Distr.: General
05 October 2017

English

United Nations Economic Commission for Europe

Conference of European Statisticians

Work Session on Migration Statistics
Geneva, Switzerland
30-31 October 2017
Item 3 of the provisional agenda
Data integration and administrative data

Migration estimates from Portuguese tax records combined with other administrative registers and estimation methods

Note by Statistics Portugal*

Abstract

Statistics Portugal is currently studying the use of administrative data in the production of official population statistics, based on a variety of sources, which were first acquired in order to develop a prototype for getting population counts based on the sign-of-life methodology but are equally relevant for other purposes, such as the improvement of migration statistics. The tax register stands out as one of the most significant and comprehensive data source to be considered at this level, since it provides information on national and foreign individuals of all ages, residing in Portugal or abroad. The transition method will be adopted for the measurement of migration flows. After the necessary adjustments (including weight calibration), the results thus obtained will then be used to calculate the distribution ratio amongst sub-national areas (the actual number of immigrants/emigrants at the national level will be estimated according to the current methods). The study in question will also cover internal migrations through the comparison of the postal codes of tax payers who are present in the country for two consecutive years. In spite of the limitations of the available data, the results obtained so far appear to point to the possible use of tax registers within the production of migration estimates. In a first phase, however, this process will be limited to the improvement of existing migration statistics, namely in what concerns the production of estimates at a more detailed geographic level. A more ambitious approach will require a close cooperation with

*Prepared by Ms. Paula Cruz

the Portuguese Tax and Customs Authority in order to optimize the use of tax registers for statistical purposes.

I. Introduction

1. Statistics Portugal is currently studying and exploring the use of administrative data in the production of official population statistics, based on a variety of sources, namely: national population civil register, foreign population register, education attainment register, tax register, social security register, employment and unemployment registers, etc.

2. These data sources were first acquired in order to develop a prototype for getting population counts based on the *sign-of-life* methodology but they are equally relevant for other purposes, such as the improvement of migration statistics, due to the availability of residence related information for two consecutive years.

3. We recall that, currently, the annual estimates of immigrants and emigrants released by Statistics Portugal - based on the Labour Force Survey (LFS) combined with foreign population registers, regarding immigration flows and on the Emigration Survey (an LFS module) in what concerns emigration flows - are only available at the national level, due to the lack of reliability of existing estimates at less aggregate levels.

4. According to the analysis already carried out, the tax payers' register (including fillers, non fillers and dependents) stands out as one of the most significant and comprehensive data source to be considered in the production of sub-national estimates, since it provides information at postcode level on national and foreign individuals of all ages (it is mandatory to register newborns since 2011), residing in Portugal or abroad.

5. As defined in article 16 of the Personal Income Tax Code, an individual is qualified as a resident of Portugal if he spends more than 183 days on Portuguese Territory, counted in any 12-month period starting or ending in the calendar year concerned. When staying for a shorter period of time, the individual still qualifies as a resident of Portugal if he has, at any time of that same period, a dwelling with the intention of maintaining it as his usual residence.

6. Unless proven otherwise, an individual shall also be considered a resident of Portugal if he moves his residence to another territory only to benefit from a clearly more favourable tax regime included in a list approved by the Government.

7. This concept of tax residency, which has entered into force on January 1, 2015, is more strongly related to the period of the individual's effective residency in the country than the previous one. In fact, until December 31, 2014, if an individual spent, for instance, 9 months on Portuguese Territory he would be considered a resident of Portugal for the whole year; under the new rules, he may be considered a resident of Portugal in the first 9 month period and a non-resident during the remaining time of the year [1, 2].

8. Since the available data covers the years 2014 to 2016, it will be interesting to analyse the impact of these conceptual changes on the sub-national relative distribution of potential migrants.

9. This document aims to present an outline of the methodology that will be adopted within the study in question, alongside a brief description of the data sources being used, including quality and coverage issues already detected. Some preliminary results concerning the regional breakdown of migration flows, as well as the main limitations and conclusions of the study carried out will also be presented.

