Matching Census Data

LUFIALUISO SAMPAIO VELHO, Instituto Superior Técnico, Portugal

A feasibility study is under way to enable Statistics Portugal to obtain part of the census information through administrative data sources. The process becomes complex because there is not a personal unique number, inconsistencies in the data and anonymised/pseudonimized data by determination of the Data Protection Authority (CNPD). This work presents an approach based on matching available data using Machine Learning methods. With the developed system, it was possible, for example, to detect 244,903 new matches between records of the databases of the Civil Population Register (Citizen's Card) and Tax Authority (IRS), representing an increase of 64.94%, and 47,836 new matches, an increase of 19.21%, with the Social Security database. The obtained results support the feasibility of the methodology and software developed for pairing the administrative data that are now available at Statistics Portugal.

Additional Key Words and Phrases: Record Linkage, Data Quality, Machine Learning, Census

ACM Reference Format:
Lufialuiso Sampaio Velho. 2017. Matching Census Data. 1, 1 (October 2017), 7 pages.

1 INTRODUCTION

Censuses are the source of much of the sociodemographic statistical information on the population and housing around the world [INE,Gabinete dos Censos 2021 2016b]. According to the [Office for National Statistics 2016], it is the only means of providing accurate information about the number of the population living in a given country, the size of the households, the characteristics of the population, the conditions and types of housing in which they live.
In most countries in the world, whether in Africa, Europe, Asia, or the Americas, this survey is conducted every 10 years.
Considering the financial costs, the burden on citizens to respond to surveys and the frequency required for these surveys, several countries, such as Germany, Poland, United Kingdom, New Zealand, Canada and Italy seek to migrate from the traditional survey-based model door to door, for new census models, based solely on administrative files or even hybrid models that combine the two.
A feasibility study is under way to enable Statistics Portugal, to migrate in Census 2021 to a combined model, based on administrative data sources complemented by surveys. Currently, nine administrative data sources from nine public administration agencies are available at Statistics Portugal:

- Civil Population Register (IRN),
- Tax Authority (AT),
- General Retirement Account (CGA),
- Institute of Informatics of Social Security (IISS),
- Ministry of Education - General Statistics of Education and Science (DGEEC),
- Immigration and Borders Service (SEF),
- Office of Studies and Planning (QP),
- Institute of Employment and Training (IEFP),
- Ministry of Health - Central Administration of the Health System (ACSS).

The purpose of Statistics Portugal is to integrate the above data sources to create a Resident Population Database (BPR), a relation in which each record contained in it corresponds to a citizen residing in Portugal in a given year. There are multiple hurdles to the creation of this database: records have inconsistencies and errors due to manually inserted data, and the Data Protection Authority (CNPD) imposes anonymisation criteria on the datasets. Attempting the record linkage between sources using exact comparison methods would leave out many potential matches. In these cases, it is common to use similarity measures to compare records complemented with rules or probabilistic classification methods to determine whether two records refer to the same person or not.
The data of all sources are coded in a way that protects the citizens privacy. Namely, for the quasi-identifier attributes of the individuals, the following transformations were made:

(1) Identifiers such as the Civil Identification Number (NIC), Tax identification Number (NIF), Social Security Identification Number (NISS) and Resident Permit (AR), are encrypted with a hash SHA256;
(2) The name of each individual is represented with the three first letters of the first name and the three last letters of the last name;
(3) For the address of the individual only the Postal Code information is available.

This dissertation proposes a methodology for matching the records of the different administrative data sources available at Statistics Portugal, taking into account that they generally do not share a common unique personal identifier, and quality metrics for the process.
The contributions of this work include a new classifier, to match records using a probabilistic method (*Logistic Regression*) and its use to evaluate the quality of previous matchings performed by Statistics Portugal during the construction of the prototype of the BPR in the 2011 and 2015 versions.
In the remaining of this paper, I start by describing the implemented methodology for matching the records available at Statistics Portugal, next the data quality monitoring metrics, then the obtained results, and finally the conclusions of this research. More details about this work can be found in my dissertation [Velho 2017].

