Sâmela Batista Arantes

Disclosure risk assessment for frequency tables of a Brazilian economic survey

Sâmela Batista Arantes [1]; Maysa Sacramento de Magalhães [2]; José André de Moura Brito[3]

[1] Brazilian Institute of Geography and Statistics, samela.arantes@ibge.gov.br
[2] Brazilian Institute of Geography and Statistics, maysa.magalhaes@ibge.gov.br
[3] Brazilian Institute of Geography and Statistics, jose.m.brito@ibge.gov.br

Abstract: In the dissemination of survey results, the protection of confidential information is an important step that guarantees the credibility of the statistics offices and, consequently, the quality of the published results. Tables are the main form of dissemination of economic surveys. In this sense, the objective of this work is to compare methods of secondary suppression for frequency tables from the Innovation Survey (IS) of Brazilian Institute of Geography and Statistics (IBGE). The results show that different methods and criteria for assessing disclosure risk must be considered in the construction of an approach to statistical disclosure control for tables.

Keywords: statistical disclosure control; risk assessment; suppression; frequency tables; economic survey

1. Introduction:

Tables are the most common form of dissemination of surveys results carry out by National Statistics Offices (NSOs), mainly in economic surveys. In this case, the application of confidentiality protection methods aims to avoid disclosing information in cells or information that may be obtained from the relationships between cells in the same table or between tables. Hereupon, this work presents some results related to the disclosure risk assessment for frequency tables of the Brazil's Innovation Survey 2012-2014 (IBGE, 2016) considering cell suppression as a method of protection. The methods presented here considers the available literature on the topic, international guides and international practices adopted by other NSOs. In section 2, the data source is briefly described, whereas the disclosure risk assessment with results obtained for two frequency tables from the Innovation Survey (IBGE, 2016) is presented in section 3. The final remarks are provided in section 4.

2. Data source

The Innovation Survey is one of the IBGE's economic surveys and deals with innovation practices in Brazilian companies which are selected by sampling. The responding units are the companies with 10 people or more whose main economic activity belongs to industrial, electricity, and gas sectors, and some selected services. The survey is conducted

every three years and the latest dissemination refers to the triennium period 2012-2014. The selected sample is probabilistic and stratified by company size and by main economic activity. The two tables used to exemplify the SDC approach proposed in this work are two-dimensional hierarchical tables from this Innovation Survey. The row's variable referring economic activities and the columns refer to a variable that represents the degree of novelty of the innovation adopted, with each table having 621 cells. The first table refers to product innovations and the second refers to process innovations. The two tables will not be detailed, as they represent only examples of applications.

3. Disclosure risk assessment

The application of disclosure risk assessment methods consists of two stages, namely: the primary suppression and the secondary suppression.

3.1. Primary supression

The risk assessment in tables consists first of an analysis of the disclosure risk of confidential information applied to each cell separately. In the case of frequency tables, the disclosure risk is related to the concept of rareness or uniqueness of the units in relation to the counts of the categories of the grouping variables, that is, the lower the frequency presented in the cell, the greater the risk of disclosure. For certain types of information, rareness may encourage respondents or not to find out the identity or attributes of other survey respondents (HUNDEPOOL *et al.*, 2012). In this sense, the rule commonly used in frequency tables is the minimum frequency rule, in which cells with frequencies below a certain threshold are suppressed.

Regarding the minimum number of respondents required in each cell so that it is not considered sensitive, it varies according to several institutions and guides. In European Statistics System in Field of Statistical Disclosure Control - ESSNetSDC (Brandt *et al.*, 2009), for example, there is an agreement between NSOs that the number of unweighted respondents (f) in each cell must be at least three for frequency tables with information from economic or demographic surveys. For the Europa Business Statistics Manual (Eurostat, 2019), in the publications that aggregate European countries, there is an agreement of a minimum value of five respondents in each cell. The UK Office for National Statistics (ONS) uses the value f = 10 (Griffiths *et al.*, 2019). Considering that the minimum values recommended in the literature are between three and ten unweighted respondents, the minimum number of respondents considered in this work varies between three and ten as presented in Section 3.3.

3.2. Secondary suppression

Analyzing the tables individually considering only the disclosure risk in each cell is insufficient to protect the sensitive information presented in the table, because the values of some cells can be recalculated or estimated using the totals available in the same table. Considering that cell suppression is the most common method of statistical disclosure control to protect tables, the methods of secondary suppression are used to solve the so-called secondary cell suppression or secondary suppression problem.

The problem of secondary suppression involves finding a set of additional cells to be supressed for the protection of sensitive cells such that the number of suppressed cells is minimum.