II. Data sources description

10. The main data source consists of tax payers' registers (including fillers, non fillers and dependents) for the years 2014, 2015 and 2016 (the reference date being 31 December of each year), with the following totals:

Table 1 - Tax registers

	2014	2015	2016
Country of Residence = PT	14 029 321	14 022 860	13 958 720
Country of Residence = Other	1 005 260	1 703 751	1 610 837

11. It should be noted that the late release of the data concerning the year 2016 resulted in a more limited analysis of the information in question.

12. These files contain basic demographic and geographic information concerning the tax payer, with a coverage of 100%, namely:

- Sex;
- Date of birth;
- Country of citizenship (PT/Other);
- Place of birth (district, municipality and parish);
- Address (locality, postal code);
- Country of residence (PT=True or False).

13. The tax registers show an obvious overcoverage when compared to annual estimates of resident population (10 341 908 on average, in the period 2014-2016). However and as illustrated in the graphs below, tax payers who are signalled as residents of Portugal show a similar relative distribution by sex; as far as age is concerned and with the exception of the first one (0-4 year), all age groups evidence an overcoverage by the tax register, especially in the 85 and more years group. This results from the non-exclusion of deceased tax payers, who are never deregistered for fiscal purposes, but rather maintained as inactive tax-payers through a status field which has not been provided and would be needed to correctly identify these cases. However, it should be noted that this omission doesn't interfere with the study of migration flows since these imply changes in the individuals' residence and, as such, are less prone to being related to deceased or inactive tax payers.

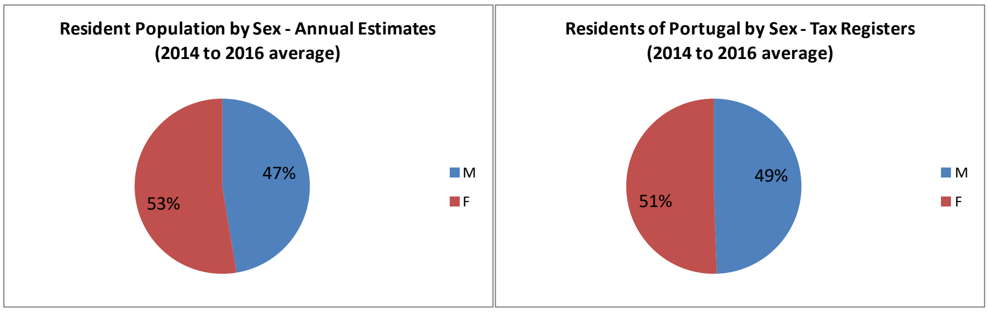
14. Another meaningful finding concerns the great similarity between tax registers and resident population estimates relative distributions by NUTS 2 region, in the period under consideration.

15. It is expected that, in future releases of these files, additional data will be provided, concerning, namely:

- Events that affect the register's objects such as the date of last update;
- Quarterly releases of information which will enable to determine more effectively the type of migration movement (temporary or permanent).

16. Lastly and as previously mentioned, other data sources will also be considered in order to deal with overcoverage issues, based on the *sign-of-life* methodology, which basically aims at finding individuals who are active in at least two administrative registers. An already available prototype for getting population counts based on several administrative registers concerning the year 2015 will be used for that effect [3].