2 RECORD MATCHING APPROACH

The machine learning method developed for record matching is implemented as a set of software scripts written in *Python* and *SQL*, which constitute the Information Production Module (IPM).
The purpose of the IPM is to produce evidence of residence of the population by means of successive matching of various pairs of data sources available at Statistics Portugal. It produces as a result the

Author's address: Lufialuiso Sampaio Velho, Instituto Superior Técnico, Lisboa, Portugal, lufialuiso.velho@tecnico.ulisboa.pt.

© 2017 Association for Computing Machinery.

Base Population Matrix, a relation that contains the register of the residency evidence for each individual potentially residing in Portugal in each year. It is from this *Matrix* that the Statistics Portugal derives its its *Resident Population Base* (BPR).
The Figure 1 illustrates, in Business Process Model and Notation (BPMN) [1], the process of generating and updating the BPR. Without loss of generality, this section the process used to match the records of the Civil Identification Database (BDIC)[2] and the Tax Authority Database (AT) [3], coming from the Civil Population Register and Tax Authority respectively. For other pairs of data sources, the process is similar.

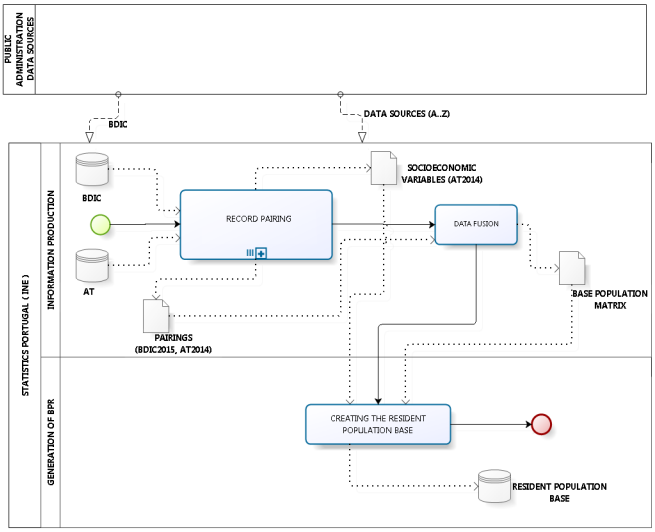

Fig. 1. BPR generation Process

The IPM has two submodules that correspond to two sequentially executed activities: Record Pairing and Data Fusion.
First, data is loaded from source into the Statistics Portugal DBMS and stored in tables with the originally received schema, which is variable from source to source. For simplification of information management, it is also useful to have the data of all years of residence standardized and stored in tables, one for each source. For example, in the case of BDIC and AT, the records of all years are stored in tables standardized with the same name. A standard scheme was defined for processing with the same *scripts* all the information sources.
The Record Pairing activity is responsible for receiving a pair of relations (e.g. BDIC and AT of a given year) and fill the output relation with the pairings made. After the Record Pairing, is necessary to assign an unique record for each individual. This is performed by the Data Fusion Activity. The output of this activity will allow to add new records to the *Population Base Matrix* or update it with new personal identifiers.
The Data Fusion module was not implemented within the scope of this work, since it would be trivial, because the Statistics Portugal has already established its own Data Fusion strategy based on an established hierarchy of sources reliability. The records in the Population Base Matrix filtered using residence rules established by Statistics Portugal to decide whether they will be added or not to the BPR [INE,Gabinete dos Censos 2021 2016a]. This process is represented in Figure 1 as BPR Generation, a software module for *Creating the Resident Population Base*.
Figure 2 illustrates also in BPMN, the sub-activities of the Record Pairing activity, as a relation between them during the Record Pairing process. The remainder of this section describes each one of the sub-activities, as well as how each was implemented: *Data Cleaning and Standardization, Blocking, Record Comparison, Training and Classification of Pairings*

2.1 Data Cleaning and Standardization

This is the first step to be executed in the Record Pairing Process ensuring that the data of successive years from each source are cleaned and stored in a common schema, with a provenance label associated; It receives two *inputs:* a relation with data obtained from the source (e.g BDIC) and a label to be assigned in order to identify the provenance of each record (e.g " BDIC2015 " to the data of the year 2015). The results are inserted in the corresponding normalized relation. e.g: BDIC -> BDIC _NORM.
The execution of the Data Cleaning and Standardization activity is performed by a *Python* script that loads the data from the original data source to the relation with the standardized data.

