Study of Record Linkage Software for the 2010 Brazilian Census Post Enumeration Survey

Diniz da Silva, Andrea
Brazilian Institute of Geography and Statistics, Brazil
Avenida Republica do Chile 500 – 10 andar
Rio de Janeiro, CEP 20.031-170, Brazil
E-mail: adiniz@ibge.gov.br

SantAna Martins Romeo, Otavio
Brazilian Institute of Geography and Statistics, Brazil
E-mail: otavio.romeo@ibge.gov.br

Silva Soares, Thiago
Brazilian Institute of Geography and Statistics, Brazil
E-mail: thiago.s.soares@ibge.gov.br

Layter Xavier, Vinicius
Brazilian Institute of Geography and Statistics, Brazil
E-mail: Vinicius.Xavier@ibge.gov.br

1. Introduction

One of the biggest improvements of the 2010 Brazilian Census Post Enumeration Survey (PES) conducted by the Brazilian Institute of Geography and Statistics (IBGE) is the incorporation of new methodologies and technologies developed or improved throughout the last decade. The use of handheld devices for the data collection, already experienced in the 2007 population count, was one of the successful innovations in the Census project, allowing improvement of quality and timeliness in the data collection process. The use of a management system for data collection made possible to transfer data from the field to the main data processing units, providing indicators that allowed following up the data collection process and controlling its quality while still in course.

The use of new technologies had an effect on the structure and format the data collected was delivered immediately after its collection and facilitate to perform automatic matching of census to the post enumeration survey data.

The development of software packages that implements computational models for record linkage and matching has increased over the last 10 years, making possible to find several software that perform similar tasks. Such packages may improve the process and its results but, on the other hand, it requires a great deal of time and effort to decide which one better meets the Organization needs. Thus, to be able to choose the best software for linking records from the Brazilian Census and the PES, IBGE concentrated efforts in studying related software over the last two years.

The studies involved assessment of corporate and free (open source) software, taking into account not only operational aspects but also methodological issues of each package. Studies of mathematical models for matching were also conducted in parallel. The experiment was based on some documentation and on the available software trial licenses, considering their suitability to the Brazilian Census PES project.

The software assessed are Data Quality (by SAS), Quality Stage (by IBM), RecLink (by University of Rio de Janeiro State and Federal University of Rio de Janeiro), LinkPlus (by CDC-Atlanta), FEBRL (by Australian National University) and RELAIS (by Italian National Statistics Institute). The main findings on

this study are presented in the next sections. After considering the suitability of these packages to the Brazilian PES project, the tool finally used in the matching process of the Brazilian PES is presented in section 4.

2. Review of Some Record Linkage Software

Data Quality

The Data Quality software was developed by SAS Institute Inc. Its evaluation was based on the available documentation for the period of May 2009 and December 2010 and on a presentation made by a representative of SAS to staffs of IBGE.

Very little literature was found about Data Quality so that the SAS 9.1.2 Data Quality Server: Reference was the main source used. Data Quality is a package that enables you to analyse, cleanse, transform and standardise your data. Its documentation offers a comprehensive description on the functions, system options, macros and procedures available, all accompanied by a rich amount of examples on programming.

However, no detailed information on the methodological aspects used for matching or deduplication was found. Some information was also obtained from Data Quality website, but again no details about the methodological aspects of "Matching and deduplication" were found. Besides, the SAS representative could not clarify issues related to methodological aspects during his presentation, so that no information on the available methodologies and models for deduplication, space reduction or matching could be obtained for further evaluation.

Thus, the literature review and the exchange during the presentation led to the conclusion that the software is well developed for data cleanse, but information to allow evaluation of methodological aspects of the software is still needed. No trial license was available for the Data Quality, so it was not possible to assess the interface, usability and the different features of the software.

QualityStage – IBM

Developed and traded by the International Bussiness Machines (IBM), QualityStage is a component of IBM Information Server package integrated to a platform called DataStage. This software includes a server version and a client version. The latter only works with the server version installed and configured to the database that will be used by the client.

The QualityStage is supposed to be a tool for analysis and data standardization, as well as matching jobs, deduplication and survival processes. For data analysis, the software allows to set type equal to "default" or to a particular character of the variable of analysis. It is useful for checking the composition of the database prior to perform the matching or deduplication. Besides, concatenation of variables can be done for example concatenating variables "sex" and "name" for frequency analysis.

