Paper Title Outlier detection for compositional data by municipality and disabled people in Brazil

Paulo Tadeu Meira e Silva de Oliveira[1,2]

[1] EESC-USP. São Paulo, Brazil

Abstract:

Outliers are observations with a unique combination of characteristics identifiable as being different from the others. Disabled people are considered to be those who have physical, hearing, intellectual or sensory disability, factors which, in interactions with various barriers, can obstruct their full and effective participation in society like other people. According to the IBGE in the 2010 Demographic Census, there were 45.6 million disabilities people in Brazil distributed in different municipalities. Data were considered by municipality justified by the fact that the level of service provided by these people varies according to the infrastructure and availability of existing resources in the most diverse locations. The data sets of the 2010 Demographic Census, aggregated by municipality on topics related to disabled people, identification, education, family, work and income, housing occupation, housing conditions, housing basic improvements, other goods, life quality and all for each of the 645 municipalities of the São Paulo state. Compositional data are those that establish the relative information, they are parts of a whole, the sum of these data in each line is a constant, in such a way that it represents 100%. In this work, for compositional data with logarithmic transformation were applied in conjunction with the variable selection procedure in which the variables for each model then, they were applied for detection method Distance from: Mahalanobis, Euclidean, standardized Euclidean, and, score of the first two main components using the corresponding data from the municipalities of the State of São Paulo, and finally, a comparative study will be carried out between the results obtained by these different methods and topics. The objective was the comparative study between these methods for detecting outliers. In this work it was verified that the outliers are more concentrated in work and income between the topics and Pearson's distance between the methods

Keywords:

disability people; compositional data; regression analysis; outlier's data

1. Introduction:

Introduction In any data set it is important, first of all, to make a very careful analysis of all the data, as there may be cases that do not match the distribution of the rest of the data. These are atypical points, known as outliers ((Giroldo, 2008).

The interest in the detection of atypical data points by the statistical community has increased since the middle of the last century. Identifying outliers is an important step prior to multivariate analyzes which are sensitive to them.

In univariate and bivariate data sets, outliers can be detected analyzing the scatter plot. Observations distant from the cloud formed by the data set are considered unusual. In multivariate data sets, the detection of outliers using graphics is more difficult because we have to analyze a couple of variables each time, which results is a long and less reliable process because we can find an observation that is unusual for one variable and not unusual for the others, masking the results.

It is currently considered as a fact that people with disabilities have always existed throughout history (Silva, 1986; Carvalho, 2001). Worldwide, disabled people have worse health prospects, lower levels of education, lower economic participation and higher poverty rates compared to people without disabilities (Hawking, 2011). Disability continues to be considered a universal challenge with social and economic costs for individuals, families, communities and nations; it varies according to a complex combination of factors, including age, sex, exposure to environmental risks, socio-economic status, culture and available resources; they are associated with chronic health problems; global aging; and finally; disabled people and households with a disabled member face the worst economic and social realities, compared to people who do not have disabilities.

The statistical analysis of compositional data is based on determining the appropriate transformation from the simplex to real space. Possible transformations and outliers strongly interact: parameters of transformations may be influenced particularly by outliers and the result of good-on-fit tests will reflect their presence (Greenacre, 2019). In section 2, outliers were defined with the methods that were used, compositional data and description of the variables; in section 3, results and discussions, and finally; 4, conclusion

2. Methodology:

2.1. Motivation

In order to better meet the needs of disabled people, it is important to take into account the situation within each municipality. For the elaboration of this work, we considered the data set of the 20800804 interviewees who composed the Sample of the respondents of the Complete Demographic Census Questionnaire of the Brazilian Institute of Geography and Statistics (IBGE) for São Paulo State aggregated in the 645 municipalities. For this work it was proposed to use the compositional data of each municipality applying linearly transformed regression using Neper logarithms.

In statistical terms, it is noted that there are few published works that take into account the situation of disabled people distributed throughout the different municipalities.

2.2. Study of outliers

It is defined as outlier when the value of a variable is different from other. This value can be higher or lower. Bivariate case

For the univariate case, the outlier can be characterized by assuming a higher or lower value in relation to the others. jointly evaluates two variables in detecting outliers, and, finally, multivariate, when jointly evaluating three or more variables.