2.2 Blocking

Blocking is a strategy to reduce the computational cost caused by the high number of computations between the various pairs of records during the comparison stage. *Blocking* was used in this work with a dual purpose:

(1) Generate candidate matchings for each pair of data sources to ensure a reduction in the number of comparisons between record pairs from those sources;
(2) Generate training examples for classifier used in the subsequent activity. The examples can be *Positives* (correspond to records with pairs labeled as true match) or *Negatives* (correspond to records in advance labeled as unmatched).

The *Blocking* activity receives as *input* a pair of normalized relations (e.g BDIC_NORM and AT_NORM) considering the selection of provenance (e.g the provenance of BDIC = "BDIC2015" and AT="AT2014"). The output relation contains a set of candidate records to be paired and the positive and negative example to train the classifier.
The Blocking is performed using the *Standard Blocking* approach[Baxter et al. 2003; Christen 2012]. This technique was chosen initially because it is simple to implement and has later shown that good results can be obtained in most of the paired sources. As *Blocking criteria* I chose the concatenation of the *3 first letters of the first name and the date of birth* for data sources paired with the BDIC, and the concatenation of *country of naturality attribute and the date of birth*

[1] http://www.omg.org/spec/BPMN/2.0.2/
[2] relation that contains the registration of all citizens of Portuguese and Brazilian nationality with safe harbor status with residence address in Portugal or unknown address
[3] more specifically for the Citizens that present the declaration of Individual Income Tax (IRS)

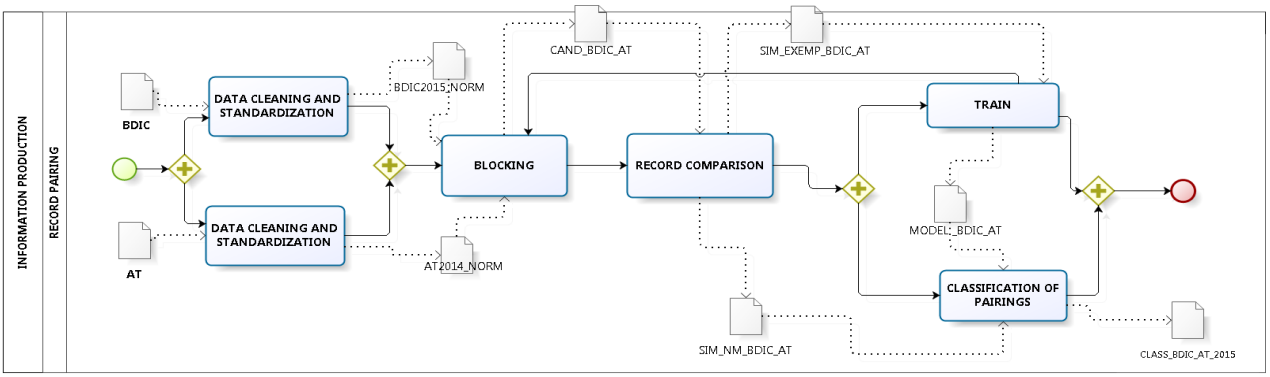

Fig. 2. Statistics Portugal Record Matching Process

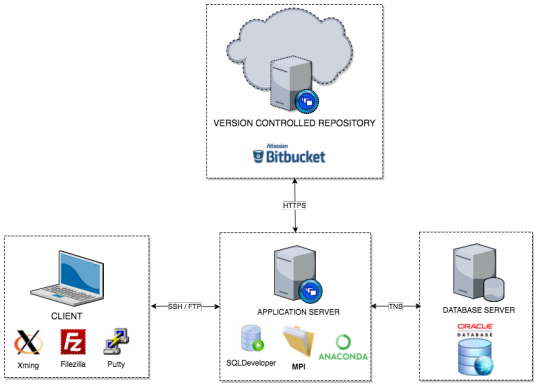

Fig. 3. IPM Execution Environment

for the data sources matched with the Immigration and Borders Service (SEF) database. This choice was based on a prior analysis of data quality of the records that could be matched by equality of identifiers, which were also considered as true matches.