Secondary suppression problem: Let C, S', and $\overline{S'}$ donate a set of cells in a table T, a set of sensitive cells (primary suppressed cells), and the cells that were initially not suppressed or not classified as sensitive, respectively, where $C = S' U \overline{S'}$. Thus, in the case of cell suppression as method of protection, it is necessary to determine, from the set $\overline{S'}$, which cells should be suppressed to minimize the disclosure risk of sensitive cells S'. At the same time, the loss of information must be minimized, which consists of minimizing the total number of additional cells to be suppressed. The set of cells that should be suppressed from $\overline{S'}$ is denoted by S'' (adapted from Minami and Abe, 2019).

In Fischetti and Salazar-González's (2001) proposal, the model for the problem of the secondary suppression is defined by an objective function that allows measuring the loss of information expressed by the total number of suppressed cells weighted by the cost of each cell.

In this model an exact algorithm is used, more specifically, the branch-and-cut algorithm. The objective function is minimized under the table's additive restrictions. A disadvantage of applying this model is that the convergence of the algorithm is not guaranteed or can take a very long time which depends on the size of the table. Thus, for large tables (number of cells), there is no guarantee that this algorithm will present a solution in a reasonable time. Therefore, in addition to this model, the HITAS algorithm (Heuristic approach to cell suppression in hierarchical tables) and the Hypercube algorithm are also used in this work. Both are heuristic algorithms. The objective function is the same and the goal is to minimize the number of additional cell suppressions.

The HITAS algorithm is heuristic proposed by De Wolf (2002) for hierarchical tables. In this approach, the table is divided into non-hierarchical subtables within the original table and the algorithm used by Fischetti and Salazar-González (2001) is applied to each subtable. In this way, a suboptimal solution is obtained in a shorter execution time. These two approaches use integer programming.

The Hypercube algorithm was proposed by Repsilber (1994) and produces solutions in a shorter computational time than the previous algorithms. According to Giessing (2013), the sensitive cell is considered sufficiently protected if it is contained in a hypercube where all other cells (vertices belonging to the hypercube) are suppressed. For each hypercube formed, the loss of information associated with the suppression of its vertices (cells) is calculated. The hypercube that leads to the minimum loss of information, according to the objective function, is selected and all its vertices are suppressed cells. The algorithm subdivides D-dimensional tables with hierarchical structure into a set of subtables without substructure. These subtables are protected successively in an iterative procedure that starts from the highest level, i.e., the most aggregated level. In comparison to the first two algorithms, this algorithm has a low computational cost, but it tends to suppress more cells to obtain safe tables.

3.3. Results of disclosure risk assessment for two tables

The methods were performed in the R software, more specifically with sdcHierarchies package, which implemented the defined hierarchies of the hierarchical variables, and sdcTable package (Meindl, 2019) that contains the three algorithms cited, to obtain the number of suppressed cells in each case. Figures 1 and 2 show the results of the primary and secondary suppressions for two-dimensional frequency tables containing 621 cells each (69 rows and 9 columns). The variables that make up the analyzed tables (economic activity - table rows, degree of innovation of the activity - columns) are hierarchical. In Figures 1 and 2, the red line corresponds to the percentage of sensitive cells obtained in the primary suppression, the other lines correspond to the percentage of cells suppressed in the secondary suppression after the application of the algorithms.

In Figure 1, the red line represents the percentage of sensitive cells in the first table for each value f, where f represents the minimum number of respondents required for the cell not to be classified as sensitive. Analyzing the red line graph, one can observed that: When f=3, more than 15% of the cells contain less than 3 respondents which means more than 15% of the cells do not meet the minimum number required, which is 3, for instance. When f = 10, the percentage of sensitive cells exceeds 40%. In Figure 2, the red line represents the percentage of sensitive cells in the second table. In this case, when the minimum number of respondents required is f=3, the number of sensitive cells exceeds 20% and when f = 10 the percentage of sensitive cells also exceeds 40%.

Regarding the secondary suppression, the Hypercube algorithm presented the highest percentage of suppressed cells, and for the first three values of f, i.e., 3,4, and 5 this method tends to be more distinct from the others and this represents an overprotection. In both tables, the Hypercube method indicated between 40% and 60% of suppressed cells, approximately. The Fischetti and Salazar-González model and the heuristic algorithm HITAS showed close

performance in the number of suppressed cells, particularly when f is 8, 9 or 10. The Fischetti and Salazar-González's model presented higher or equal percentage of suppressed cells than the HITAS heuristic algorithm in all cases.