For standardization, it is possible to build a dictionary to recode values of certain variables by pre-established values in this dictionary. For example, for variable "sex" it can be build a dictionary where "M" or "Man" corresponds to "male" and then all the original values would be replaced by the new codes. It is also possible to split the variables according to defined criteria. For example, for variable "address", a value like "street", "avenue" or "road" is "type of parks". The survival process consists of choosing a writing standard and using it for all records. For example, month of birth can be typed in the database in different ways, e.g. January, Jan, 01, 1, etc), the process then chooses a form of survival, which can be January for example, and replaces all the others by it.

Matching and deduplication are the most important functions for the 2010 PES, so they were analyzed in more detail. Deduplication is nothing more than the matching done in the same database, so all the functions described for the matching are also available for deduplication.

There are four types of matching available in the software: many-to-one, many-to-one multiple, many-

to-one duplicates and one-to-one. For the 2010 PES, only the last one is employed because it is the only one in accordance with the principle of matching research: a record from the Census can only match with a single record from the PES and vice versa.

QualityStage also has the option to reduce the comparison space by using blocking. Besides, it has comparison functions called approximate comparison algorithms. There are sixteen different functions but none of them are the comparison functions Jaro and Jaro-Winkler, which are methodologically appropriate for the 2010 PES.

In QualityStage, the matching results occur through grade cuts and use parameters similar to the Fellegi-Sunter Record-Linkage theory, but no software release shows this theory. They only explain the calculated parameters vaguely.

In general, QualityStage has several functionalities, and many of them are similar in practice producing the same results. The software is not easily accessed and understood by technicians who are not in the information technology area, what makes its use a difficult task. Also, you must have knowledge and access to the server database to get the program working efficiently. The comparison functions are not sufficiently clear either, with random scores sometimes obtained without a clear explanation.

RecLink II

This is a free software created by two Brazilian professors aiming to conduct record linkage of health data. RecLink II was developed in C++ language and works with input and output files in XBase standard (extension DBF).

The evaluation of RecLink II was based on the information reported in the literature and also on information available at its website. The terms used in the interface of the program indicate that this software was developed to assist users familiar with computing technical terms. In addition, RecLink II is not intuitive, so that non-experienced users need help from a tutorial.

RecLink II has the advantage of allowing implementation of any algorithm to make comparisons as long as the user feeds the program with an input file containing the necessary algorithms. On the other hand, RecLink II has the disadvantage of not allowing implementation of the EM-algorithm to estimate the M and U probabilities used to calculate the agreement and disagreement values.

Registry Plus™ Link Plus

Link Plus is a free software developed to perform probabilistic record linkage to support the National Program of Cancer Registries (NPCR) of the United States Center for Diseases Control (CDC). The evaluation of Link Plus was based on the examination of the User Guide version 2.0 (available through the “Help” tab or the F1 button) and on tests with simulated and real data.

Link Plus was used to detect duplicates in both Census and PES databases and to link records between the two databases. Although it has been originally developed to be used in cancer registries, the program can be used with any type of data with fixed width or delimited format, which allows its use in the Brazilian PES.

The Graphical User Interface of Link Plus is very friendly, showing all setting fields on one screen. Its output files show the matched records together, showing also the scores associated to them, which facilitates the verification of the quality of the matches.

The blocking mechanism of Link Plus has only one exact method and two phonetic methods (Soundex and NYSIIS – New York State Identification and Intelligence System) to make comparison blocks. The program has the advantage to implement the scheme “OR blocking”, which compares pairs formed in blocks based on at least one blocking variable. Most programs use the scheme “AND blocking”.

Since Link Plus was designed to be operated by non-advanced users, who are usually not familiar with the technical terms of matching theories, the field for choice of comparison methods presents illustrative terms instead of metric names. The available matching metrics for comparing alphabetical strings can be summarized in two methods: Jaro-Winkler distance (Last Name and First Name methods) and Levenshtein

distance (Generic String method).

The most basic comparison method implemented by Link Plus is the Exact, which returns all the agreement weights when the field agree or returns the full disagreement weight when the field does not agree. The Middle Name method only accounts for the occurrence of the middle name initial character versus the full middle name. The Value-Specific method associates lower weights to frequent values and higher weights to rare values in a database. SSN and Zip Code methods are designed to compare the Social Security Number and the Zip Code of the United States. This information is not collected by the Brazilian PES and its usefulness for data from other countries is not known. The Date method is used to verify if the dates compared are the same, checking also for omissions or transpositions.