Graph 1 - Resident Population by Sex



Graph 2 – Resident Population by Age Group

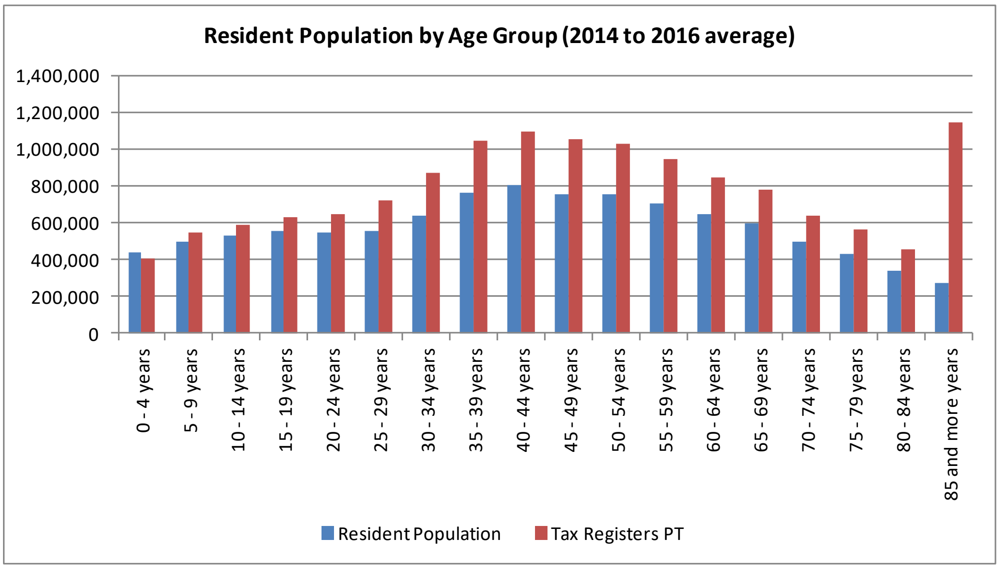


Table 2 - NUTS 2 relative distribution: resident population estimates (ResPop) vs. tax registers (TaxReg)

	2014		2015		2016	
D_NUTS 2	ResPop	TaxReg	ResPop	TaxReg	ResPop	TaxReg
Norte	34.91%	33.91%	34.85%	33.79%	34.77%	33.71%
Centro	21.82%	22.10%	21.82%	21.94%	21.76%	21.79%
Lisboa	27.08%	26.46%	27.20%	26.72%	27.37%	26.94%
Alentejo	7.07%	6.93%	7.00%	6.92%	6.97%	6.88%
Algarve	4.26%	5.01%	4.27%	5.03%	4.28%	5.07%
RAA*	2.37%	2.70%	2.38%	2.72%	2.38%	2.73%
RAM*	2.49%	2.89%	2.48%	2.88%	2.47%	2.88%

*RAA – Região Autónoma dos Açores; RAM – Região Autónoma da Madeira.

III. An outline of the methodology

17. The proposed methodology will follow closely the one adopted by Statistics Canada, as described in several documents, dating from 1998 to 2016, after successive changes [4, 5]. The basic idea will be to use tax registers as a source of detailed geographic information in what concerns changes in the tax payer’s residence from Portugal to abroad and vice-versa (international migrations) or within Portugal (internal migrations).

A. International Migrations

18. First of all, it should be noted that, due to overcoverage issues and lack of relevant meta information concerning the tax register, the actual number of immigrants/emigrants at the national level will continue to be estimated according to the current methods.

19. The *transition* method – according to which the population of a country is compared at two points in time (t and t+1), instead over a given period (t, t+1] as in the event approach [6] - will be adopted for the identification of migrant tax payers, that is, a person will be considered:

- A potential emigrant when he or she is present in the country at t but absent at t+1 (excluding deaths) and
- A potential immigrant if absent at t but counted in the country at t+1 (excluding newborn infants).

20. In the first case, only tax payers who are registered in two consecutive years with a change in the variable indicating whether the tax payer is a resident in Portugal (from “True” to “False”) will be considered (this update by itself will, in principle, exclude possible deaths). At this phase of the project, it won’t be possible to consider non-reported residence changes.

21. Immigrants, on the other hand, will be identified in one of two ways:

- Tax payers who are registered in both of the years with a change in the variable indicating whether the tax payer is a resident in Portugal, from “False” to “True”;
- Tax payers who are only registered in the second reference year born before that same year.