The candidate records to be paired are generated by applying the Traditional Blocking approach to the record pairs that could not be matched by equality of identifiers. We add to each an attribute named *CLASS* to distinguish the matching candidate records from the examples for the classifier training. The unlabeled candidate matches have *CLASS=-1*.

The examples for training the classifier are generated applying the traditional blocking approach, which will produce records of two types: *Positives* (records labeled with CLASS=1) and Negatives (records labeled with *CLASS=0*). The records labeled with *CLASS=1* are the records paired after we apply the blocking strategy; the remaining records are labeled with *CLASS=0*.

2.3 Record Comparison

This activity performs the comparison between the record pairs coming from the *Blocking* step, using a similarity measure. The records comparison is preceded by the selection of a set of relevant attributes which may vary from relation to relation for instance, the attributes that identify the naturalness of the individual are relevant for pairing BDIC and AT records but are not relevant for the case of BDIC and General Retirement Account (CGA) database by the fact that in the data source of CGA these attributes are not fulfilled. After the selection of attributes, the records are exported as a CSV file and the comparison are made by the specific *Python* module. The similarity measure used for the comparison is the *Edit Distance* (Levenshtein distance)[Christen 2012; Rieck 2011].

2.4 Training and Classification of Pairings

The training of classifier (*model*) is performed based on a set of examples *Positive* and *Negatives*. The model is used as a support for deciding what *CLASS* to assign to unlabeled record pairs or the unmatched record set. The model was trained using the *Logistic Regression* technique[Alpaydin 2004; Hastie et al. 2009]. The choice of this technique was due to the fact that it is widely used for exploratory data analysis, simple to implement and fast to execute.

The Classification of Pairings activity is intended to predict the *CLASS* where each pair of unmatched registers lies. This decision is made based on the trained classifier as described in the training step.

2.5 IPM Execution Environment

Fig.3 illustrates the execution environment of the Information Production Module(IPM). It consists of a relational database server, an application server for running Python code, and a software repository with version control.

Access to the Application Server is made through a Client terminal configured with the Windows environment (Windows 7), in which the following tools are installed: Xming, FileZila and Putty.

The Application Server is a Linux environment, where *Anaconda* (Python programming environment) and *SQLDeveloper* for accessing the Oracle Database Management System (DBMS) are installed.

In the Oracle Database server, are stored the tables created from the data files coming from the various Public Administration institutions, and all other tables created for the operation of the IPM. The IPM code is stored in a Version Controlled Repository (*Bitbucket*)[4].

[4]https://www.bitbucket.org/

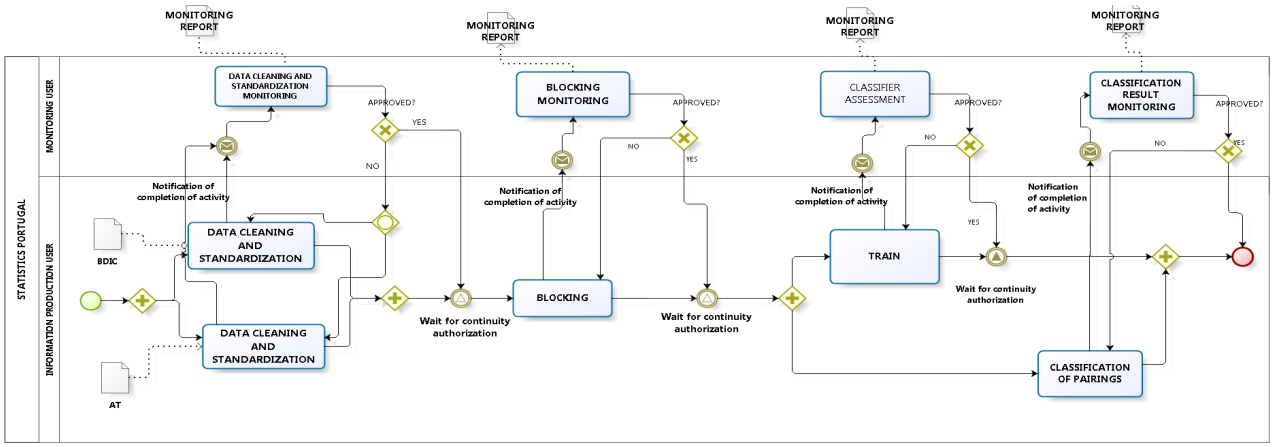

Fig. 4. Information Production monitoring process

3 DATA QUALITY MONITORING

The data quality of each Information Production activity is evaluated by the Monitoring Process. It allows to evaluate the quality of each Information Processing step.

