LINKING CENSUS OF AGRICULTURE AND BUSINESS REGISTER IN BRAZIL: SOME DEVELOPMENTS AND MANY CHALENGES

Andrea Diniz da Silva[1], Raphael Molina Guimaraes[2], Geremias de Mattos Fontes Neto[2]

Abstract: To improve agricultural statistics production the Brazilian Institute of Geography and Statistics - IBGE considers the integration of data from censuses and administrative registers. However, as not all records from such sources have a common unique identifier, data integration depends on a record linkage method that helps to identify which records belong to the same establishment using only quasi-identifiers. A complete record linkage exercise, using data from the census of agriculture and the business register, was performed. Here we present the final design of a three-step method for linking the census of agriculture to administrative registers, as well as some challenges to perform record linkage on agricultural data in Brazil. We also provide an illustration of the problems caused by different accuracy standards in the registration of information, by the use of different concepts and definitions, and also by the different levels and patterns of coverage of the sources, which, ultimately, introduce a level of error that can prevent the identification of many records belonging to the same establishment, even if they exist.

keywords: Agricultural statistics, data integration, record linkage

INTRODUCTION

To improve agricultural statistics production, as a way of meeting growing national and international demands, the Brazilian Institute of Geography and Statistics - IBGE considers the integration of data from censuses and administrative registers such as business register and other sources under the public administration. However, as not all records from such sources have a common unique identifier, data integration depends on the development of a record linkage method that helps to identify which records belong to the same establishment, using quasi-identifiers such as name and address.

Three experiments were carried out between August 2018 and December 2020, using data from the 2017 Census of Agriculture, State department of agriculture of the states of Pará and Tocantins, and the Business Register - CEMPRE / IBGE. In addition to a complete record linkage exercise, the experiments include 3 studies: one concerning the optimal tolerance limit for the difference between the coordinates when these are used as a linking variable (Silva et al., 2019); a second referring to the generation of training data for using supervised methods for record linkage; and a third one involving the comparison of two variables of the name of the agricultural establishment.

Here we present the final design of a three-step method for linking the census of agriculture to administrative registers as well as some challenges to perform record linkage on agricultural data in Brazil. The complete report as well as the R codes used are available on demand.

A THREE-STEP METHOD

The record linkage method used can be described as a three-step method: preparing, comparing and classifying record pairs as link or non-link, as described by Chambers and Silva (2019). After preparing the data, three rounds, adopting different strategies of comparison and classification, were performed. This approach allowed a better use of the predictive power of all the variables available in the databases used. Round 1 has a hybrid approach, combining exact comparison plus deterministic classification and comparison using similarity (Winkler, 1990) plus probabilistic classification (Fellegi and Sunter, 1969; Dempster1977); round 2 is entirely deterministic with a tolerance limit, with the use of algorithms for calculating geodesics on an ellipsoid WGS84 (Karney, 2013); and round 3 is non-deterministic with a comparison using similarity plus classification using classification tree algorithm (Breiman, 1994).

[1] National School of Statistical Sciences – ENCE/IBGE.
[2] Brazilian Institute of Geography and Statistics - IBGE.

For the first round, the number in the tax administration, name and address of the establishment were used as linking variables; for the second it was the geographic coordinates; and, in the third, the name and the address of the establishments were the linking variables.

In the proposed method, the rounds are performed sequentially, and the records linked in one round are not submitted to the next one. Precisely, records linked in round 1 are not submitted to round 2 and those linked in round 1 or 2 are not submitted to round 3. Besides, in each round only records that have information for the linking variable are considered, i.e., record with missing values in the linking variables are excluded.

LINKING CENSUS AND BUSINESS REGISTER

The experiment included the adoption of rounds 1 and 3, as described in the previous session. Round 2 was skipped as no geographic coordinates were available in the Business Register. To perform those rounds, ten linking variables were used: the number in the tax administration; first to fourth part of the establishments' name; first to fourth part of the street name; and the gate number.

The data sources were 2017 Census of Agriculture and Business Register, from now on called just Census and Cempre database. Census database consists of 5,073,316 agricultural establishments, distributed in 5,284 municipalities, in the 27 Federation Units while Cempre database includes 131,921 registered agricultural establishments, distributed in 4,304 municipalities, in the 27 Federation Units.

The complete process, after the completion of the two rounds, would allow the updating of 18,229 records from the Census database and the incorporation of 113,692 new records of establishments coming from the Cempre database (Table 1).

Table 1: Summary statistics of linking Census and Cempre

Source	Records
Census database	**5.073.324**
Cempre database	**131.921**
Updating in Census	18.229
Round 1	1.785
Round 2	-
Round 3	16.444
To be included in the Census base	113.692
Updated Census database	**5.187.016**

Data sources: Brazilian Institute of Geography and Statistics – IBGE, 2017 Census of Agriculture and 2019 Business Register (CEMPRE).

Remarks

The Cempre base is made up of formally constituted companies, so it represents a subset of existing agricultural establishments in Brazil. To be more accurate, most of the agricultural establishments registered in the Census are not formalized. This limitation does not invalidate the use of such a source for updating the Census base, although its contribution may be modest. However, of the approximately 130 thousand companies registered with Cempre, only about 8,400 could be associated with an establishment in the Census when the number in the tax administration was solely used. In other words, just over 6% of such numbers were found in the Census. This situation is further aggravated when the name and address are added as a condition. Under this restriction, less than 2% of Cempre's records are found in the Census.

The results are quite inconsistent with those expected, as there is an understanding that Cempre includes at least all large companies, which would also have been included in the Census. Large companies are usually less prone to registration and coverage errors.