Among the possible causes for the occurrence of outliers, we can mention: measurement errors; typing or transcription errors; errors for considering one or more samples that do not belong to the population of interest. The importance of studying outliers lies in the fact that their presence can lead to false alternatives and interpretations (Barnett and Lewis, 1994; Rosado, 2006).

In this proposal, for procedures will be studied for the identification of outliers in a database:

a) Principal component analysis.

It was introduced by Pearson (1901) and developed independently by Hotelling (1933). It is a technique that linearly transforms a set that explains a substantial portion of the information in the original set. The original variables ($\boldsymbol{X_1}$, ..., $\boldsymbol{X_p}$) are transformed into $\boldsymbol{p}$ variables ($\boldsymbol{Y_1}$, ..., $\boldsymbol{Y_p}$), called main components, so that $\boldsymbol{Y_1}$ is the one that explains most of the total variability of the data, $\boldsymbol{Y_2}$ explains the second largest share and so on. The objectives of the principal component analysis are:

- Reduction of the dimensionality of the data;
- Obtaining interpretable combinations of variables;
- Description and understanding of the correlation structure of the variables.

The analysis is carried out in order to summarize the correlation pattern between the variables and it is often possible to arrive at sets of variables that are not correlated with each other, thus leading to a grouping of them. Develop the interrelationship between variables, that is, obtain factors common to all $\boldsymbol{p}$ variables describing their dependency structure through the construction of factors and seek latent variables that represent linear combinations of a group of variables under study that are, for related.

To do a main component analysis we have the following steps:

i) Encode the variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, ..., $\boldsymbol{X_p}$ to have zero mean and variance one (standardization);

ii) Calculate the covariance / correlation matrix;

iii) Find the eigenvalues $\boldsymbol{\lambda_1}$, $\boldsymbol{\lambda_2}$, ..., $\boldsymbol{\lambda_p}$ and the corresponding eigenvectors $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, ..., $\boldsymbol{\alpha_p}$. So that the coefficients of the i-th main component are then the elements of $\boldsymbol{\alpha_i}$, while $\boldsymbol{\lambda_i}$ is its variance, and finally;

iv) Discard components that explain only a small proportion of the variation in the data.

In this work, the points of the dispersion diagram of the first versus the second main component located outside the 95% confidence ellipse will be considered as discrepant cases.

b) Mahalanobis Distance

For each of the $\boldsymbol{n}$ samples and p variables, the Mahalanobis distance ($\boldsymbol{Di}$) is calculated by the expression:

$$D_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})} \qquad (7)$$

for $\boldsymbol{i = 1,...,n}$.

where $S=\sum_{i=1}(x_i-\bar{x})'(x_i-\bar{x})$ is the sample variance-covariance matrix; and, $(x_i-\bar{x})$ it is the vector of difference between the concentrations measured in one group and the average of the concentrations in the other group (Oliveira and Munita, 2003).

c) Euclidean Distance

Be $\underset{\sim}{X}=(X_1,\cdots,X_p)'$ a random vector with E(*X*) $=\mu=(\mu_1,\cdots,\mu_p)'$ e **Cov(*X*) = *Σ*.** We have the Euclidean Distance squared (***DE***^{2}) between a vector of values from the sample X and the vector of means µ is given by. $DE^2=(x-\mu)'(x-\mu)$.

d) Standardized Euclidean Distance (Pearson)

This quadratic distance is considered appropriate when heteroscedasticity occurs and can be calculated by the following expression:

$$d^2_{(x,y)}=(x-\bar{x})^T D^{-1}(x-\bar{x}),$$

where $\bar{x}$ is the centroid of the group and ***D***$^{-1}$ is the inverse of the diagonal of the covariance matrix ***S***.

e) Box plot

Box plot It is a graph showing characteristics the data sorted as Q_1 (first quartile), which corresponds to the position of the first 25% of data; M_d (median), which represents the position of first 50% of the same data, and finally, Q_3 (third quartile), which corresponds to the position first 75% of the data already sorted. Consider a rectangle on the basis determined by Q_1 and Q_3, as shown in Figure 1. Marked with one segment a median position. Consider the limits *liminf* = Q_1 - 1.5 (Q_3 - Q_1) and *limsup* = Q_3 + 1.5 (Q_3 - Q_1) where *liminf* and *limsup* are the limits upper and lower, respectively, for concentrations of each variable separately. the observations that are above the upper limit (*limsup*) or below the lower limit (*liminf*) observations are inconsistent with the established too, are called outliers in the case univariate analysis (Baxter, 2003; Bussab and Moretin, 2009).

