Use of Record Linkage to build a Brazilian List Frame for Agricultural Statistics

Andrea, Diniz da Silva
National School of Statistical Science – ENCE/IBGE. Rio de Janeiro, Brazil
andrea.diniz@yahoo.com

Flavio, Pinto Bolliger
The Food and Agriculture Organization of the United Nations – FAO. Rome, Italy
Flavio.bolliger@fao.org

Jose Andre, de Moura Brito
National School of Statistical Science – ENCE/IBGE. Rio de Janeiro, Brazil
jambrito@gmail.com

Seventh International Conference on Agricultural Statistics
ICASVII
2016 Rome, 26 | 28 October

ABSTRACT

An important part of the Brazilian agricultural statistics is produced based on census and subjective surveys. Despite the number of good quality statistics produced, having the official agricultural statistics relying only on decennial census, which gets obsolete quickly, and non-probabilistic surveys, which does not allow for associating any margin of error, is critical. To move away from the present model, the national statistical office is working on implementing an integrated agricultural survey system, which is a probabilistic sampling based system. A key component of such system - a master sampling frame – is still missing. A number of available quality source allow for building a list frame using data integration technics. The paper presents a proposal of using record linkage methods for bringing together data from survey and register in order to obtain a list frame for agricultural surveys.

INTRODUCTION

An important part of the Brazilian agricultural statistics is produced by the National Institute of Geography and Statistics - IBGE. Since 1920, a census of agriculture has been conducted and since 1938, agricultural statistics have been produced regularly from subjective surveys (IBGE, 2002, p.7). These two sources are on the pillar of official agricultural statistics. Together, they account for a number of quality information, as frequent as monthly, for national and sub-national levels.

However, having the official agricultural statistics relying only on decennial census, which gets obsolete quickly, and non-probabilistic surveys, which does not allow for associating any margin of error, is critical. Such state of art imposes limitations to timeliness and quality of statistics and prevents from meeting accordingly the current and emerging demands in order to support sounding decision at local level and ensuring both, comparability at international level and compliance with international standards.

Looking for best practices for improving agricultural statistics, IBGE works in the meaning of meeting Global Strategy - GS recommendations. Under such framework, is working on implementing an integrated agricultural survey system: National System of Agricultural Surveys - SNPA, which is a probabilistic sampling based system (Proposta, 2011). When implemented, the system will produce information on rural activities and production, every year and quarter respectively.

At present, the project still misses a key component: a master sampling frame – MSF. To run SNPA considerable effort on finding a suitable and feasible way of building a MSF is highly desirable. Considering already existing frames as starting point will allow for using multiple frame approach, improving sample efficiency, making surveys more cost-effective.

Nowadays IBGE holds four important frames: 1) area frame built as part of last population census (2010); 2) list of establishments from last census of agriculture (2006); 3) list of establishments from last population census (2010); and 4) list of establishments in business register. In addition, lists of establishments and rural producers are available from State Tax Administration and from Ministry of agrarian Development.

Despite good quality of available area frame the same consideration does not apply to list frames. There is no one list that can be considered complete and up to date. Limitation on coverage and quality highlights the need of using them in a complementary way to obtain a single improved list frame.

The combination of available lists requires intensive use of data integration techniques. A number of methods to solve record linkage problems have been developed, but none of them can be considered the most appropriated regardless the structure and quality of data. To find out what method gives the best fit to de available data is the challenge approached in this paper.

BUILDING A LIST FRAME

Four years ago, Bolliger and colleagues made a first exercise aiming to build a list frame, applying record linkage methods. Linking 2006 Census of Agriculture, National Directory of Addresses for Statistical Purposes, Central Register of Enterprises and Annual Social Information Report, was attempted by applying sequential procedures such as standardization, deterministic matching and probabilistic matching. “Although advanced linkage techniques have been judiciously applied, the results achieved were quite frustrating due to problems of compatibility and quality of information found in different registers.” (Bolliger, 2012, p. 6)

This is an extension of the work done by Bolliger and colleagues. In addition to inclusion of two sources – tax administration and Ministry of Agrarian Development - refinements on record linkage method are improvements towards the previous experience.

Data Sources

Five data sources will be linked:

- 2006 Census of Agriculture
- 2010 Population Census (addresses file)
- Business Register
- Register of family farmers (Ministry of Agrarian Development)
- Taxpayers on State Tax Administration

The 2006 census of agriculture is the most comprehensive one and holds the best coverage (5.2 million establishments and holders). However, important addition can be obtained linking data from the 2010 population census (2.6 million); the business register (99 thousand); register of family farmers (2.6 million contracts) and register of taxpayers. In total, we expect to deal with about 15 million records to be linked.

For data available at first moment, regardless potential information of each source, name and address are the common information on all of them. Address is not directly available only for family farming but can be retrieved by linking its records to the ones in State Tax Files. In such way, names and addresses will be at the base of record linkage system, as follows.

