**­Record linkage methods for administrative data: The Portuguese Census Transformation Program**

Sandra Lagarto, Statistics Portugal, 2021 Census Unit, [sandra.lagarto@ine.pt](mailto:sandra.lagarto@ine.pt)

Anabela Delgado, Statistics Portugal, 2021 Census Unit, [anabela.delgado@ine.pt](mailto:anabela.delgado@ine.pt)

R. Silva, Instituto Superior Técnico/INESC-ID, [ruimenezessilva@tecnico.ulisboa.pt](mailto:ruimenezessilva@tecnico.ulisboa.pt)

L. Sampaio, Instituto Superior Técnico/INESC-ID, [lufialuiso@fc.uan.ao](mailto:lufialuiso@fc.uan.ao)

M.J. Silva, Instituto Superior Técnico/INESC-ID, [mario.gaspar.silva@tecnico.ulisboa.pt](mailto:mario.gaspar.silva@tecnico.ulisboa.pt)

P. Calado, Instituto Superior Técnico/INESC-ID, [pavel.calado@tecnico.ulisboa.pt](mailto:pavel.calado@tecnico.ulisboa.pt)

**Abstract**

*From 2014 to 2016 Statistics Portugal (SP) has been studying the usability of available administrative data for census purposes. Unlike other countries that have already made the transition to a register-based or a combined census model, in Portugal there is neither a central population register nor a unique personal identification number. SP created a methodological framework to build a Statistical Population Dataset (SPD), integrating linked registers in a “potential” resident database and then applying a ‘signs of life methodology’ to estimate true residents. For 2015, the population estimated from the SPD has a deviation of 0.9% to the official resident population estimates for the same year. Results are promising at a national level, but there are multiple hurdles to the creation of this dataset with high accuracy at the detailed geographical level: records have inconsistencies and errors due to manually inserted data, and the National Data Protection Authority imposed anonymization criteria on the datasets, restricting access to the full name and address of the persons in registers. Exact comparison methods performed by SP left out many potential matches (roughly more than 5% for most sources). With the goal of identifying the highest number of linked pairs of records, we developed an alternative linkage model, using logistic regression, which added thousands of new pairs of linked registers to those found by SP. The precision of the method is about 99% on a large set of linked records used as gold standard.*

**Keywords:** Census, Administrative data, Record Linkage, Machine learning applications.

**1. Background**

From 2014 to 2017, Statistics Portugal (SP) developed a feasibility study on a new census model: linked data coming from administrative sources is to be integrated in the new Portuguese Statistical Population Dataset (SPD). From that point, SP has been also committed to continue to study the usability of the available administrative data for census purposes, considering the advantages of changing to a more efficient model and the need to prepare for the expected release of annual census data, according to Eurostat plans on future population statistics.

To estimate the resident population in Portugal, purely from administrative data sources or using a combined model, a specific objective of SP is to increase the micro data match rates, already achieved in exploratory studies, to better integrate records from administrative registers (***Table 1***).

**Table 1 – Administrative registers**

|  |  |
| --- | --- |
| **Acronym** | **Designation** |
| BDIC | Civil Register (Portuguese citizens) |
| SEF | Immigration Register |
| IRS | Taxes Register (Personal Income) |
| ISS | Social Security Register |
| CGA | State Pension/ Work Fund Register |
| QP | Private Employment Register |
| IEFP | Unemployment Register |
| EDUC | Education Register |
| ACSS | Patient Register (Hospital assistance) |

SP created a methodological framework to build a SPD integrating linked registers in a database with the maximum number of candidates to be resident and then applying a signs of life methodology to estimate actual residents, as described in references (SP, 2015; 2016). For 2015, the population estimated from the SPD was 10 434 161 individuals, with a deviation of 0.9% to the official resident population estimates (PE) for the same year. Exact comparison methods left out many potential matches (roughly more than 5% for most sources), as presented in ***Table 2***.

**Table 2 – Registers integrated and not integrated in 2015 SPD, by administrative source**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Registers** | **No** | **Registers integrated in the 2015 SPD** | | **Registers not integrated the in 2015 SPD** | |
| **No** | **%** | **No** | **%** |
| 2014 IRS | 9 370 879 | 8 969 050 | 95,7 | 401 829 | 4,3 |
| 2015 ISS | 6 927 720 | 6 678 767 | 96,4 | 248 953 | 3,6 |
| 2015 EDUC | 1 777 732 | 1 667 252 | 93,8 | 110 480 | 6,2 |
| 2014 QP | 2 609 046 | 2 584 267 | 99,1 | 24 779 | 0,9 |
| 2015 CGA | 1 032 133 | 1 001 865 | 97,1 | 30 268 | 2,9 |

To identify the highest number of linked pairs of records, cooperation between SP and the Academy – Instituto Superior Técnico/INESC-ID, proposed an alternative linkage model (Silva *et* al., 2017). In the next sections we present the processes of matching and the improvements in probabilistic record linkage methods that resulted from that cooperation.

