Brazilian Journal of Biometrics, **43**, 1-23 (2025)
doi:10.28951/bjb.v43i1.712

BRAZILIAN JOURNAL OF
BIOMETRICS
ISSN:2764-5290

ARTICLE

Procrustes analysis, multivariate regression, variable selection and outlier detection in compositional data for social vulnerability

Paulo Tadeu Meira e Silva de Oliveira*[1]

[1]Department of Statistics, University of São Paulo, São Paulo – SP, Brazil
*Corresponding author. Email: poliver@usp.br

(Received: December 17, 2023; Revised: May 23, 2024; Accepted: June 6, 2024; Published: February 11, 2025)

Abstract

Vulnerability means delicate and weak in the behavior of people, objects, situations and ideas. People considered "socially vulnerable" are those who lose their representation in society and generally depend on help from third parties to ensure survival. The main characteristics that mark this vulnerability are precarious housing conditions, sanitation, non-existent means of subsistence and the absence of a family environment. Among the different types, they highlight youth in the area of health, marginalization, exclusion and territorial. Social Vulnerability Index (SVI) is composed of indicators of income and social impairment in dimensions such as identification, housing, education, income, poverty, family, work and other assets. Variable selection is finding a subset of variables that best explains a response vector, without losing relevant information. Procrustes Analysis is a method that aims to determine how much a subset of variables best represents the structure of the original data. Compositional data are quantitative descriptions of the parts of a whole, which convey information in a relative way. Principal components are linear combinations of all original variables, independent of each other and estimated with the purpose of retaining, in order of estimation, the maximum amount of information to explain the total variance. Univariate outliers are observations that differ greatly from the others. Multivariate outlier corresponds to cases involving two or more variables. In this work we use the Procrustes method and other regression methods to select variables formed from compositional data after detecting multivariate outliers using Mahalanobis Distance and comedian approach.

Keywords: Procrustes analysis; Compositionals data; Variables selections; Multivariate outliers; Multivariate regression; Social vulnerability.

1. Introduction

Social vulnerability is a multidimensional concept that concerns a condition of material or moral fragility of individuals or groups in the face of risks produced by the economic-social context. In etymological terms, it rescues the connection between the Latin words *vulnerare*, which means to hurt, injure, harm, and *bĭlis* – susceptible, originating the word vulnerability (Carmo *et al.*, 2018). Social vulnerability is understood as the negative result of the relationship between the availability of material or symbolic resources of actors, whether individuals or groups, and access to the structure of social, economic and cultural opportunities originating from the State, the market and society (Pessalacia *et al.*, 2010). Some of the main characteristics that mark the state of social vulnerability are precarious housing and sanitation conditions, non-existent means of subsistence and the absence, for example, of a family environment (Nunes and Andrade, 2009).

Among the different types of vulnerability, the following stand out: marginalization and exclusion, in the health, territorial and youth. A situation of social vulnerability is related to the exclusion of citizens, representation lack and opportunities. Furthermore, it is a multifactorial concept, that is, it can occur due to issues of housing, income, education, among others. Among the different consequences, fragility in family relationships, social isolation of young people, illegal behavior and/or other violence types that compromise Life Quality (LQ) stand out. LQ is understood as the relationship between the environment, psychological and physical aspects, independence level, personal beliefs and social relationships (Costa, 2012). Social risks are not limited to situations of poverty, but also factors such as unemployment, difficulties in social integration, illnesses and abuse, among others.

The IVS is a quantitative analysis expression composed of income indicators, urban infrastructure, human capital, commitment and social factors. Understood as aspects that interfere, for example, with the permanence and success of students in schools and whose average characterizes a situation of vulnerability. For this work, variables related to this topic will be considered, obtained from data from the IBGE 2010 Demographic Census, whose ranges are defined, as shown in Figure 1 below.

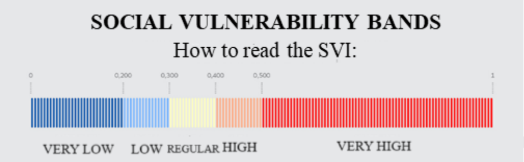


Figure 1. IVS Bands.

From Figure 1 it is possible to see that the IVS is considered very high when this index is greater than 0.5; high when this index is greater than 0.4 and less than 0.5; medium when this range is between 0.3 and 0.4; low when it is between 0.2 and 0.3, and, finally; very low when it is less than 0.2.

Selecting variables means choosing a subset that retains the most important predictor variables while excluding the others and that this subset fits as well as the model that includes all predictor variables (Oliveira, 2015).

According to Oliveira (2008) and (Boogaart and Tolosana-Delgado, 2013) a model must be as simple as possible and as complicated as necessary and that no statistical procedure can identify a true model.