Link Plus presents two options for obtaining the values of agreement and disagreement, which are used in its linkage process. The first option is related to the "Direct Method", which uses default or user-defined M-probabilities. The second option calculates the M-probability using the EM-algorithm based on the current data.

Link Plus is distinguished by the speed that it performs the entire process, but the program does not allow setting the parameters of the EM-algorithm. This may be one of the reasons why the values of M and U probabilities calculated by Link Plus do not coincide with those calculated by other programs that also implement the EM-algorithm.

Freely Extensible Biomedical Record Linkage - FEBRL

This is open source software originally developed to form probabilistic pairs in biomedical records. The assessment of FEBRL was based on the examination of the Graphical User Interface Manual - Release 0.4.1 and on tests with simulated and real data. The experience of Brazilian researchers and staff from the Australian Bureau of Statistics were also accounted for in the evaluation process.

This software was developed in Python by the Data Mining Group of the Australian National University, being able to perform deduplication (same database), linkage between two databases and data standardization.

The interest of the PES team was to assess the deduplication and linkage tools. The standardization tool was not studied because data collection was planned to allocate each variable in its respective field, which has greatly reduced the need for treatment of the data.

The deduplication and linkage tolls use the same method to compare pairs of records. The difference is basically that the deduplication tool does not compare pairs of records that are in the same position in the sequence, since it is exactly the same record in the same database. Moreover, the linkage tool compares all pairs of records within each comparison block.

The FEBRL Graphical User Interface uses a system of tabs to help the user to set rules and parameters to perform the record linkage. These tabs open only after confirmation of the data entered via the "Execute" button, which is tricky and makes the users to spend time to get used to it.

Through the "Index" tab, FEBRL can implement six different methods for holding the blocking, also allowing the comparison of records without perform the blocking. Through the "Compare" tab, FEBRL can implement twenty six metrics to perform the comparison of each field in the record. The "Classify" tab can implement six weight vector classification methods.

The main advantage of FEBRL is to implement many functions for blocking, comparing and indexing. Thus, the software meets the needs of the Brazilian PES based on its methodological aspects. In addition, FEBRL has three other advantages: the availability of a graphical interface (GUI), the option to import TXT and CSV file formats and the comparison of information from different fields.

The main disadvantage of FEBRL is that does not implement the EM Algorithm required for the calculation of the M and U parameters, which are used to determine the agreement and disagreement values for each field. Another disadvantage of FEBRL is the need for further processing of the output files for use in the Brazilian PES.

RELAIS – Record Linkage at Istat

Developed by the Italian National Statistics Institute, RELAIS is free software that has been examined and tested by the 2010 PES team. This section presents general information about the software, the Brazilian experience with the program and considerations on its usage.

The computer languages used by RELAIS are: Java, which is a language associated with objects, and R, a functional language for computing techniques associated with matching data process. Moreover, RELAIS has an architecture based on MySql environment, which allows the use of different databases.

The software matches data from two different databases, so there are several options available for that, such as: reduction of the comparison space, choice of matching methods, comparison functions, among others. There are also three matching methods: deterministic, deterministic rule-based and probabilistic based on the Fellegi and Sunter theory. In addition, there are seven different comparison functions.

The focus of the program is the matching, therefore no other processing functions, such as data standardization, cleansing, deduplication, merge, etc., are available. Any adjustment in the databases used in the matching process should be done through another tool. This requires creating new databases and making another upload RELAIS.

The matching results are displayed in tables and can be saved in text file format (.txt). These tables show all the variables common to the two databases, even if they were not used in the process of matching. Furthermore, it is possible to generate files with records that were not matched.

RELAIS 1.0 was used at the 2001 Italian PES. It was also presented at seminars and events at the Eurostat, at the Federal Committee on Statistical Methodology and at European and American institutes. It was also used by the Spanish National Statistics Institute to integrate two different databases: the examination of living conditions database and the central population register database.

In the Brazilian 2010 PES, RELAIS was tested with real data of two cities where the dress rehearsal of the 2010 Census and PES were conducted in April and July. The entire automatic matching for this exercise was done using RELAIS and its effectiveness was verified.