Figure 4 illustrates the interaction between the activities of the Information Production Process and the corresponding monitoring activities. At each stage of the Information Production Process, a Monitoring activity takes place. The stages after the execution of Data Cleaning and Standardization, will require prior approval of data quality from previous steps. For example, to perform the *Blocking* activity requires prior validation of the data quality in the Data Cleaning and Standardization step.

The monitoring module is responsible to query the data produced by the Information Production Process, calculates Data Quality metrics, and provide the monitoring reports in steps of *Data Cleaning and Standardization, Blocking, Training and Classification of Pairings*.

The activity of Record Comparison discussed in Section 2 is not validated in the Monitoring Process, since the quality evaluation in this step can be determined from the monitoring results of the Blocking or later stages.

3.1 Data Cleaning and Standardization

The evaluation of data quality in this step is performed based on the metrics of *completeness [Batini and Scannapieco 2016; Sidi et al. 2012; Taleb et al. 2015], null count* and *uniqueness* [Abedjan et al. 2015] for each data source indicated in the Input.

Taking into account that this stage is key to the success of the later stages of Information Production, it should be verified that the level of quality in the data to be used for the new pairings is maintained or improved. To do this, we can compare the metrics obtained with the data batches from previous years/loads with the new batch to be processed. As an example, after executing the monitoring module with the Civil Population Database (BDIC) for 2014 and 2015 data, we could observe that:

(1) The number of distinct records found in the 2014 batch is higher than in the batch of 2015 in about 0.497%. This is due to the fact that the birth ratio is lower than the mortality ratio during this period;
(2) The Completeness ratio in the *Last name, Date of birth and Residence* attributes are below 100% (e.g. in terms of the date of birth, we could find about 54 individuals with unknown birth date (Null values)). The choice of any of the attributes to form the blocking criterion will lack a thorough analysis of the impact of attributes with completeness less than 100% in subsequent steps and in the final outcome of the process.

3.2 Blocking

The measurement of the quality of this step is based on the the Reduction Ratio [Baxter et al. 2003; Elfeky et al. 2002] at which a given blocking criterion is able to affect the number of comparisons between each pair of relations to be paired. Considering as an example the pairing of BDIC and Tax Authority (AT) with 11,825,786 and 9,370,879 records respectively, applying the Blocking strategy we were able to obtain a Reduction Ratio of 99.99%. This means that, we were able to select with great accuracy the potential candidates for pairing.

The applied approach allows to obtain on average about 3 similar candidates in each block. Although it has a higher computational cost (it is necessary to analyze the triple of the pairs of registers compared to the exact method), it remains nevertheless treatable and with the advantage of guaranteeing higher number of correct pairings with respect to the exact method without significantly degrading the error.

3.3 Training

The quality of the Training stage is measured based on the accuracy of the Classification model. The technique used for this purpose is a *2-fold Cross Validation* [Han and Kamber 2006] applied to the set of examples extracted from the pair of relations to be paired.

Once the *2-fold Cross Validation* is performed, the *Precision* and *Recall* measures are calculated for each class of the examples used[Christen 2012]. The measure of accuracy of the model will dictate whether