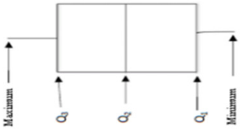


Figure 1. Box-plot

Box plot Figure 1 shows the portion of data (50%) between the first and third quartile and position of median. The Box plot gives an idea of the position, asymmetry, tails and univariate outliers. The central position is given by the mean and the dispersion is given by $Q_3 - Q_1$

2.3. Data compositional

In statistics, compositional data are quantitative descriptions of parts of a whole, which communicate information exclusively in relation to the whole (Greenacre, 2019). The most striking feature of this type of data is that its sum is always equal to a constant (1 for proportions and 100 for percentages). Such data is very common in areas of research such as geology and soil science. Examples of compositional data are the size distribution of mineral particles (sand, saltpeter and clay) in a soil or the concentration of cations in the soil solution. In economics, compositional data can be found in the form of family budgets, which are the proportions spent on food, education, travel, entertainment, among others. Government budgets that allocate the proportions to be spent on health, education, science, defense, culture and so on. For this article, data from the 2010 IBGE Demographic Census are being considered for the different municipalities considering proportions of different variables such as disabled people, education level, type of main job, among others.

The first recommendations related to the statistical analysis of compositional data, refer to an article by Karl Pearson from 1897 on spurious correlations. The article points out problems arising from the use of traditional statistical methods, as parts of a whole. But his warnings were ignored until around 1960, when the geologist Felix Chayes (1960) also warned against the application of standard multivariate analysis for compositional data, in order to avoid inconsistencies due to the unit sum constraint.

It was not until 1980 that John Aitchison systematized the theory for compositional data. Presenting a sample space adequate to its restrictions with its own probability distribution and the possible transformations for the real sample space. Thus, according to Aitchison (1986), a set of n vectors $\boldsymbol{y_i}$, $\boldsymbol{i = 1}$, ..., $\boldsymbol{n}$, whose elements have the restrictions $\boldsymbol{y_{i1}} > 0$, ..., $\boldsymbol{y_{iD}} > 0$, and, where D is the number of components (part) of the vector. These sets of vectors are defined as compositional data and show variability from vector to vector (Oliveira and Munita, 2011).

Compositional data are those that establish the information in a relative way, they are parts of a whole, in most cases they are considered as closed data, the sum of the data in each row or column is constant, in such a way that it represents 100%, archaeological data are classic examples. Compositional data has important particular properties that assist in the application of standardized statistical techniques in such concentration data. These statistical techniques are standardized for use in interval data ranging from **- ∞** to **+ ∞.** If one component increases, another must remain constant and another must decrease. This means that the results of standard statistical analysis of the relationship between concentration data components or parts in a compositional data set can be overshadowed by spurious effects (Bucciantti, 2006).

Applying the logarithmic transformation converts the ratios to an interval scale, reduces the effect of outliers and symmetrizes the distributions of the ratios. Therefore, logratios will generally be used as the basic statistical variables for data composition analysis (Greenacre, 2019).

In this study, proportions of variables related to disability were used as a function of variables related to education, family, work, housing conditions, other assets and quality

2.3. Variable's descriptions

Figure 2 illustrates the variables used in this research and notice that divided into identification, disabled peope, education, family, work and income, occupation housing, conditions housing, basic improvements and life quality.

3. Reult:

For this work, data available in the 2010 Demographic Census was considered and the following procedures were passed for each of the different variables:

Step 1: aggregate for each of the different levels and municipalities;

Step 2: obtaining the compositional data;

Step 3: for each of the compositional data, the Neper logarithm was calculated;

Step 4: for each of the different topics were calculated:

- i) Score of the main components and 95% confidence ellipse;
- ii) Euclidean distance;
- iii) Distance from Mahalanobis, and, finally;
- iv) Pearson distance.

