of the Social Security and *Caixa Geral de Aposentações* databases.

benefiting from the elimination of sampling errors. However, since the monthly administrative data on earnings still holds some limitations regarding the distribution and territorial classification of data at establishment level, the observed statistical unit changed from 'establishment' to 'enterprise', the LCI data is now published only at country level, and the regional analysis at NUTS 2 level is no longer available.[2]

Besides reducing the burden on the respondents, INE has used the information received to produce a new statistical product, starting the quarterly release of statistics on the average gross monthly earnings per employee (RBMMT) in May 2019 with data dating back to 2014, still within the scope of StatsLab.[3]

In that same year, Statistics Portugal signed an agreement with *Caixa Geral de Aposentações* (CGA[4]) to receive data on earnings for civil servants still enrolled in this social protection scheme, and incorporated this information into the quarterly release from November 2019 onwards.[4] The combination of these two sources made it possible to cover almost all employees in the economy, comprising around 445 thousand companies and approximately 4.6 million employees in 2023.

In November 2021, with the consolidation of the data flows and the methodological procedures, the results released quarterly moved from experimental to official statistics.[5]

2.1. Data sources: DMR-SS and CGA

Although both Social Security and CGA send data to INE on a monthly basis, the RBMMT monthly data is released quarterly in the seventh week after the end of the natural quarter.

The CGA data is final, while the DMR-SS data can be 'permanently' updated due to the existence of a non-negligible portion of undelivered declarations or corrections after delivery, especially in recent months. Since each month is received four times, it is assumed to be final by the fourth time. Therefore, the published information for the last three reference months is provisional in the first press release and is revised and deemed final in the following release.

To reduce the size of these revisions, DMR-SS values are imputed in two situations: 1) companies that systematically delay sending data; and 2) companies that regularly make substantial corrections to values reported in previous months. The detection of these companies is ensured by a combination of two methods: 1) ad hoc criteria; 2) supervised machine learning algorithm in the Support Vector Machine (SVM) version.[5]

Since the data is received at company level, it is possible to link the databases using the employers' tax identification number (NIF) or the legal person identification number (NIPC).[3] The final database covers almost the entire universe of employees, and those with more than one job are counted as many times as the number of jobs they have.

[4] Portuguese civil servants' retirement and survivor pensions fund.
[5] For more information on this procedure, see the Appendix to this article.

2.2. Computation and results

The RBMMT corresponds to the ratio between the sum of the volume of earnings paid by companies and the total number of workers in those companies. For this reason, its change reflects variations in the volume of payments (such as bonuses, holiday pay or overtime), but also in the number of workers and their composition, especially in terms of characteristics not observed in the database (part-time vs. full-time, level of education, occupation, years of experience, and hours worked, among others).

Data published only consider earnings subject to income tax and to Social Security or CGA deductions and, to minimise the impact of provisional months' revisions, the results are published as three-month moving averages (quarters ending in the reference months). Finally, to ensure statistical confidentiality, data based on less than six companies and/or less than eleven employees are not published.

The estimates are computed using Stata software and obtained using the following formula:

$$RBMMT_{Q,c} = \frac{\sum_i^k Earning\ paid\ by\ the\ company\ i_{Q,c}}{\sum_i^k Number\ of\ the\ company's\ employees\ i_Q}$$

where:

Q: moving quarter ended in month M, covering months *{M-2, M-1, M}*

c: earning component $\{Total, Regular, Base\}$

i: company number

k: total number of companies

$\sum_i^k Earning\ paid\ by\ the\ company\ i_{Q,c}$: volume of earnings paid in the period Q in the component c

Together with the release of the 4th quarter of each year, annual results are made available, corresponding to 12-month averages, even though salaries in Portugal are typically based in 14 payments.

The RBMMT is presented in various breakdowns, with Figure 1 providing an example of the information disseminated: (i) earnings component (total, regular, and basic); (ii) breakdowns of economic activity (NACE-Rev. 2 sections; groups by type of goods and services; characterisation of the technological intensity of 'Manufacturing' and the knowledge intensity in 'Services'); (iii) company size (measured by the number of employees); (iv) institutional sector (Public Administration vs. Private).

All these breakdowns are available in nominal and real values, deflated to January 2014.

Figure 1: Total gross monthly earnings per employee in 2023 and rate of change since 2014, in real terms, by economic activity (NACE-Rev. 2), employment size class and the institutional sector (public and private). [6]

3. One step further: experimental statistics on gross monthly earnings per employee based on administrative data at worker level

In Portugal, employers must report, for each employee, the amount of earnings that is subject to taxation/contribution, both to the Tax Authority (AT) and to the Social Security (SS). However, the scope of the information is different, as summarised in Table 1[7].