**2. SPD matching overview**

* 1. *Initial conditions*

Unlike other countries that have already made the transition to a register-based or a combined census model, in Portugal there is neither a central population register nor a unique personal identification number (Id number). Portugal does not have a unified identification system: some registers use a single Id number, other use two or even three id numbers. The most representative identifiers are the Civil Register Identification Number (NIC), the Tax Identification Number (NIF) and the Social Security Identification Number (NISS). The Resident Permit Number, for immigrants, is an independent number of the Civil Register and doesn´t usually have an acronym. Considering these three Ids, the Civil Register only uses the NIC; the Taxes Register only uses the NIF; other registers have one or several numerical identifiers. Additionally, the fact that some sources don´t have full coverage on these numbers, as shown in ***Table 3***, introduces additional difficulties.

**Table 3 – Coverage of available numeric identifiers, by Register (%)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Administrative Register** | **NIC** | **NIF** | **NISS** |
| 2015 BDIC | 100,0 | - | - |
| 2015 SEF\* | 100,0 | 62,2 | 50,8 |
| 2015 ISS | 81,5 | 98,8 | 100,0 |
| 2014 QP | - | - | 99,1 |
| 2015 IEFP | 100,0 | 99,6 | 98,8 |
| 2015 CGA | 77,0 | 81,9 | - |
| 2015 EDUC | 90,9 | - | 66,6 |
| 2014 IRS | - | 100,0 | - |

\*\* We consider SEF resident permit number equivalent to NIC for this purpose

Thus, for matching purposes, we use the numeric Id keys (preferably) and, when not available, use the remaining individual attributes.

* 1. *The anonymization process*

To guarantee the privacy of the individual data, the National Data Protection Authority (CNPD) allowed SP to access the registers (Deliberations 06.2014 and 01.2017), after some transformations:

* Personal numeric identifiers (NIC; NIF; NISS) encrypted with a hash function to convert them into a condensed representation of fixed value in a one way irreversible process;
* Person names had to be truncated to the first three letters of the first name and the three final letters of the family name.

SP provided software to all source owners to perform these transformations on datasets with personal identifiable information attributes. These were applied before the transfer of the datasets to SP.

A final note to the implications of the anonymization process in the matching quality: as it is not possible to revert to the original data once the transformations are applied, a significant additional effort is required to validate the data, given that there is no integrated system and each one of the registers is independently managed.

* 1. *Matching methods*

Every matching process starts by applying a set of record linkage methods to the data, including data cleaning, standardization and indexing. After these procedures, a similarity vector is created for each pair of records; results can be classified as matching, no matching or possible matches. The matching methods used to link together individual personal records were:

* Deterministic, identifying identical keys;
* Probabilistic, matching similar (approximate) keys.

After applying these automatic methods, the application of a manual matching process, based on clerical searching would further improve the record linking activity. However, the truncations on names and addresses provided to SP preclude that final step.

* + 1. *The use of matching keys*

The record linkage process runs in two main stages:

* Using hashed numerical matching Ids (NIC, NIF or NISS);
* Using other matchable attributes (for example, first three letters of the first name, last three letters of the family name, sex, date of birth, place of residence).

A pair of records is classified as matched if the two records agree in all characters of the key and that key is unique (exact deterministic).

After linking together pairs of records from different registers, it was possible to create a table of candidate residents in Portugal – this is the foundation for the SPD. For the 2015 edition of the Portuguese SPD, there were about 14 million candidates (the resident population estimates for the same year are 10,3 million).

* + 1. *Probabilistic matching*

The 2011 and 2015 SPD were created using essentially deterministic matching keys. A previous exercise in the 2011 SPD (using the 2011 Census as a benchmark) was made using *IBM Quality Stage* software, but the obtained results were not statistically robust. That method was no longer applied.

Recently, back in 2016/2017, an independent research was conducted to answer this problem: optimize the record linkage between administrative data sources (see ***Table 2*** back in Section 1). These recent developments in matching methods will be presented in the next section.