Step 5: For method i) the points located outside the ellipse were considered as outliers and for methods ii), iii) and iv) the ordering was done by box-plot and the number of outliers by the box plot criterion previously defined;

Step 6: determining the number of outliers considering for each topic and for methods i), ii), iii) and iv), and finally;

Step 7: Comparative study between these quantities of outliers by method and topic.

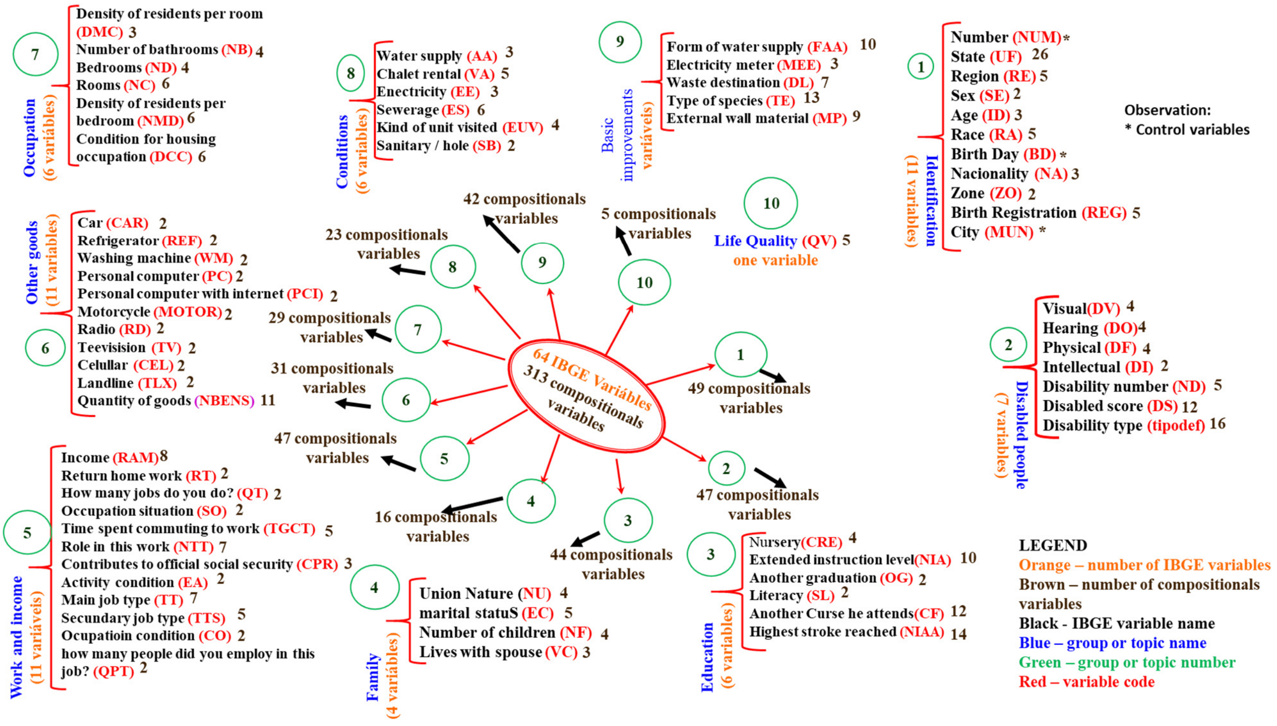
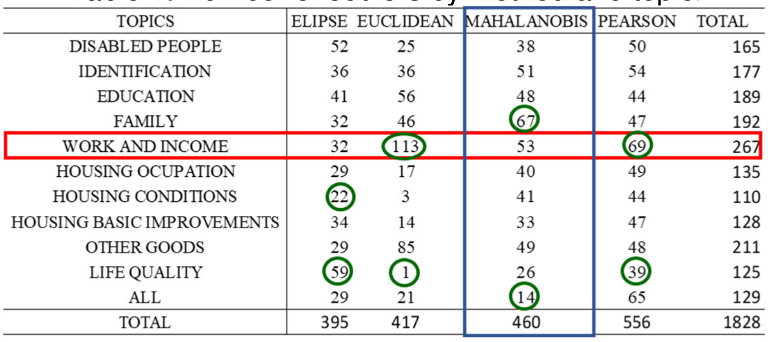