22. Some preliminary results have already been obtained and confirm the expected overcoverage, in terms of absolute total values, when compared to the existing estimates of international migrants (2015).

23. Through the use of additional administrative data sources and the encrypted VAT number of each individual, based on the previously mentioned *sign-of-life methodolog*y, it was possible, however, to reduce those values to the following ones:

Table 3 - Potential international migrants based on tax registers

	Tax register (2015)		International migration estimates (Statistics Portugal, 2015)
Potential emigrants	226 558	Emigrants (total)	101 203
Potential immigrants	31 519	Immigrants (permanent)	29 896

24. The obtained flows still point to an overcoverage, especially in what concerns the total number of emigrants, which may result from the following:

- Inclusion of potential emigrations with a duration below the 3 months threshold;
- Residence changes related to the new concept and not with an effective migration;
- Incoherence between variables pertaining to the same domain (residence, for instance).

25. Further corrections for overcoverage may still be made thus, based on additional administrative sources and coherence rules. However, due to the lack of relevant information, at this phase of the study only this last type of validation will be made, namely in what concerns residence-related data.

26. After these additional corrections and based on the existing estimates[1] by sex, age, citizenship (PT/Other) and or type of migration, calibration methods will be used to adjust the weight of each register (before adjustment, all weights are equal to 1; after adjustment, the weights for the different categories for which there is overcoverage will be less than 1), following closely the methodology proposed in [7].

27. The results thus obtained will then be used to calculate the distribution ratio amongst sub-national areas (namely, NUTS 2), which will be determined based on the administrative divisions corresponding to valid postal codes. These last ones will be determined based on the latest file provided by the postal services of Portugal. Postal codes corresponding to post-office boxes will be discarded; postal codes associated to more than one administration division will be imputed to the most significant one, based on relevant criteria.

28. In the case of emigration flows, destination countries will also be analyzed based on the registered residence concerning the year 2015. It should be noted, however, that the field in question consists of a text field with no particular format. Existing registers contain references to countries, regions, cities, street names and/or

[1] Emigration Survey microdata and official permanent immigration estimates.

postal codes, manually inputted in various languages. Hence, the use of such data implies a lengthy process of data cleaning in order to match the resulting string with standardized registers of countries, cities and postal codes.

29. The main software that will be used for the extraction, pre-processing and analysis of the available data consists of R Studio and the following packages and libraries: ROracle, maps, ggmaps, plyr, sqldf, data.table, stringr, dplyr, tm and DescTools. Google's API for geocoding will also be used.

B. Internal Migrations

30. The identification of internal migrant tax payers will be based on the comparison of the postal codes of tax payers who are present in the country (PT="V") for two consecutive years, namely 2014 and 2015 and who simultaneously show "signs of life" in this last year.

31. The resulting data will be analysed at a more aggregate level (namely, by NUTS 2) based on the method of social network analysis (SNA) [8] and Gephi software [9, 10], available at http://gephi.org.

IV. Some preliminary results

A. International Migrations

1. Emigration flows

32. The non-calibrated register of potential emigrants accounts for 194 533 individuals, that is, almost double of official estimates (101 203), which may result from the reasons presented earlier, especially in what concerns the concept of resident of Portugal underlying the tax registers which was in force in the year 2014. In fact, an analysis of potential emigrations between 2015 and 2016 based on the new concept of tax residency, shows a significantly lower difference in relation to official estimates: 127 561 vs. 97 151, which is also in line with the observed decline of these type of migration flows.

33. The variables that were used to calculate the calibrated weights of each potential emigrant contain no missing values and consist of the following ones: age, sex and country of citizenship. Two calibration exercises were carried out: (1) one based on official estimates by five-year age groups, sex or nationality (groups of countries); (2) another, by sex broken down by functional age groups (0-14,15-24,25-64,65+), based on available microdata (emigration survey).