RELAIS is easy to use, provides good results for matching deterministic models and has a good interface to visualize the results. The software implements the Jaro-Winkler comparison functions and also optimizations for one-to-one match, following the premise of all the Post Enumeration Survey, i.e. a record from the Census refers to only one record from PES and vice versa.

The problem occurs when the probabilistic match is done. In this case, RELAIS presents rigid requirements related to the size and quality of the data file and to the quality of the comparison variables used in the matching process. SCANNAPIECO (2008) shows that the parameter estimation used in the probabilistic match is not reliable when the agreement weight equals to zero and disagreement weight equals to 1 for at least one of the comparison variables. Thus, RELAIS sends an error message, stops the match processing and doesn’t show the results.

Moreover, even when the estimation is performed, the match is processed and the results are presented, there are discrepancies among different data blocks. For example, consider a matching process with three comparison variables. Suppose that in a block there are two records that agree on all three variables and they have scores close to 1 (maximum value). In another block two records with the same characters may have another score, say close to 0.9. There may be cases where two records that agree on two out of three variables have score greater than two records that agree on the three variables.

RELAIS presents a clear methodology and has most of the comparison functions necessary for the Brazilian PES built in. Besides, the software allows estimation of the U and M probabilities using the EM algorithm and provides output that can be used in the subsequent stages of the process, such as assisted matching and reconciliation.

3. Some Remarks on the Assessment of the Packages

Most of the software considered here were developed for general purposes of record linkage, e.g. producing a comprehensive data base of clients or users of health services so that not only the requirements of the packages but also the profile of the users were taken into account. According to their scopes, some functionalities were more or less developed. RELAIS is an exception since it has been developed to be used in a PES project and, for this reason, it was the most friendly software for the Brazilian PES context.

The corporate software (SAS and IBM) provide more information on their own structure and results, whereas non corporate ones (Reclink, Likplus, Febrl and Relais) present a clearer view on the methodology. Thus, users can evaluate which one can meet better the methodological requirements of the project in course. Besides, there are also some differences on the interface, documentation and requirements among the corporate and open and free source software. The non-corporate software present friendly interface and allow the users to perform a record linkage without the need of extra technical assistance.

Most of the non-corporate software provides good documentation that can guide users not only in the use of the software but also on the choice of the functions and metrics to be used. Besides, all of them are available by download, allowing to try their suitability to the project in course.

Considering the Brazilian PES needs and the profile of the its team, a comparison between corporate and non-corporate software was performed, accounting for the following aspects:

- Documentation: assessment of the information available on both the use of the software and the mathematical models implemented. It was specially taken into account the clear explanation about the available functions for blocking, comparing and indexing.

- Interface: evaluation of how friendly and intuitive the users graphical interface is.

- Features: assessment of functions to perform complete matching considering a project such as a PES, which includes the need of deduplication and matching features.

- Requirements: accounting for the need of additional costs such as other software, training and technical assistance.

Fig 1 – Rating of software studied

Software	Documentation	Interface	Features	Requirements
Data Quality	☹	😐	😐	☹
Quality Stage	☺	☺	☺	☹
RecLink	☺	☺	☺	☺
LinkPlus	☺	☺	☺	☺
FEBRL	☺☺	☺☺	☺☺☺	☺
RELAIS	☺☺☺	☺☺☺	☺☺☺	☺

Positive: ☺ Negative: ☹ Not evaluated: 😐

Among the software studied, FEBRL and RELAIS deserve some highlights. FEBRL makes available a large number of functions for blocking, comparing and indexing. RELAIS allows performing a complete matching, which fits the Brazilian PES since implements the EM algorithm and provides fitted output for assisted matching and reconciliation. However, in the Brazilian PES project the matching is done by enumeration area as soon as the data collection is finished. This is due to the need of immediate reconciliation. Thus, it was necessary to develop a program that could be inserted in the production process

by running in batch. The program developed is presented in the next section.

4. R: a language and environment for statistical computing

Due to the need of a tool to perform the matching of the Brazilian Census and the Post Enumeration Survey records, IBGE developed a program in R language to find probabilistic record pairs.

R is a free and open source software for statistical computation and graphics, which consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files (see http://www.r-project.org/ for more information).