7
Density of residents per room (DMC) 3
Number of bathrooms (NB) 4
Bedrooms (ND) 4
Rooms (NC) 6
Density of residents per bedroom (NMD) 6
Condition for housing occupation (DCC) 6
Occupation
(6 variábles)
8
Water supply (AA) 3
Chalet rental (VA) 5
Enectricity (EE) 3
Sewerage (ES) 6
Kind of unit visited (EUV) 4
Sanitary / hole (SB) 2
Conditions
(6 variables)
9
Form of water supply (FAA) 10
Electricity meter (MEE) 3
Waste destination (DL) 7
Type of species (TE) 13
External wall material (MP) 9
Basic improvements
variáveis)
1
Number (NUM)*
State (UF) 26
Region (RE) 5
Sex (SE) 2
Age (ID) 3
Race (RA) 5
Birth Day (BD) *
Nacionality (NA) 3
Zone (ZO) 2
Birth Registration (REG) 5
City (MUN) *
Identification
(11 variables)
Observation:
* Control variables
Other goods
(11 variables)
6
Car (CAR) 2
Refrigerator (REF) 2
Washing machine (WM) 2
Personal computer (PC) 2
Personal computer with internet (PCI) 2
Motorcycle (MOTOR) 2
Radio (RD) 2
Teevisision (TV) 2
Celullar (CEL) 2
Landline (TLX) 2
Quantity of goods (NBENS) 11
42 compositionals variables
23 compositionals variables
5 compositionals variables
10
Life Quality (QV) 5
one variable
29 compositionals variables
31 compositionals variables
64 IBGE Variábles
313 compositionals variables
49 compositionals variables
2
Visual(DV) 4
Hearing (DO)4
Physical (DF) 4
Intellectual (DI) 2
Disability number (ND) 5
Disabled score (DS) 12
Disability type (tipodef) 16
Disabled people
(7 variables)
47 compositionals variables
5
Income (RAM)8
Return home work (RT) 2
How many jobs do you do? (QT) 2
Occupation situation (SO) 2
Time spent commuting to work (TGCT) 5
Role in this work (NTT) 7
Contributes to official social security (CPR) 3
Activity condition (EA) 2
Main job type (TT) 7
Secundary job type (TTS) 5
Ocupatioin condition (CO) 2
how many people did you employ in this job? (QPT) 2
Work and income
(11 variáveis)
4
16 compositionals variables
3
44 compositionals variables
47 compositionals variables
4
Union Nature (NU) 4
marital statuS (EC) 5
Number of children (NF) 4
Lives with spouse (VC) 3
Family
(4 variábles)
3
Nursery(CRE) 4
Extended instruction level(NIA) 10
Another graduation (OG) 2
Literacy (SL) 2
Another Curse he attends(CF) 12
Highest stroke reached (NIAA) 14
Education
(6 variables)
LEGEND
Orange – number of IBGE variables
Brown – number of compositionals variables
Black - IBGE variable name
Blue – group or topic name
Green – group or topic number
Red – variable code

Figure 2. Variables considered in this analysis.

Table 1 shows the quantities of outliers detected by topic and method. Note that the red strip contains the largest amount of outliers per topic, blue by methods and in green, minimum and maximum in each method.

Table 1. Number of outliers by method and topic.



TOPICS	ELIPSE	EUCLIDEAN	MAHALANOBIS	PEARSON	TOTAL
DISABLED PEOPLE	52	25	38	50	165
IDENTIFICATION	36	36	51	54	177
EDUCATION	41	56	48	44	189
FAMILY	32	46	67	47	192
WORK AND INCOME	32	113	53	69	267
HOUSING OCUPATION	29	17	40	49	135
HOUSING CONDITIONS	22	3	41	44	110
HOUSING BASIC IMPROVEMENTS	34	14	33	47	128
OTHER GOODS	29	85	49	48	211
LIFE QUALITY	59	1	26	39	125
ALL	29	21	14	65	129
TOTAL	395	417	460	556	1828

The confidence ellipse allows the detection of divergent bivariate data, with the advantage of graphic visualization and the disadvantage of losing information when working with ***p*** greater than two variables for each of the different topics. To illustrate, follow the example of Figure 3, which considers disabled people with 19 compositional variables. Note that the data considered outliers are those that are outside the ellipse.