34. The resulting distribution ratios by NUTS 2 region of previous residence are presented in the table below, alongside the ones based on the non-calibrated observations:

36.

Table 4 - Distribution ratios by NUTS 2 region of previous residence (potential emigrants, 2015)

NUTS 2 region	Calibrated observations		Non-calibrated observations
	(1)	(2)	
Norte	40.30%	39.19%	38.92%
Centro	25.24%	25.40%	25.96%
Área Metropolitana de Lisboa	23.43%	23.65%	22.98%
Alentejo	3.53%	3.54%	3.57%
Algarve	3.89%	4.89%	5.36%
Região Autónoma dos Açores	0.34%	0.31%	0.29%
Região Autónoma da Madeira	3.27%	3.02%	2.92%

37. It is possible to observe that the three distributions are very similar, as shown by a high correlation coefficient (around 99%, in all cases). They are also highly correlated (around 98%) with the distribution ratio of resident population in 2015 and with one resulting from the emigration survey.

38. These results hold true for residence status changes between 2015 and 2016, pointing to an apparently low impact of the new concept of tax residency on the regional distribution of the underlying individuals:

Table 5 - Distribution ratios by NUTS 2 region of previous residence (potential emigrants, 2016)

NUTS 2 region	Non-calibrated observations
Norte	40.73%
Centro	26.97%
Área Metropolitana de Lisboa	20.76%
Alentejo	3.39%
Algarve	4.82%
Região Autónoma dos Açores	0.76%
Região Autónoma da Madeira	2.57%

39. In what concerns the main countries of destination, the results obtained so far appear to be in line with figures taken from Portuguese consulates' registers [11], with the exception of Switzerland and Luxembourg, as shown in table 6. However, it should be noted that almost 18% of the registers under consideration consist of unknown cases due to insufficient information.

Table 6– Main countries of destination (Europe) ranking

Country	Emigration Report (2015)	Tax Registers (2015)
UK	1	2
France*	2	1
Switzerland	3	7

Germany	4	4
Spain	5	5
Belgium*	6	6
Luxembourg	7	3

*Emigration Report data relates to the year 2013.

2. Immigration flows

40. As presented earlier, the figures for potential immigrants based on tax registers are very close to the official estimates concerning the same period (2015). It should be noted, however, that this result was greatly achieved at the expense of the *sign of life methodology*, which enabled to reduce the initial number of observations. As in the case of emigrations, this may result from the tax residency concept that was in force in the year 2014.

41. Only one calibration exercise was carried out, based on the sex variable broken down by five-year age groups and nationality (group of countries), resulting in the following distribution ratios by NUTS 2 region of residence:

Table 7 - Distribution ratios by NUTS 2 destination region (potential immigrants, 2015)

NUTS 2 region	Calibrated observations	Non-calibrated observations
Norte	20.84%	20.09%
Centro	14.68%	15.43%
Área Metropolitana de Lisboa	46.79%	48.90%
Alentejo	4.48%	4.42%
Algarve	10.67%	8.48%
Região Autónoma dos Açores	1.13%	1.21%
Região Autónoma da Madeira	1.41%	1.47%

42. As in the case of emigrations, it is possible to observe that the two distributions are very similar, as shown by a high correlation coefficient (around 99%, in all cases). They are also highly correlated (around 90%) with the distribution ratio resulting from the LFS and less correlated to the corresponding distribution of resident population (75%).