Table 1. Monthly Pay Statement from Social Security and Tax Authority

	Social Security[(a)]	**Tax Authority**	
Desagregation level	Enterprise	Employee	
Observation unit	Job [(b)]	Employee	Job [(b)]
Annual average in 2021	*4.2 millions*	*4.6 millions*	*5.5 millions*
Covered social protection schemes	Social Security [(a)]	All schemes of social protection [(c)]	
Typo of income/earning	Only those subject to income tax/contribution	All income or earnings	

Notes:

(a) This includes data from CGA.

(b) Corresponds to the unique combination of employee and company.

(c) This includes the various pension funds, the Social Security and the CGA, for example.

With the signing of a protocol with the AT, INE started receiving information from the DMR-AT at employee level. This fact broadened the scope of possibilities for statistical production, since it allowed combination with other data sources available within the National Data Infrastructure, thus enabling the sociodemographic characterisation of employees and resulted, in 2023, in the availability of two StatsLab studies based on these data.

3.1. Data source: DMR-AT

As data are not collected for statistical purposes, the analysis at employee level required the definition of some assumptions, of which we highlight the following[6]:

- The earnings include all the amounts earned that year in all jobs.
- The economic activity and occupation represent the main activity/occupation of the worker in that year.
- For the analysis by institutional sector, employees who worked in both sectors during 2021 were excluded, which covered 80 thousand workers (1.7% of the total of 4.6 million).

Record matching between the data received from the AT and that available in the NDI varied according to the dimensions and years being analysed, as shown in Table 2.

Table 2. Coverage rates of the dimensions characterising individuals and companies

Portugal	2019	2020	2021
Total	**100.0**	**100.0**	**100.0**
Sex	95.8	96.5	94.7
Age group	95.8	96.5	94.7
Educational level	86.1	83.8	77.8
Region of residency (NUTS 2)	96.9	96.4	89.9
Economic activity (NACE)	99.7	99.9	99.9
Occupation (ISCO)	75.0	73.6	63.1

3.2. Computation and results

Analysing the integrated database by employee resulted in two StatsLab Press Releases: a more complete one, with numerous dimensions for analysis and covering three years (2019, 2020 and 2021); and another focused on the distribution of earnings by institutional sector in 2021.

The first included a dynamic Excel file in which the user can choose the characteristics of the workers (sex, age group, level of education, NUTS 2 region of residence, and occupation

[6] For more information on the assumptions made, see the Appendix to this article.

group) and companies (economic activity) for which they want to know the earnings figures, their distributions (mean, median, deciles and percentiles) and inequality indicators (Gini coefficient, S80/S20 and P50/P10). Figure 2 is an example of the figures automatically generated according to the selected dimensions.[7]

The second publication focussed on comparing the distribution moments (percentiles and mean) between the two institutional sectors disaggregated by sex, age group and level of education. Figure 3 is an example of the published analysis.[8]

Figure 2. Distribution of the gross monthly earnings per employee in NACE I and NACE K in 2021.