it can be used or not for the classification of new records. For the case of this work, I consider as satisfactory a level of recall with an error of 3%. This value was established based on the coverage error verified in the Quality Surveys (IQ) of the Door-to-door Census of 2011 in Portugal, set at 2.5% for the whole country [INE 2012].
Given that data sources do not have attributes with a completeness ratio of 100% in most cases, for example in the case of EDUC (Education and Science), I choose to establish a value not much higher but equated to the 2011 Census as *baseline* for the quality of the intended Classification model, although it has passed a certain time up to the present date.
From the data sources available at Statistics Portugal, eight pairs were paired. A model was trained for each pair, since each one has specific characteristics make it differ from the others. Considering the Blocking Strategy approach used to generate the training examples, some discrepancies were found relating to the number of positive examples *versus* negatives for each pairing observed during the training of the classifier (e.g. From the Classifier model of BDIC and Social Security (IISS) used 44.912 positive and 955.088 negatives examples). The unbalancing is due to the fact that there are blocks with a large number of negative examples, each having only one positive example. This approach for negatives generation could be improved by choosing at most two to three negative examples for each positive example, whose similarity is closest to the positive example.
All the classification models trained presented an optimal accuracy, generally having Precision and Recall between 97% to 99% for the positive examples and 97% to 100% for negative examples. These measurements are as intended, with the exception of the models generated with EDUC and General Retirement Account (CGA) sources which achieved a Precision and Recall under 97%.

4 RESULTS OF PAIRINGS

Table 1 presents the statistics of the records paired by Data Source and the results of the pairing using the probabilistic methods presented in this Work.
Statistics Portugal has obtained before pairings between the different data sources available, using exact methods.
In addition to finding new matches, one of the purposes of this work was to evaluate the pairings previously made by Statistics Portugal. To do this, the total records found in the Matrix in Table 1 correspond to the records with common pairings between the probabilistic method and the exact methods in the Population Base Matrix.

The Data Quality in the Classification of Pairing stage is analyzed from the New Matches found. At this stage we measure the contradictions and uncertainties as described by Bleiholder and Naumann observed in the results of the pairings performed, and see if these new matches are good or not. Table 2 presents in greater detail the Data Quality results of the new matches obtained from BDIC and AT based on the data quality measures referred. Among the various pairs of attributes that can be compared, those that have a high rate relative to the others in the two metrics are:

(1) Degree of contradictions between attribute pairs:
Concerning contradictions, it is worth mentioning those observed in the attribute RESID_POSTAL_LOC because they have a higher value compared to the others. These are mostly data insertion problems. For example:
   (a) In **BDIC**: São João dos Montes and in **AT** is written as: São João Montes
   (b) In **BDIC**: Duas Igrejas PNF and in **AT** is written as: Duas Igrejas
(2) As for uncertainties, the greatest degree lies in the attributes of residence, of which it was possible to observe:
   (a) For the total existing in attributes DISTRICT, COUNTY and PARISH of residence (RESID) consist of pairings where the value of these attributes in the AT data source is undefined (NULL);
   (b) A total of 61.2016 records whose value of attributes representing the first four and last three numbers of Post Code (POST_COD) are undefined in the AT data source;
   (c) A total of 62.091 records whose value of attributes representing the location of Post Code (POSTAL_LOC) is undefined in the BDIC data source.

5 CONCLUSIONS

Five of the eight pairs of data sources involve the Civil Identification Database (BDIC) and the other three pairs refer to pairings performed with the Immigration and Borders Service Database (SEF).
The quality of the data is good in most of the sources, with the exception of some where the completeness is very low in several attributes (in particular the General Retirement Account (CGA) and Education and Science (EDUC)).
The trained classifiers, which use logistic regression, achieve in general an F1-score of 97%, which is within what would be necessary, considering as baseline the coverage error rate of the quality survey of 2011 Census (2.5%).
The new pairings found are expected to add a substantial number of links to the Statistics Portugal Resident Population Base (BPR), in the order of 64.94% of the 401,829 records not matched with the Tax Authority (AT) and 19.21% of the 248,953 records not matched with the Social Security Database (IISS). The total of unmatched records referred in the two cases are referenced in the Statistics Portugal Technical Report for updating the Resident Population Base on the *Metodologia de atualização da Base de População Residente - Construção da BPR 2015 (QUAR 2016)* [INE,Gabinete dos Censos 2021 2016a].
Using the pairings produced with the process presented in this paper to validate the pairings previously obtained by Statistics Portugal using exact methods, the following conclusions can be made:

- **Pairings with BDIC:** were validated in the best case 98.75% of total pairings made by exact methods (pairing between sources BDIC - CGA) and in the worst case 93.35% (data sources BDIC and AT). Since AT is a transversal data source to others data sources such as the IISS, EDUC, the Institute for Employment and Vocational Training (IEFP) and CGA, many of pairings validated by these pairs can improve the total