1. **Matching developments**
   1. *Score based matching with logistic regression*

The methodological framework and the results presented in this section are explained in detail in two master thesis developed during almost one year of institutional cooperation (Sampaio Velho, 2017; Silva, 2017).

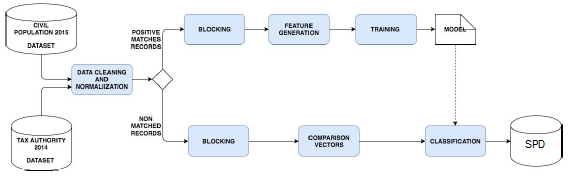
* + 1. *Methodological options*

Considering machine learning techniques for classifying record pairs as matching or not, it seems appropriate to explore supervised learning techniques: in the context of this work, a training set is generated, with each record labelled as matches or non-matches. A logistic regression model was chosen, given the linearity of the distance metrics used. This technique has been previously applied in similar context by other Statistics Institutes (ONS, 2013).

Using a logistic regression, we want to estimate the probability of a pair of records belonging to the same individual, by taking the edit distance between each records content as input features (Silva *et* al., 2017).

The proposed solution has two main phases: a learning or training phase and then a testing or classification phase. ***Figure 1*** illustrates the system architecture for matching between the BDIC and ISS datasets. The process starts by performing data cleaning and standardization and then selecting pairs of records to be matched. After these steps, there´s the training of the model step: a model capable of classifying matches and no matches.

**Figure 1. Record linkage System Architecture proposed by IST**



Source: Silva, 2017

Considering blocking (grouping records based on a common key), it was applied a standard blocking technique was applied, to speed up record matching (grouping records that are similar using a blocking key). For measuring string similarity Levenshtein edit distance was used (Levenshtein, 1966).

* 1. *Recent results*

With the methodological framework presented in the previous sub-section, eight models were generated for matching records from pairs of administrative registers. ***Table 4*** presents the results for precision and sensitivity metrics. Finally, ***Table 5*** presents the results of classification applied to the trained model (logistic regression).

**Table 4 – Quality of the models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Precision (%)** | | **Sensitivity (%)** | |
| **Matches** | **Non-matches** | **Matches** | **Non-matches** |
| BDIC\_IRS | 98 | 97 | 97 | 98 |
| BDIC\_ISS | 99 | 100 | 98 | 100 |
| BDIC\_EDUC | 93 | 99 | 94 | 99 |
| BDIC\_IEFP | 99 | 99 | 98 | 99 |
| BDIC\_CGA | 88 | 99 | 95 | 98 |
| SEF\_IRS | 97 | 97 | 97 | 98 |
| SEF\_ISS | 98 | 98 | 97 | 98 |
| SEF\_EDUC | 95 | 97 | 96 | 96 |

**Table 5 – Matching results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data sources matched** | **No. records** | **Records without common key** | **Exact matches** | **Exact matches validated** | **No. records not integrated in SPD** | **New matches** |
| **2015 BDIC** | 11 825 786 | 6 933 267 | 3 726 375 | 3 262 651 | 401 829 | 77 649 (19,3%) |
| 2014 IRS | 9 370 879 | 4 414 595 |
| **2015 BDIC** | 11 825 786 | 6 283 141 | 599 935 | 582 237 | 248 953 | 29 813 (12%) |
| 2015 ISS | 6 927 720 | 1 385 062 |
| **2015 BDIC** | 11 825 786 | 10 230 736 | 40 769 | 8 224 | 110 480 | 51 138 (46,3%) |
| 2015 EDUC | 1 680 018 | 84 968 |
| **2015 SEF** | 383 764 | 265 609 | 34 381 | 2 249 | 248 953 | 30 120 (12,1%) |
| 2015 ISS | 6 927 720 | 6 809 565 |

An extra result was the validation, through the probabilistic method, of previous deterministic matching performed by SP. ***Table 6*** presents the results obtained in the process of validation and evaluation, by logistic regression model, of the deterministic match keys process done by SP for 2015 SPD (see section 2.3.1). The average percentage for validated record pairs from BDIC – other sources is 96,2% (79,5% for SEF – other sources).