The Euclidean distance is recommended for the identification of samples distant from the sample set, appropriate for independent and heteroscedastic variables. On the other hand, the Mahalanobis distance is recommended for the identification of samples out of its tendency. Therefore, all different outlier samples obtained in at least one of the methods must be considered Euclidian and Mahalanobis Barroso and Artes 2003). In many experimental situations, this type of quadratic distance has an intuitive appeal because it contemplates the covariance structure between the different variables.

Mahalanobis distance and lever statistics are also widely used to detect outliers, especially in the development of models based on linear regression. A point that has a greater Mahalanobis distance than the rest of the sample population of points is said to have greater leverage since it has a greater influence on the slope or the coefficients of the regression equation. The Mahalanobis distance is also used to determine multivariate outliers. Regression techniques can be used to determine whether a specific case in the context of a population is an outlier or not by combining two or more scores of variables. A case does not need to be a univariate outlier in one of the variables to be a multivariate outlier. The statistical significance of the Mahalanobis distance in detecting multivariate outliers can be assessed by a chi-square test with ***k*** degrees of freedom.

It allows to be used as a comparison technique regarding the separation between different groups allowing to evaluate the extent and the direction of the gaps between the average values of the variables used in the discrimination. The differences between each pair of groups being compared are thus examined simultaneously through the various variables, which can be correlated, so that the information provided by one of them may not be independent from that provided by the others.

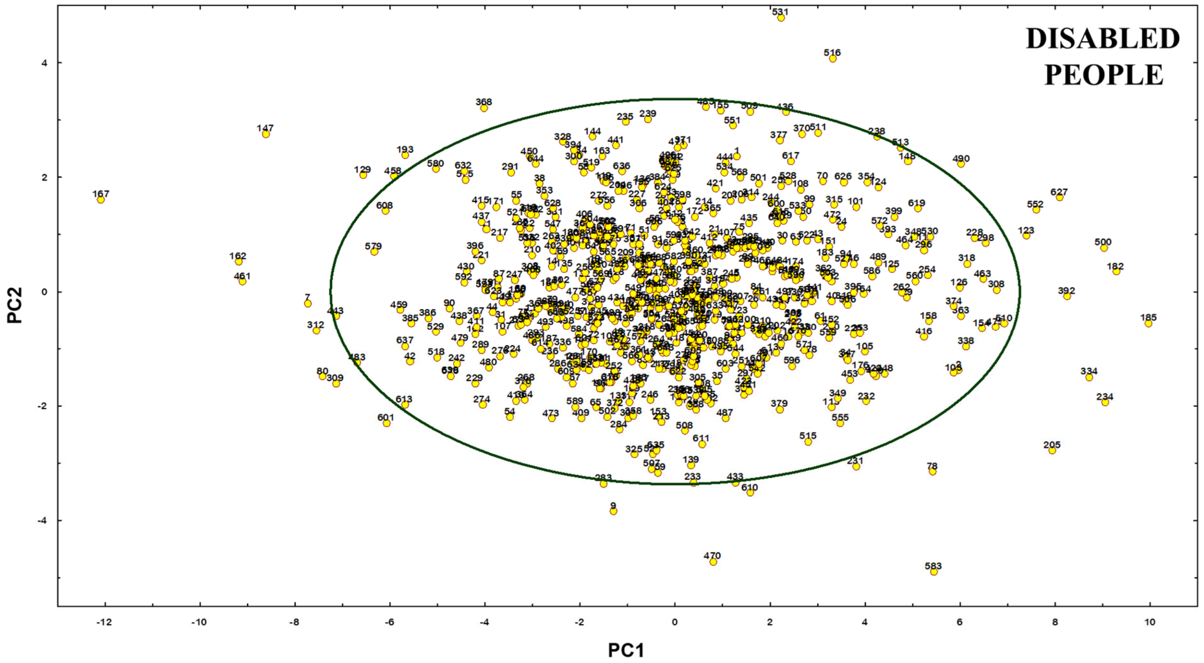


Figure 3. 95% confidence ellipse for two first principals' components.