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 itself. It is possible for the user to use procedures written in C, C++, or FORTRAN languages to improve efficiency. The R distribution contains functionality for a large number of statistical procedures, and additional modules ("add-on packages") are available for a variety of specific purposes (see R Add-On Packages at http://www.r-project.org/).

A specific library called RecordLinkage provides functions and data structures that make evaluation of record linkage methods easier, facilitating the application of record linkage to different data sets. For the use in the Brazilian PES it was necessary to implement stochastic methods based on the framework of Fellegi and Sunter for record linkage.

Assisted by the RecordLinkage library, many methods were tested before the final model was found. The method EM that is applied to the Brazilian PES is provided by the package RecordLinkage. The library allows also the implementation of blocking mechanism as well as the use of matching metrics for comparing alphabetical strings, such as Jaro-Winkler.

Due to a requirement of the Brazilian PES, the automatic processing had to be implemented through a script identified as final model. This final model is based on the following steps:

- Implementation of a function to select the variables that will be used in the model.
- Use of the compare.linkage function from the library RecorLinkage to make all possible pair combinations.
- Use of the similarity function JaroWinkler provided by the RecorLinkage library for each variable.
- Use of the function emWeights provided by RecordLinkage library to obtain weights of agreement for the pairs through the EM algorithm.
- Use of solve_LSAP function from library clue to perform a one-to-one association. The procedure solves linear sum assignment problem originating the pairs.
- Implementation of a function to assess the quality of the pairs and to exclude possible false-positive pairs. The pairs left out are recorded in a text file.

The R Record Linkage package has a vast documentation with reference to the methods and provides the datasets used in the examples. R supplies different functions for data manipulation, cleaning and transforming, making the reading and writing of different data formats possible. Compared to other programs, R system does not have a good interface due to the fact that it is an environment of programming. Besides, to make a good use of R, it is necessary to know about the methodological aspects of the process and to have programming skills. On the other hand, it is possible to make an automatic processing in R by using looping functions and changes can be made to improve the functions or to fix errors. As R is an open source software, it allows to access the script of some functions and modify them. The system was highly successful and fulfilled all the PES needs.

REFERENCES

ALUR, N., et al. IBM WebSphere QualityStage Methodologies, Standardization, and Matching. Disponível em: http://www.redbooks.ibm.com/redbooks/pdfs/sg247546.pdf. Último acesso em: 06/12/2010. Armonk, 2008.

ANDREAS Borg and Murat Sariyar, (2010). RecordLinkage: Record Linkage in R. R package version 0.2-2. http://CRAN.R-project.org/package=RecordLinkage.

CAMARGO Jr, KR & Coeli, CM: Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad. Saúde Pública, Rio de Janeiro, 16(2):439-447, abr-jun, 2000. Available at: <http://www.scielo.br/pdf/csp/v16n2/2093.pdf>.

CIBELLA, N., et al. Sharing Solutions for Record Linkage: the Relais Software and the Italian and Spanish Experiences. Disponível em:http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S7P2_SHARING_SOLUTIONS_FOR_RECORD_LINKAGE_CIBELLA_TUOTO_.pdf. Último acesso em: 18/12/2009. Bruxellas, 2009.

CHRISTEN, P. Febrl – Freely Extensible Biomedical Record Linkage. Release 0.4.1. Canberra, Australia. 2008.

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.

INTERNATIONAL BUSSINESS MACHINES. InfoSphere QualityStage for z/OS. Disponível em: http://www-142.ibm.com/software/products/br/pt/ibminfoqualforzos/ . Último acesso em: 06/12/2010.

LINK PLUS User's Guide. Version 2.0. Center for Diseases Control. Atlanta. June, 29th 2007.

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.

RECLINK Dowloadable Files, Publications Related to RecLink II and Educational Material. Available at: <http://www.iesc.ufrj.br/reclink/>.

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

SCANNAPIECO, M., et al. Relais user's guide 2.0. Disponível em: http://www.istat.it/strumenti/metodi/software/MTSFload/ALTRIload/RELAISload/manual_relais_2_0.pdf. Último acesso em: 21/12/2009. Roma, 2008.

TUOTO, T., et al. RELAIS: Don't Get Lost in a Record Linkage Project. Disponível em: http://www.fcsm.gov/07papers/Tuoto.VI-C.pdf. Último acesso em: 21/12/2009. Arlington, 2007.