# A system developed for solving the matching problem in the Brazilian Census Post Enumeration Survey

Layter Xavier, Vinicius
*Brazilian Institute of Geography and Statistics, Brazil*
*Avenida Republica do Chile 500 – 10 andar*
*Rio de Janeiro, CEP 20.031-170, Brazil*
*E-mail: viniciuslx@gmail.com*

## ABSTRACT

In order to estimate the coverage rate of the 2010 Brazilian Census, a Post Enumeration Survey (PES) was conducted. To obtain the estimates using the Dual System Estimation method registers from the Census database and the PES database have to be matched. To do this, some functions of the library RecordLinkage in R system were used. These functions implement Fellegi-Sunter approach to calculate weights using the EM method. After that, a one-to-one association was performed using the solve_LSAP function of the library clue, that solves the linear sum assignment problem. To implement the method a large-scale experiment was performed to select informative variables, string metrics and respective thresholds for each variable.

## 1. Introduction

Brazil is a country with continental dimensions and social and cultural diversities. Some problems arise from these diversities since unanswered questions are more often in areas with concentration of high income families. Besides, the difficult access to places that are distant from the urban areas makes the data collection a complex task.

The quality of population and housing census data is very important and the purpose of the census evaluation is to provide users with a level of confidence when utilizing the data and to explain possible errors in the census[1].

It is well known that population census is not perfect and that errors can and do occur at all stages of the census operation. Errors in the census results are classified into two general categories: coverage errors and content errors. Coverage errors are the errors that arise due to omissions or duplications of persons or housing units in the census enumeration. Content errors are errors that arise in the incorrect reporting or recording of the characteristics of persons, households and housing units enumerated in the census. Many countries have recognized the need to evaluate the overall quality of their census results and have used various methods for evaluating census coverage as well as certain types of content errors[1].

The Post Enumeration Survey (PES) has been conducted by the Brazilian Institute of Geography and

[1] Post Enumeration Surveys, Operational guidelines, Technical Report, World Population and Housing Census Programme, UNITED NATIONS SECRETARIAT DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, STATISTICS DIVISION.

Statistics (IBGE) in all census since 1970 in order to provide information to estimate the coverage levels as well as to provide subsidy to plan the following census.

The Dual System Estimation Method is used to calculate the coverage rate of the census. To be able to apply this method it is necessary to know if each PES dwelling and its residents were either counted or omitted in the census. To do this, a matching among the registers from the census database and the PES database is made.

Many procedures have changed from the previous PES´s to the 2010 Demographic Census PES. The main innovation is the use of computational tools to make the data comparison process automatic. For the matching process, a large number of computer programs were studied and tested. Also, an extensive study was conducted to choose the best statistical model to be used.

IBGE developed a tailor-made computer system to execute the matching process since the assessed programs have not fulfilled IBGE needs. This paper aims to describe the development of this system and to present the system itself.

## 2. General Aspects of the Matching System

### 2.1 The Programming Environment

The system that performs probabilistic record pairs was developed in the R environment. R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger and access to certain system functions. Also, R has the ability to run programs stored in script files. (http://www.r-project.org/)

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. Most of the user-visible functions in R are written in R. It is possible for the user to interface to procedures written in C, C++ or FORTRAN languages to obtain more efficiency. The R distribution contains functionality for a large number of statistical procedures, e.g. linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing. There is also a large set of functions which provide a flexible graphical environment for creating various types of data presentations. Additional modules ("add-on packages") are available for a variety of specific purposes (see R Add-On Packages). (http://www.r-project.org/). Many packages were extremely important for the development of the system that performs probabilistic record pairs and they will be mentioned throughout the paper.

### 2.2 The Record Pairs, Standardization and Variable Selection

The data standardization is essential in the record pair process because it directly affects the data collection quality. The training of the census-takers has an important point on this, avoiding collection and classifications errors.

There were pre-defined answers for some questions in the questionnaires and no standardization was needed. When editing was allowed, a few standardization was made such as standardization to deal with abbreviations.

There are different aspects in the pair matching among rural housing and urban areas. On one hand it is not so difficult to find a dwelling by the address only in the urban area, without extra information from the family. On the other hand, it is sometimes difficult to find addresses of farms and properties in the rural area because there is no standardized addressing and they are known by their own names without numbers in many cases. The information about the family is extremely important in the rural areas.

The initial idea was to develop the system with two different processes for rural and urban areas. The two main differences were in the input variables of the models and in the criteria to consider a pair.

However, problems have arisen because neither rural sectors nor urban sectors had all their specific characteristics. Also, there were sectors presenting both urban and rural features. The solution found was to make a variable selection function that depends only on the sectors data and not on the origin. In none of the programs tested during the studies there was the variable selection function. Neither were they defined previously.

For each variable the number of intersection distinct value at census and at PES about the minimum amount of distinct values at census or at PES is calculated. There is a limit of exclusions which is compared to this fraction and, based on this, exclusion of variables is made. It was also adopted a criterion of exclusion based on entropy.

The variables that presented no changes in the PES or census files were excluded from the model because they do not have any discrimination influence.

The number of distinct intersections at the Census and at the PES file is calculated for each variable. There is a limit of exclusions which is compared to this fraction and, based on this, exclusion of variables is made. It was also adopted a criterion of exclusion based on entropy.

There are variables in which it is common to occur variations for the answers from the census and the PES for the same household. The use of these variables has to be made with caution. One example is to change the name of the head of the household by the name of his spouse.

There are also variables that agree with many others but this agreement does not mean that real pair was found. There are different households presenting the same first name for the head of the household.