This method of representing differences between groups takes into account any correlation that exists between the variables used and also regardless of the units of measurement with which the variables are expressed.

Euclidean distance is more appropriate for independent and homoscedastic variables, if heteroscedasticity occurs, the alternative is to use Pearson distance and when the variables are jointly correlated and heteroscedastic the alternative becomes the use of Mahalanobis distance.

Box-Plot allows to easily evaluate the typical values, the asymmetry, the dispersion and the discrepant data of data sets referring to the quantitative variables for the univariate cases.

For the different municipalities, the largest number of outliers per method are: Ellipse of confidence (Águas de São Pedro); Euclidean (Babinos and Pracinha); Mahalanobis (Babinos, Pracinha and Ribeira), and finely; Pearson (Barra do Turvo and Iporanga).

4. Discussion and Conclusion:

Among the topics, the largest number of outliers was work and income with 267 cases, and, among the different methods, the Mahalanobis distance in 490 cases.

For the different methods, the topics with the largest and most outliers respectively are: 95% Confidence Ellipse (life quality and basic improvements); Euclidean Distance (work and income, and, quality of life); Mahalanobis Distance (family and all topics), and finally; Pearson's Distance (work and income, and life quality).

Sensitivity for outlier detection varies according to the topic and method of detection resulting in a different city profile from the others.

To continue this research, consider creating, simulating and using more robust methods for detecting multivariate outliers.

References:

AITCHINSON, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, London, UK.

BARNETT, V.; EWIS, T. (1984). Outliers in Statistical data. Jonh Wiley, New York, USA.

BARROSO, L. P.; ARTES, R. (2003). Multivariate Analysis. 1. ed. Lavras: Região Brasileira da Sociedade Internacional de Biometria.

BAXTER, M.J. Compositional data analysis in archaeometry. Universitat of Girona, 2003.

BUCCIANTTI, A. (2006). Compositional Data Analysis in the Geosciences from Theory to Practice. Chapman & Hall/CRC Press, London, UK.

CARVALHO, J.O.F. (2001). Technological solutions to enable the visually impaired to access distance education in higher education. Facudade de Engenharia Elétrica e Computação, UNICAMP, Campinas-SP.

CHAYES, F. (1960). Correlation in closed tables: Annual Report of the Director of the Geophysical Laboratory, Carnegie Institution of Washington, no. 59, p. 165-168.

GREENACRE, M. (2019). Compositional Data Analysis in Practice. Chapman & Hall/CRC Press, London, UK.

GIROLDO, F.R.S. (2008). Some robust methods for detecting multivariate outliers. Masters' dissertation, IME-USP, São Paulo-SP.

HAWKING, S. (1986). World report on disability WHO,. Genava, Switzerland

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441. https://doi.org/10.1037/h0071325

MORETTIN, P.A. (2009). Basic Statistics. Editora Atual.

OLIVEIRA, P. T. M. S.; MUNITA, C. S. Influence of critical value in detecting outliers in Archaeometry. In: 48a Annual Meeting of RBRAS/10a SEAGRO, 2003, From July 7 to 11, 2003(1):545—550. Text in Portuguese. Link: https://www.researchgate.net/publication/282809986_Influencia_do_valor_critico_na_deteccao_de_valores_discrepantes_em_arqueometria

OLIVEIRA, P. T. M. S.; MUNITA, C.S., Comparative studies of the Biplot and Multidimensional Scaling Analysis in experimental data. In: 58th ISI World Statistics Congress in Dublin, 2011, Dublin, Ireland, From August 21 to 26, 2011. Link: https://www.researchgate.net/publication/282778475_Comparative_studies_of_the_Biplot_and_Multidimensional_Scaling_Analysis_in_experimental_data

PEARSON, K. (1897). Mathematical Contributions to the Theory of Evolution – on a Form of Spuriouscorrelation Which May Arise When Indices Are Used in Measurement of Organs.Proc. Royal Societ., London, 60:489—498.

PEARSON, K. (1901). Mathematical Contributions to the Theory of Evolution VII. On the correlation of Characters not Quantitatively Measurable. Phil. Trans. Royal Society London A, 190;1—47.

ROSADO, F. (2006). Outliers em Dados Estatísticos. SPE, Lisboa, Portugal.