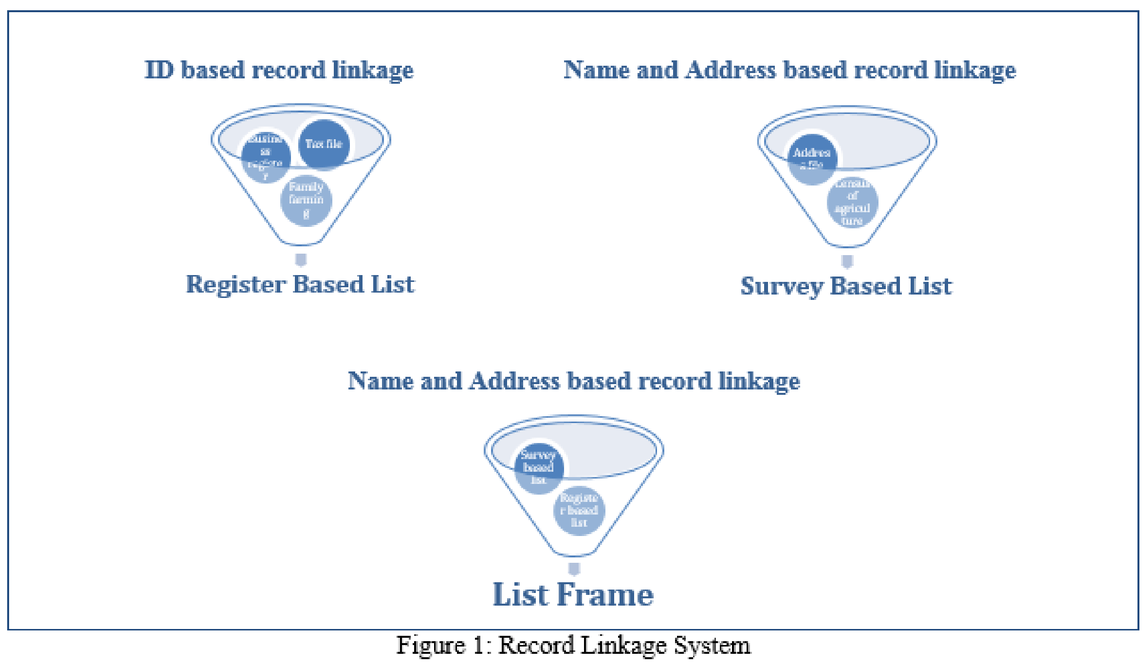

Figure 1: Record Linkage System

Two classes of theoretical model is under consideration: one stochastic and one non-stochastic. In the first class method follows concept introduced by Newcombe (1959) and formalized in the mathematical model of Fellegi and Sunter (1969). The second class comprises machine learning methods, such as decision trees.

In both cases, empirical model will consider municipalities as natural blocks. Within the blocks, Jaro-Winkler comparator shall be used to compare names and addresses.

Empirical evaluation on real data, applying proper procedures for the record linkage task under each of the methods, will be performed. This evaluation will provide a deeper insight into empirical similarities and differences between modelling frames of the record linkage problem.

Record linkage tasks were implemented in R programing language. Preprocessing rely on basic functions of packages stringr and plyr and the record linkage tasks count on package Record Linkage.

DISCUSSION

Working with regular government sources increase possibility of ensuring maintenance of resulting list frame. In addition to continuity of data production, such sources offer data at relative low cost. However, at first moment, when a culture of cooperation is still in its way, many challenges are present. Despite an expected easy cooperation among government sectors, data acquisition is still an issue. A complete run of the system was not possible yet due to still unavailability of taxpayer data from most of the States. Besides, an interruption of negotiation with Ministry of Agrarian Development, due to sensitive political situation, obliged to work with much less informative data such as the ones available at internet, in addition to increasing of effort to obtain data.

Different quality, structure and kind of available information among sources lead to an extensive work to clean and standardize data for running record linkage tasks. Such preliminary job were expected, however working with five different sources and five states resulted in a number of preprocessing steps greater than expected.

A number of method for record linkage are now available. Both stochastic and non-stochastic ones have shown advantages in specific experiment. However, none of them can be considered superior regardless the data used. Despite proved validity of Fellegi-Sunter for linking data based on name and address, limitation remain when estimating parameters; otherwise, machine learning methods lead to better result when a lot of training data is available.

Despite providing had hoc model, present exercise shall allow for an extension to update data in subsequent moments of time, when using same data sources. In addition, it represents and a starting point to add other sources when available. It is important to consider that exercise made cover northwest part of the country, whose data usually have lower quality when compared to other region´s data. In that way, it is expected to face less problem when extending the system to other regions.

Discussion presented here refers to an on-going work, so general performance of proposed record linkage system is not yet available. Expected results should present performance under the general results obtained in the 2010 Census Post Enumeration Survey, once dealing with rural addresses is harder than urban ones. However, optimistically, such performance should overcome the ones obtained by the experience reported by Bolliger and colleagues.