The variable transformation was a way to improve the variable discrimination power. Knowing the variables name of the head of the household and the spouse, two new variables are created because there is a change between them. In the census database two equal variables are created, which are the concatenation of the first name of the head of the household and the first name of the spouse. In the PES database two variables are also created: the first one is the concatenation of the first name of the head of the household and the spouse and the second one is the concatenation of the first name of the spouse and the head of the household. Having the similarity string function JaroWinkler, they are then compared two by two. For example:

| Census | | PES | |
|---|---|---|---|
| VAUX1 | VAUX2 | VAUX1 | VAUX2 |
| JoaoMaria | JoaoMaria | JoaoMaria | MariaJoao |

JaroWinkler(VAUX1, VAUX1 ) = 1

JaroWinkler(VAUX2, VAUX2) = 0.54

The highest JaroWinkler score represents the real combination and only this one is used. Due to this, the variable discrimination power increased a lot. The household matching of some sectors was tested with this variable and with the variable number of men and the variable number of women in each household. Good quality of household matching was obtained without any address information.

### 2.3 The Model used in the Matching System

A specific library called RecordLinkage provides functions and data structures that make the evaluation of record linkage methods easy and facilitates the applications of record linkage to different data sets.

According to the authors of the RecordLinkage packages, these methods can be divided into two

classes. One class consists of Stochastic Methods which are based on the framework of Fellegi and Sunter for record linkage. The other class comprises non-stochastic methods from the machine learning context. Using this library, many methods were tested before the final model.

The EM was applied in the Brazilian PES and it is provided by the package RecordLinkage. To use EM is necessary to make all possible matching combinations and for each pair the JaroWinkler similarity function is used for all variables. For each variable there is a limit for the JaroWinkler function, and if the value of the function is higher than this limit it is considered that the variable agrees.

EM has the function to provide weight of agreement for the pairs and, according to the Fellegi and Sunter theory, it is an important concept developed in an optimal decision procedure for record linkage. A record pair is classified as a match if the composite weight is above a threshold value, and a non-match if the composite weight is below another threshold value. An undecided situation is found when the composite weight is between these two thresholds.

It is not possible to apply only one classification limit because there is variation of weight among the tracks. The limits for the pairs which are considered matches are given based on the larger possible number of pairs for a track.

### 2.4 Assignment

One record on census file can be assigned to one and only one record on PES. For this reason an optimal scheme was chosen to maximize the sum of the composite weights of record pairs assigned. The function solve_LSAP of the library clue was used for this. This function applies a Hungarian algorithm which performs fast and efficiently one-to-one associations. Based on the pairs classified as a matches a matrix containing the composite weights is built. This matrix is an argument of the function solve_LSAP.

### 2.5 A Function to Avoid False Pairs

Based on the knowledge of correct matches, it was build a function which assesses the pairs based on some agreement combination in the variables. If one pair respects one agreement combination, the pair is considered true, otherwise it is considered false.

### 2.6 Undecided and Non-matched Pairs

The pairs whose scores provided by the EM are smaller than the matches, the undecided or the non-matched have not been assessed yet. The association is repeated but now with the matrix with all EM weights forming new pairs. Some of the new pairs are similar to the ones previously formed considered real pairs. For all the others, the function that avoids false pairs is applied again but now using more flexible agreement, allowing some false pairs but keeping the function to eliminate the really false ones. These pairs are viewed by a system which makes the verification of the pairs easier. The pairs considered false are unmatched. For all unmatched observations a manual pair matching is tried.

## 3. Conclusions

The knowledge about the variation in the addresses and the regional differences made a better unde rstanding about the data structure, allowing the system to be adapted to the regional differences.

Many tests and adjustments were made to improve the system until its current version. The automa ted matching system greatly exceeded the expectations in terms of both match rate and accuracy, fulfillin g the PES needs.

## Acknowledgements

The author wishes to acknowledge the help of others on the team that played a key role in this work:

Andrea Diniz da Silva and Professor Djalma Galvão Carneiro Pessoa who was responsible for the implementation of the system.

## REFERENCES (RÉFERENCES)

R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

UNITED NATIONS. Post Enumeration Surveys, Operational guidelines, Technical Report
Word Population and Housing Census Programme , Secretariat Department of Economic and Social Affairs, Statistics Divisions

Andreas Borg and Murat Sariyar,. RecordLinkage: Record Linkage in R.
R package version 0.2-2,(2010), http://CRAN.R-project.org/package=RecordLinkage
Murat Sariyar and Andreas Borg , The RecordLinkage Package: Detecting Errors in Data, R journal, 2010-2

Diniz da Silva, A. , et al. Inovações no Sistema de Pareamento de Domicílios e Pessoas para a Pesquisa de Avaliação da Cobertura da Coleta do Censo 2010. Diretoria de Pesquisas, IBGE. Rio de Janeiro, 2010.

Rainer Burkard, Mauro Dell'Amico, Silvano Martello, Assignment Problems, Society for Industrial and Applied Mathematics, Philadelphia, 2009, http://www.assignmentproblems.com/

Romeo, O. Situação do Pareamento Automático após a PA Experimental de Rio Claro/SP. Diretoria de Pesquisas, IBGE. Rio de Janeiro, 2010

Fellegi and A. Sunter. A theory for record linkage. Journal of the American Statistical Association, 64:1183–1210, 1969.

Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer and Andreas Weingessel. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-24. 2010
http://CRAN.R-project.org/package=e1071

Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457, 2002 http://cran.r-project.org/web/packages/class/index.html

Matthew A. Jaro, Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association Vol 84 No 406

M. Culp, K. Johnson, and G. Michailidis. ada: Performs boosting algorithms for a binary response, R package version 2.0.1 http://www.stat.lsa.umich.edu/~culpm/math/ada/img.htm