CHALLENGES

The proposed record linkage method can be adopted for several purposes. In the present work, the focus was on its use to integrate data on agriculture, from business register. The task is quite challenging because, despite the fact that there are already numerous methods that are proven to be efficient in linking data with some level of inaccuracy, in the case studied the linking faces problems caused by **different accuracy standards** in the registration of information, by the use of **different concepts and definitions**, and also by the **different levels and patterns of coverage** of the sources, which, ultimately, introduces a level of error that can prevent the identification of many records belonging to the same establishment, even if they exist.

Important information for linking records is usually declaratory, when obtained through surveys, while in administrative registers it is documentary. This results in different standards in the recording of information, since the respondent often does not know the "real" name of the establishment or even in whose name the business is registered in the government agencies. This type of problem is not restricted to the name of the establishment, but it also affects the address and number in tax administration, especially the last one. While the address of the establishment is registered in surveys as it is known by the respondent, this usually needs to be verified by means of documents (deed, electricity bill, rental contract, etc.) when informing public agencies. The result is that there are "different" addresses for the same establishment depending on the source of the data. The same occurs with the tax administration number, which is registered as informed, often without at least going through a consistency check, allowing the registration of sequences of numbers manufactured on the spot. In addition, it is not uncommon to register a number in tax administration belonging to a different person or company, with some or even no relation to the agricultural establishment. The three aforementioned information, name, address and number in tax administration, are the main identifiers of the agricultural establishment, widely used in record linkage. Therefore, important errors in these variables can make it impossible to identify the real link. In addition, depending on the level of error, verification cannot be done even by clerical review, requiring the return to the field to verify the information.

Since the present work involves dealing with different data sources, the need for compatibility and harmonization of concepts and definitions is not surprising. However, working with linking variables or even the main unit with compatible concepts instead of the same ones, adds another source of linking error. When conceptual differences are in the definition of the main unit of work, in this case the agricultural establishment, the implication is that different units such as, establishment, producer, company will be compared. This increases the risk of false positive in addition to affecting the method efficiency. When conceptual differences are in the linking variables, they can, in the end, weaken a variable that in theory would be an excellent predictor. Silva et al. (2019) studied the distance between coordinates, searching for a limit below which the two coordinates would be considered as belonging to the same establishment, but the optimal tolerance limit found was too high. The conclusion was that this high limit was too excessive to be attributed to the variation in accuracy of the equipment used for capturing the coordinates. These differences would be better explained by the conceptual differences adopted for capturing the coordinates. Thus, geographic coordinates, which can have an excellent predictive power for the identification of agricultural establishments in two databases, cannot be adopted as a link variable when there are important differences on concept or instructions such as getting the coordinates from the gate, or from the building, or even from the fields.

The Census is the only operation fully inclusive, that is, it is meant to enumerate all units dealing with agricultural production of all sizes in all geographic levels. The authors of this work do not know any other source that is not composed only by selected units. Depending on the level of coverage of the source to be integrated with the census, the "needle in a haystack" effect can be experienced, that is, having the challenge of finding a few links in a large number of comparisons. This affects not only the efficiency of the method, but also increases the risk of false positives. If the coverage pattern depends on the type of activity or other characteristics of the establishment, for example, greater coverage for soybean than corn producers, there will be biases in estimating quantities from linked data. That is the case when linking census to business register. When linking the Brazilian Census and Cempre, this partial coverage meant searching for the approximately 132 thousand establishments present in the register among the more than 5 million establishments present in the census. Certainly, a strategy to improve linking efficiency is to build blocks taking into account characteristics of the units present in the registry. Such variables have to be used in addition to the municipality of activity, adopted to define

the blocks in the mentioned exercise. However, the choice of new blocking variables needs to take into account the response level, which should ideally be close to 100% for the blocking variables. Reflecting on the quality of the information regarding production, whether on the basis of the census or records, is fundamental for this type of variable to be used in blocking.

The use of the proposed record linkage method to update Census database using administrative registers requires caution and, when possible, additional actions to reduce or to mitigate effect of errors in the linked data. Active search for new establishments, telephone contact, and visiting the establishment's operating location to update information regarding the establishments during the inter-census period are some of them. The complementary work allows: 1) to add data on establishments already registered in updated Census database; and 2) to identify establishments that have ceased to operate or have been included improperly in the original databases. Whether it will be applied to all establishments or to a sample depends on the available resources. Additionally, a sample of establishments can be used in order to provide inputs to assess the quality of the data in the updated Census database. As such activities are carried out only for known establishments, therefore already included in the integrated base, they do not result in the inclusion of new establishments. The inclusion of new establishments must be carried out by active search or when updating the database using administrative records, as mentioned above.

REFERENCES

Breiman, L. (1994). Bagging Predictors. Technical Report No. 421. Department of Statistics University of California.

Chambers, Ray & Silva, Andrea. (2019). Improved secondary analysis of linked data: a framework and an illustration. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Dempster, A. P.; Laird, N. M.; Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B.

Fellegi, I.; Sunter, A. (1969). A Theory for Record Linkage, Journal of the American Statistical Association, Vol. 64, 1183-1210.

Karney, Charles F. F. (2013). Algorithms for geodesics. Journal of Geodesy. Volume 87, Issue 1, pp 43–55.

Silva, Andrea Diniz; Guimaraes, Raphael Molina; Fontes Neto, Geremias Mattos. (2019). Adjusting Tolerance Limits for Difference Between Coordinates Used to Link Agricultural Producers' Records. Proceedings of the 62nd ISI World Statistics Congress, Kuala Lumpur.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. U.S. Bureau of the Census, Stat. Research Div.