Table 1. Pairings of the Probabilistic Method by Data Source Pair

Data Sources	Records to pair (a)	Records Paired (b)	Records found in Matrix (c)	New Matches (d)
BDIC (2015)	6.933.267	3.631.740	3.262.651 (84,88%)	244.903
AT (2014)	4.414.595			
BDIC (2015)	6.283.141	4.667.315	582.237 (68,4%)	47.836
IISS (2015)	1.385.062			
BDIC (2015)	10.230.736	61.943	8.224 (20,17%)	51.138
EDUC (2015)	84.968			
BDIC (2015)	11.203.211	74.468	55.260 (61,68%)	11.974
IEFP (2015)	63.622			
BDIC (2015)	11.014.943	300.823	168.158 (93,18%)	60.545
CGA (2015)	209.642			
SEF (2015)	253.742	32.823	2.249 (8,71%)	30.120
IISS (2015)	624.118			
SEF (2015)	220.315	66.028	7.990 (35,72%)	52.177
AT (2014)	9.023.088			
SEF (2015)	375.872	23.381	2.691 (30,02%)	12.796
EDUC (2015)	79.132			

(a) Total records to pair by Data Source pair
(b) Total records matched by Probabilistic Model per pair of Data Sources.
(c) Total of common pairings between probabilistic and exact methods, accompanied by the total percentage that it represents.
(d) Total records added by Probabilistic Model. Refers to new pairings found by pair of data sources.

Table 2. Assessment of the quality of pairings BDIC - AT

Attribute	Contradiction (Records)	Contradiction (%)	Uncertainty (Records)	Uncertainty (%)
NAME_3FIRST	0	0.0	0	0.0
NAME_3LAST	2527	1.03184	1	0.00041
SEX	1380	0.56349	0	0.0
Y_BRITH	0	0.0	0	0.0
M_BIRTH	0	0.0	0	0
D_BRIRTH	0	0.0	0	0
NAT_DISTRICT	399	0.16292	1790	0.73090
NAT_COUNTY	664	0.27113	1790	0.73090
NAT_PARISH	1433	0.58513	1790	0.73090
NAC_ISO	2664	1.08778	0	0.0
RESID_DISTRICT	0	0.0	244903	100.0
RESID_COUNTY	0	0.0	244903	100.0
RESID_PARISH	0	0.0	244903	100.0
RESID_POST_COD_4	3625	1.48018	61206	24.99194
RESID_POSTAL_COD_3	5163	2.10818	61206	24.99194
RESID_POSTAL_LOC	81191	33.15231	62091	25.35330
RESID_ISO	7050	2.87869	0	0.0

percentage of pairings between BDIC and AT, also reflecting for the cases of the remaining pairs of pairings.

- **pairings with SEF:** were validated in the best case 92.25% of all pairings made by exact methods (refers to pairing between sources SEF - AT) and validated 62.77% in the worst case (refers to pairing between sources SEF - EDUC).

Among the contributions resulting from this work, it was possible to produce a probabilistic pairing method that, as a result of the applied blocking technique, allows to select with great accuracy the potential candidates for pairing. The approach allows to obtain on average about three similar candidates in each block. Although it has a higher computational cost, it remains nevertheless treatable and with the advantage of guaranteeing higher number of correct pairings with respect to the exact method, without significantly degrading the error.

The evaluation of the quality of the pairings produced allows to gauge how good or bad were the methods used for pairing records. For example, the evaluation of the new pairings produced between

the data sources BDIC and AT, based on the level of inconsistencies between the pairs of more significant demographic attributes of the individual, allows to conclude that of the total of 244,903 new pairings, no more than 1.087% of these records may be excluded if the attribute that identifies the nationality of the individual is considered to be of greater relevance in terms of the individual's identification or a smaller percentage if the non-residence attributes are considered, since that the individual residence information can be changed periodically.

Based on the results obtained, it is possible to support the feasibility of using the methodology and the probabilistic matching software for the administrative data sources available to Statistics Portugal. It is also worth mentioning that the present work is in its first stage of development, so significant improvements in the methodology, as well as the entire software component, will be necessary in order to enable Statistics Portugal to construct its BPR in fast and effective way.