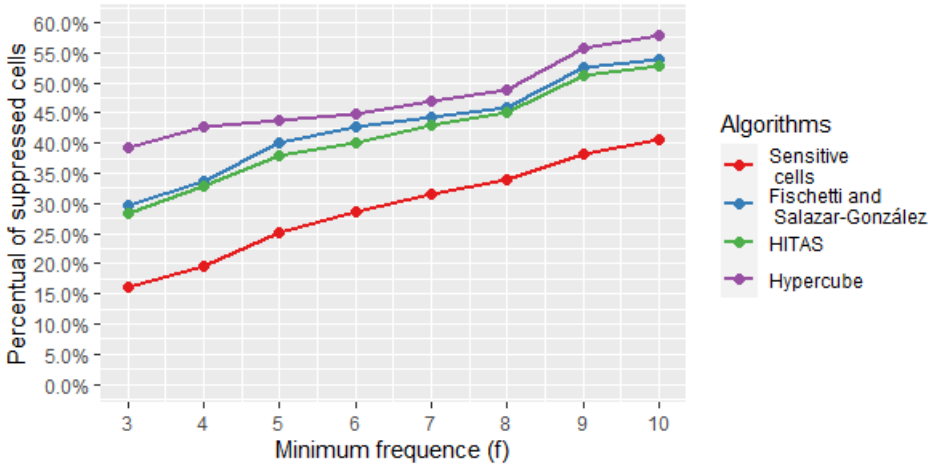

Figure 1. Percentual of suppressed cells by algorithms for the first table

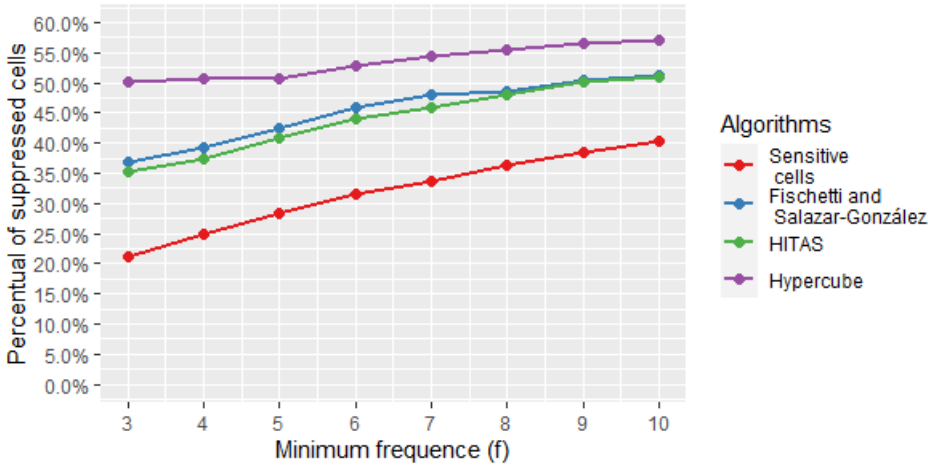

Figure 2. Percentual of suppressed cells by algorithms for the second table

4. Final remarks

Tables are the main product of economic surveys. In this regard, it is important to study and compare risk assessment methods for confidential information to preserve the confidentiality of respondents and the credibility of the institutes. Future studies will include more details regarding the performance of different methods applied to different types of tables from Brazilian economic surveys.

References:

1. Brandt, M; Franconi, L; Guerke, C; Hundepool, A; Lucarelli, M; Mol, J; Ritchie, F; Seri, G; Welpton, R. Guidelines for the checking of output based on microdata research. European Statistics System in field of Statistical Disclosure Control ESSNet SDC, 2009.
2. De Wolf, P. P. Hitas: A heuristic approach to cell suppression in hierarchical tables. In Inference Control in Statistical Databases, FromTheory to Practice, pages 74–82, London, UK, Springer-Verlag, 2002.
3. Eurostat. Europa Business Statistics Manual - Statistical Disclosure Control, Eurostat Statistics Explained, 2019.
4. Fischetti, M.; Salazar-González, J. |J. Solving the Cell Suppression Problem on Tabular Data with Linear Constraints. Management Science, Vol. 47, p. 1008-1027, 2001.
5. Giessing, S. Software tools for assessing disclosure risk and producing lower risk tabular data. Data Without Boundaries Deliverable 11.1 – Part B, 2013.
6. Griffiths, E.; Greci, C.; Kotrotsios, Y; Parker, S.; Scott, J.; Welpton, R; Woods, A. Handbook on Statistical Disclosure Control for Outputs, UK, 2019.
7. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S.Spicer, K., et al. Statistical disclosure control, 2012. Wiley Series in Survey Methodology: Wiley. ISBN 9781118348222.
8. IBGE. Pesquisa Industrial de Inovação Tecnológica 2014 – PINTEC, 2016.
9. Meindl, B. Package ‘sdcTable’, 2019.
10. Minami, K.; Abe, Y. Algorithmic Matching Attacks on Optimally Suppressed Tabular Data. Algorithms Open Access Journal Volume 12, Issue 8, 2019.
11. Repsilber, R. D. Preservation of Confidentiality in Aggregated Data. Paper presented at the Second International Seminar on Statistical Confidentiality, Luxemburg, 1994.