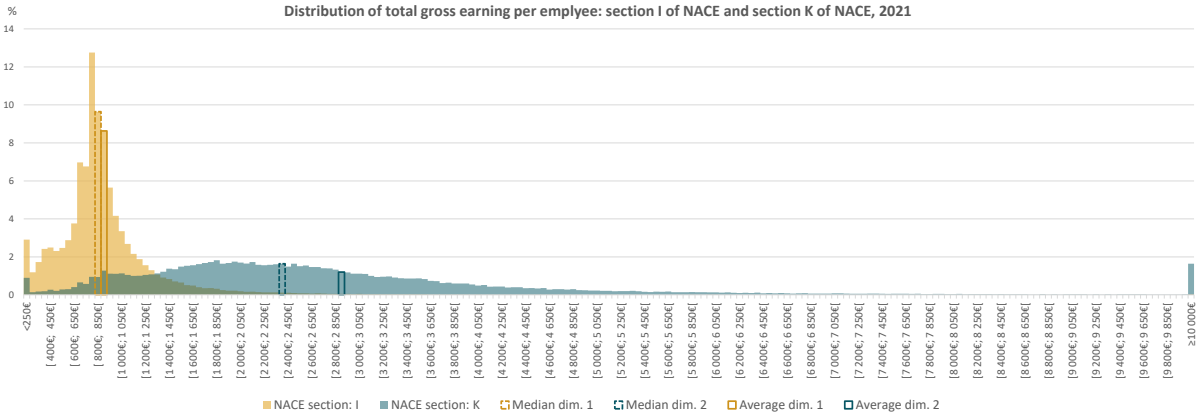
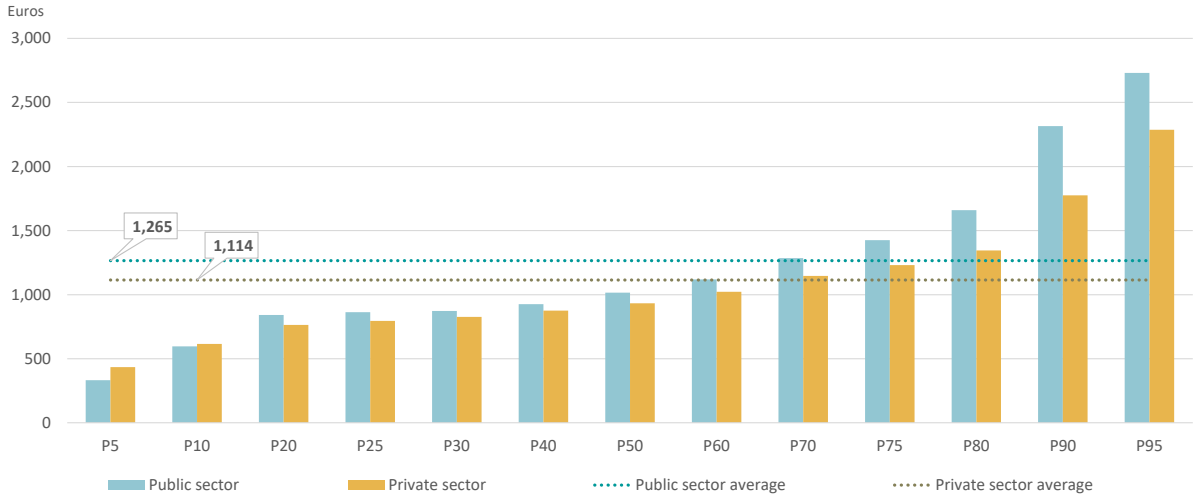


Figure 3. Distribution of gross monthly earnings per employee who has completed, at most, the lower secondary education by institutional sector.



4. Conclusions and future developments

The use of administrative data makes it possible to increase the frequency of information, reduce the statistical burden on respondents, as well as costs and produce new statistical products.

Since 2018, Statistics Portugal has been cooperating with Social Security, CGA and the Tax Authority to produce new and relevant information of public interest and for the definition and monitoring of public policies in the area of labour market and living conditions, while also generating information of academic interest for teaching and research purposes.

The management of administrative data received at enterprise level has already reached a level of maturity that allows the production of official statistics and the replacement of infra-annual surveys that used to support, along with other sources, numerous official indicators.

The integration of data at employee level with the National Data Infrastructure still requires further developments in terms of data consistency, data coverage and data quality to move from experimental to official statistics. These are, however, the focus of future projects, with the following developments being expected:

1. Data from the DMR-SS and CGA will be received at the employee level, maintaining and enriching current official statistics by allowing a sociodemographic characterisation of employees.
2. Updating the procedure to deal with non-responses to further minimise revisions to DMR-SS data.
3. Regularly analysis of earnings distribution (median, quartiles or deciles) and inequality indicators.
4. Study of movements into and out of the labour market and associated wage changes.

5. References

[1] “O desenvolvimento da Infraestrutura Nacional de Dados no INE” presented at the 27.ª Plenary Session of Statistical Council on 31 May of 2019 and available at the Portuguese Statistical Council website (only in Portuguese).
[2] Methodological document “Labour Cost Index (Enterprise)”, version 3.0, April of 2023, available at the Metadata System area at the Statistics Portugal website (only in Portuguese).
[3] StatsLab Press Release “Gross Monthly Earnings per employee (Social Security – Data analysed by Statistics Portugal) – March of 2019”, published on 9 May of 2019 and available at the StatsLab area of the Statistics Portugal website.
[4] StatsLab Press Release “Gross Monthly Earnings per employee (Social Security and Caixa Geral de Aposentações – Data analysed by Statistics Portugal) – September of 2019”, published on 7 November of 2019 and available at the StatsLab area of the Statistics Portugal website.
[5] Press Release “Gross Monthly Earnings per employee – September of 2021”, published on 11 November of 2021 and available at the Statistics Portugal website.
[6] Press Release “Gross Monthly Earnings per employee – December of 2023”, published on 15 February of 2024 and available at the Statistics Portugal website.
[7] StatsLab Press Release “Gross Monthly Earnings per employee (Tax Authority data – Statistics calculated and analysed by Statistics Portugal) – 2019-2021”, published on 12 April of 2023 and available at the StatsLab area of the Statistics Portugal website.
[8] StatsLab Press Release “Gross Monthly Earnings per employee (Tax Authority data – Statistics calculated and analysed by Statistics Portugal) – 2021”, published on 23 May of 2023, available at the StatsLab area of the Statistics Portugal website.