This work is a first approach for assessing the feasibility of using probabilistic methods in the pairing of administrative data for census purposes in Portugal. Therefore, there are many improvements and additional developments to consider before the developed system comes to be used in production. The following improvements could be considered:

- Implementation of a strategy for the combination of multiple Blocking criteria to improve the activity of candidates generation;
- Improvement of the negative examples method generation for the training of the Classifier. The two or three examples in each block that best resemble the positive example known in each block could be chosen as negative examples;
- Provide a graphical user interface with a *Dashboard* for invoking the processing/analysis and visualization steps of data statistics and made pairings and to allow administrative review (*Clerical Review*);
- Pairings from the previous years versions of the data sources could also be used in a way to be used as evidence for pairings in the current versions.

REFERENCES

Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. *The VLDB Journal* 24, 4 (2015), 557–581.

E. Alpaydin. 2004. *Introduction to Machine Learning*. Vol. 53. MIT Press, London, Engand. 1689–1699 pages. https://doi.org/10.1017/CBO9781107415324.004

Carlo Batini and Monica Scannapieco. 2016. *Data and Information Quality: Dimensions, Principles and Techniques.* Springer International Publishing. 500 pages.

Rohan Baxter, Peter Christen, Tim Churches, et al. 2003. A comparison of fast blocking methods for record linkage. In *ACM SIGKDD*, Vol. 3. 25–27.

Jens Bleiholder and Felix Naumann. 2006. Conflict Handling Strategies in an Integrated Information System. *Proceedings of the IJCAI Workshop on Information on the Web* 197 (2006), 1–13.

Peter Christen. 2012. *Data Matching: Concepts and Techniques for Record Linkage,Entity Resolution and Duplicate Detection.* Springer-Verlag Berlin Heidelberg. 272 pages. https://doi.org/10.1007/978-3-642-31164-2

Mohamed G Elfeky, Vassilios S Verykios, and Ahmed K Elmagarmid. 2002. TAILOR: A Record Linkage toolbox. In *Proceedings 18th International Conference on Data Engineering*. 17–28. https://doi.org/10.1109/ICDE.2002.994694

Jiawei Han and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques* (2nd ed.). Vol. 12. Morgan Kaufmann. 744 pages. https://doi.org/10.1007/978-3-642-19721-5

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning* (second ed.). Vol. 1. Springer. 1–694 pages. https://doi.org/10.1007/b94608

INE. 2012. *Censos 2011, Resultados Definitivos.* Technical Report. Instituto Nacional de Estatística (INE), Lisboa. URL: https://goo.gl/XuYkvZ (acesso em 27 Set. 2017), isbn = 9789892501857. 7905 pages.

INE,Gabinete dos Censos 2021. 2016a. *Metodologia de atualização da Base de População Residente - Construção da BPR 2015.* Technical Report. Instituto Nacional de Estatística, Lisboa. URL: goo.gl/fmo6Z7 (acesso em 02 Nov. 2016).

INE,Gabinete dos Censos 2021. 2016b. *Novo modelo censitário - Estudo de viabilidade Programa de Trabalho.* Technical Report. Instituto Nacional de Estatística, Lisboa. URL: https://www.ine.pt/xurl/doc/265780886 (acesso em 10 Out. 2016).

Office for National Statistics. 2016. *Census Transformation Programme. Annual assessment of ONS' progress towards an Administrative Data Census post 2021. May 2016.* Technical Report May. URL:https://goo.gl/nNja21 (acesso em 04 Out. 2016).

Konrad Rieck. 2011. Similarity measures for sequential data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, 4 (2011), 296–304. https://doi.org/10.1002/widm.36

F. Sidi, Payam Hassany Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha. 2012. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*. 300–304. https://doi.org/10.1109/InfRKM.2012.6204995

Ikbal Taleb, Rachida Dssouli, and Mohamed Adel Serhani. 2015. Big data pre-processing: a quality framework. In *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE, 191–198.

Lufialuiso Sampaio Velho. 2017. *Emparelhamento de Dados Censitários.* Master's thesis. Instituto Superior Técnico, Universidade de Lisboa. https://goo.gl/M3mzEK