**Table 6 – Validation of the matched pairs obtained by match keys**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Register** | **No records** | **Resident candidates table (No records)** | **Validated records** | |
| **No** | **%** |
| **2015 BDIC** | 11 825 786 | 8 736 100 | 8 155 177 | 93,4 |
| 2014 IRS | 9 370 879 |
| **2015 BDIC** | 11 825 786 | 6 393 870 | 6 124 895 | 95,8 |
| 2015 ISS | 6 927 720 |
| **2015 BDIC** | 11 825 786 | 1 635 819 | 1 603 274 | 98,0 |
| 2015 EDUC | 1 680 018 |
| **2015 BDIC** | 11 825 786 | 712 163 | 677 836 | 95,2 |
| 2015 IEFP | 686 198 |
| **2015 BDIC** | 11 825 786 | 991 296 | 979 001 | 98,8 |
| 2015 CGA | 1 032 133 |
| **2015 SEF** | 383 764 | 143 950 | 120 404 | 83,6 |
| 2015 ISS | 6 927 720 |
| **2015 SEF** | 383 764 | 185 605 | 171 227 | 92,2 |
| 2014 IRS | 9 186 325 |
| **2015 SEF** | 383 764 | 16 848 | 10 576 | 62,8 |
| 2015 EDUC | 87 017 |

In this process, both matched pairs of records obtained by cross validation with a common numeric identifier and also pairs of records obtained by the probabilistic method described in Section 3.1.1 were considered correct.

Finally, considering the usability of this method (record linkage with logistic regression), namely the matching results in ***Table 5***, we can conclude that this was an experimental work and there is still much work to be done, particularly:

* Data aggregation of the results obtained on each model (that only uses a specific pair of registers and one numeric Id);
* Data deduplication (resulting from aggregation).

After these steps, clean results could be incorporated in the table of candidates for resident and might be considered becoming part of the final SPD. For the reasons identified, work is still in progress and these promising results were not included in a new version of the 2015 SPD. This method should be improved and integrated in the process of creation of the next editions of the Portuguese SPD.

1. **Conclusions**

Changing the census paradigm will take time and require huge investment. Work is still in progress for the transition to a register-based census. This line of research should continue until 2021 and beyond.

The results are encouraging and research to date has shown the potential to produce good quality population estimates at a national level. Nevertheless, for going to a less aggregated geographical level, improvements must be made:

* The setting-up of a more favourable legal framework with the creation of an integrated system to produce annual editions of the SPD is the key to overcoming problems; that includes reduce limitations on data access by national data authorities and commitment, by the data owners, on
  + sending data on a regular basis;
  + provide up to dated data;
  + increase data coverage (numerical Ids);
  + data cleaning;
  + ...;
* Refinements in the methodological framework:
  + Improvements in probabilistic record linkage (blocking strategy could use some more work and also determine an optimal threshold for matching; increasing automation of the process and explore more than pair-wise matching).

1. **References**

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics-Doklady 10, 707-710.

Office for National Statistics (ONS). 2013. Beyond 2011 ‘Matching Anonymous Data’ Methods & Policies Report (M9). Available at: <http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/methods-and-policies/index.html> (Accessed: 14 May 2018).

Sampaio Velho, L. *Emparelhamento de dados Censitários*. 2017. Thesis to obtain the Master of Science Degree in Information Systems and Computer Engineering. IST, Lisboa, 68p. Available at: <https://fenix.tecnico.ulisboa.pt/cursos/meic-t/dissertacao/1972678479053522> (Accessed: 14 May 2018).

Silva, R. ‘Matching Census Data Records’. 2017. Thesis to obtain the Master of Science Degree in Information Systems and Computer Engineering. IST, Lisboa, 58p. Available at: <https://fenix.tecnico.ulisboa.pt/cursos/meic-a/dissertacao/1409728525632007> (Accessed: 14 May 2018).

Silva, R., Sampaio, L., Calado, P., Silva, M. and Delgado, A. (2017) Matching administrative data for census purposes. In Proceedings of the Data Science, Statistics and Visualization Conference (DSSV 2017), Lisbon, Portugal. Available at: <http://iasc-isi.org/dssv2017> (Accessed: 14 May 2018).

Statistics Portugal (SP). 2015. *Interligação das diferentes bases de dados provenientes de fontes administrativas, no âmbito do novo modelo censitário para 2021*. 2021 Census Unit. Available at: <https://www.ine.pt/xportal/xmain?xpgid=censos21_uiadmin&xpid=CENSOS21&xlang=pt> (Accessed: 14 May 2018).

Statistics Portugal (SP). 2016. Metodologia de atualização da Base de População Residente – Construção da BPR 2015. 2021 Census Unit. Available at: <https://www.ine.pt/xportal/xmain?xpgid=censos21_uiadmin&xpid=CENSOS21&xlang=pt> (Accessed: 14 May 2018).