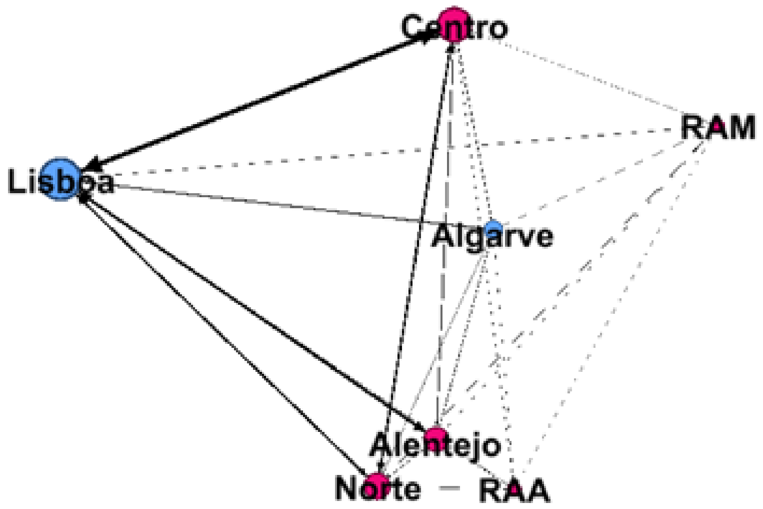
B. Internal Migrations

43. Preliminary results point to 104 216 internal migrants by NUTS 2, in the period under consideration (2014-2015).

44. In the graph below, the regions with a negative net-migration are signalled in red and the size of each node reflects the volume of migrants, being possible to observe that only the NUTS 2 regions of Lisboa and Algarve evidence a greater

number of inflows than outflows over the period 2014-2015. This pattern is somewhat similar to the one underlying the distribution of potential immigrants by NUTS 2 region (table 7).

Graph 3 – Potential internal migrants by NUTS 2 – SNA



V. Conclusions

45. The currently available tax data present some limitations, the most important of which relates to the reference period of the registers provided so far (31 December of each year). Quarterly updates will enable to determine more effectively the volume of migrations since it will be possible to measure the actual period of residence of individuals who change their residence data throughout the year, once or more. Ultimately, potential short-term circular migrations may also be captured by using the proposed methodology.

46. In spite of the current limitations, the results obtained so far appear to point to the possible use of tax registers within the production of migration estimates. In a first phase, however, this process will be limited to the improvement of existing migration statistics, namely in what concerns the production of estimates at a more detailed geographic level, according to the methodology previously presented (in which the actual number of immigrants/emigrants at the national level will continue to be estimated according to the current methods).

47. A more ambitious approach will require a close cooperation with the Portuguese Tax and Customs Authority in order to optimize the use of tax registers for statistical purposes.

VI. References

[1] Nogueira D. (2015): Tributação das pessoas singulares: residentes, não residentes e residentes não habituais. Tese de Mestrado em Direito (Direito Fiscal). Universidade Católica Portuguesa.

[2] Flores C. (2016): O Elemento da Residência Fiscal para as Pessoas Singulares. Tese de Mestrado em Direito. Universidade do Porto.

[3] Gabinete para os Censos 2021 (2015): Estudo de viabilidade para os Censos 2021 - Linhas gerais do novo modelo para os Censos 2021, a testar em 2016 (Relatório QUAR). INE.

[4] Demography Division (2016): Population and Family Estimation Methods at Statistics Canada. Statistics Canada.

[5] Small Area and Administrative Data Division (1998): Description of the methodology used to create migration data from taxation records. Statistics Canada.

[6] Wisniowski A. (2017): Combining Labour Force Survey data to estimate migration flows: The case of migration from Poland to the UK. Journal of the Royal Statistical Society. Series A: Statistics in Society.

[7] Wallgren A. and Wallgren B. (2014): Register-Based Statistics - Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd.

[8] Maier G, Vyborny M, 2008, “Internal migration between US States: A social network analysis”, in Migration and Human Capital Eds Poot J, Waldorf B, Wissen L W (Edward Elgar, Cheltenham, Glos) pp 75–93.

[9] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[10] Jacomy M, Venturini T, Heymann S, Bastian M (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9(6): e98679. https://doi.org/10.1371/journal.pone.0098679.

[11] Portal das Comunidades Portuguesas (2015): Relatório da Emigração 2015. Gabinete do Secretário de Estado das Comunidades Portuguesas, pp 49.