APPENDIX

1. Dealing with non-responses in the DMR-SS

To reduce the size of the revisions to the last three months of DMR-SS data, values are imputed in two situations:

A. companies that systematically delay sending information; and
B. companies that regularly make substantial corrections to figures reported in previous months.

In A, the process of detecting missing companies focuses only on those with 10 or more employees, considering as missing a company for which there was a response in month M-1, but not in month M (M being the last reference month).

In B, a company is considered to have made a substantial correction to the values already reported when the revisions are equal to or greater than 10 thousand Euros.

The detection of these companies is ensured by a combination of two methods:

1) ad hoc criterion;
2) supervised machine learning algorithm in the Support Vector Machine (SVM) version.

More specifically, a given company fulfils the ad hoc criterion if it falls into one of the following scenarios (or both):

i. it has made at least 9 corrections in the last 12 months;
ii. it has made at least 3 corrections in the last 4 months.

The SVM algorithm makes it possible to identify companies that systematically correct information through an optimisation process. In this procedure, a set of training data is used (records of companies that correct information and companies that don't) to which the SVM algorithm is applied to obtain a classification model that maximises the distinction between the two groups of companies, i.e. a model with the maximum success rate (accuracy) in identifying companies that revise the information provided. Although most companies are identified simultaneously by both methods (ad hoc and SVM), each of them identifies fringes of companies that the other does not. Using both methods guarantees a greater number of companies identified.

Once the companies for which values need to be imputed have been identified, the earnings volumes are imputed by company and by type of remuneration distinguishing between earnings components:

- For regular components (such as ‘Monthly prizes, bonuses or allowances’, ‘Base salary’, ‘Meal allowance’ and ‘Night work compensation’), the amount declared in the previous month is imputed.
- For non-regular components (such as ‘Non-monthly prizes, bonuses or allowances’, ‘Holiday allowance’ and ‘Christmas allowance’), the same value as the previous year

is imputed, multiplied by the year-on-year rate of change in base salary for the previous month.

- For the other categories of remuneration, the median value of the last 12 months is used, provided there are at least 6 observations; otherwise, the value of the last month is imputed.

2. Integration of the DMR-SS and CGA databases

Since the data is received at company level, it is possible to link the databases using the employers' tax identification number (NIF) or legal person identification number (NIPC), namely:

1. Each database is harmonised so that, for each year-month-NIPC/NIF trio, the number of employees and the volume of earnings (total, regular and basic) are obtained.
2. The two databases are then joined by year and month, using the NIPC/NIF as the unique identifier.
3. Their NACE-Rev. 2 section is then identified by linking the year-month-NIPC/NIF trio to the Statistical Units File (FUE) register, a daily updated list of all companies and establishments in the country. In 2023, this identification was not possible for 0.4% of companies, which corresponded to 0.1% of workers.
4. Finally, the totalisers are computed, this is for each year-month-NIPC/NIF trio, the total number of workers and the volume of earnings (total, regular and basic) are calculated.

3. Integration of the DMR-AT database and NDI

As it is not a database collected for statistical purposes, to analyse the DMR-AT data by worker, certain assumptions had to be made:

- In the total earnings, all the amounts earned that year in all jobs were included.
- When analysing by economic activity (NACE) and by occupation (ISCO), the amounts paid by the company that provided the highest volume of annual remuneration to the worker were considered.
- If the employee worked in companies with different NACE sections or worked in different occupations throughout the year, the main activity and occupation were considered. The following criteria were adopted sequentially: (i) the highest paying job; (ii) the company in which they worked the most months; (iii) the most frequent combination of NACE and ISCO in the year; (iv) the occupation worked in the largest company, measured by the number of staff employed.

- For the total, no restriction was placed on the age of employees, but individual analysis of the age group of those aged 75 and over was not allowed, due to their heterogeneity and lower participation in the labour market.
- Due to the very small number of records of workers in Sections T (Activities of households as employers; undifferentiated goods - and services - producing activities of households for own use) and U (Activities of extraterritorial organisations and bodies) of NACE-Rev.2, which would jeopardise the confidentiality of the data, individual analysis of these sections was not allowed, even though they were included in the total.
- For the analysis by institutional sector (Public Administration vs. Private), workers who worked in both sectors during the year under study were excluded from the base, which covered 80 thousand workers (1.7% of the total of 4.6 million).