In this study, the proposal is to use Procrustes analysis and multivariate regression to select variables, carry out a comparative study and determine how much the new subset of variables represents the original data structure (Kranowski, 1987); detection of multivariate outliers to evaluate municipalities considered outliers, and selection of variables with the objective of including those necessary to adjust the model, and, simultaneously, discarding those considered unnecessary as a form of simplification (Oliveira, 2008).

Compositional data are quantitative descriptions of parts of a whole that transmit information in a relative way and their measurements usually involve probabilities, proportions and percentages (Aitichison, 1986) and can suffer from pollution, mainly due to inadequate treatment during sampling or in the laboratory. This type of irregularities can affect a significant part of the data set (Barbosa, et. al, 2018).

An outlier is defined as the value of the observation of a variable that is different in relation to the others and it is one of the oldest problems in statistical analysis, and in recent years interest in this area has progressively increased.

The presence of these points can cause distortions in the results when adjusting models and estimates. After this diagnosis, a decision can be made regarding the outliers. One possibility is to correct them if there was, for example, an error in transcribing the data and if they are valid points, they should be treated differently from the others, either as a weighting, use of a more robust method of analysis or as an analysis Special.

Now, among the possible causes for the occurrence of outliers we can mention: measurement errors; typing or transcription errors; errors due to considering one or more samples that do not belong to the population of interest, as shown in Figure 2. The importance of this study lies in the fact that their presence can lead to false alternatives and interpretations (Barnett and Lewis, 1994).

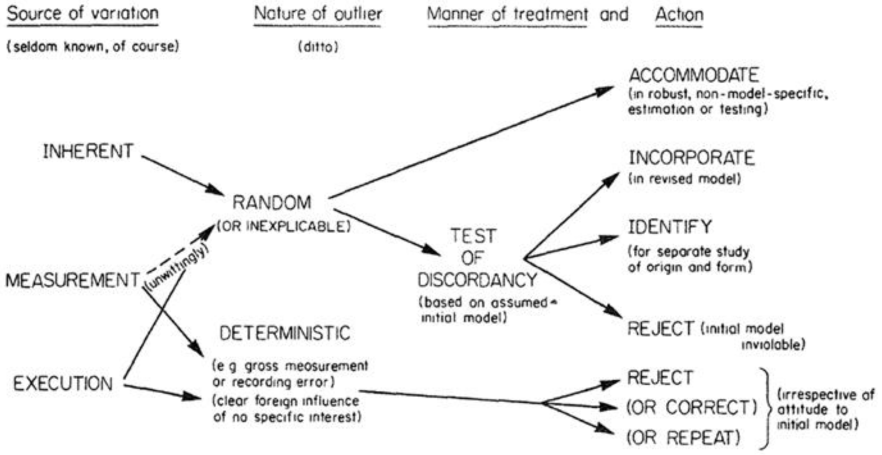

Source of variation
(seldom known, of course)
Nature of outlier
(ditto)
Manner of treatment and Action
ACCOMMODATE
(in robust, non-model-specific, estimation or testing)
INHERENT
INCORPORATE
(in revised model)
RANDOM
(OR INEXPLICABLE)
TEST OF DISCORDANCY
(based on assumed initial model)
IDENTIFY
(for separate study of origin and form)
MEASUREMENT
(unwittingly)
REJECT (initial model inviolable)
DETERMINISTIC
(e g gross measurement or recording error)
(clear foreign influence of no specific interest)
EXECUTION
REJECT
(OR CORRECT)
(OR REPEAT)
(irrespective of attitude to initial model)

Figure 2. Treatment of outliers.
Fonte: Barnett (1994).

From Figure 2, it can be seen that outliers can be a source of inherent variation (essential characterization), measurement and execution; regarding its random or deterministic nature; form of treatment, discordance test, and finally; decision to accommodate, incorporate, identify, reject, correct or repeat.

Outliers are observations with a unique combination of characteristics that are identifiable as being different from other observations. Typically, it is considered as an unusual value in a variable because it is high or low in relation to the others, or an odd combination of values across several variables that form the marginal observation in relation to the others and its presence of outliers allows conclusions to be drawn about data quality as well as atypical phenomena that may arise.

In compositional data, each observation consists of *d* parts. However, as in a *d*-dimensional space, $d \geq 2$ there are an infinite number of directions that each observation can take a situation. To detect outliers in a situation like this, distance such as Robust Mahalanobis, Comedian Approach and Adjusted Atypicality are usually used (Sousa, 2016; Leite, 2019; Maltez, 2020).

When detecting outliers in compositional data, it is assumed that the multivariate data are compositions. In a situation like this, instead of identifying outliers directly in the original space, it is customary to express the compositions in log-ratio coordinates, and then; apply the usual multivariate outlier detection methods.

In this work we are starting from a robust multivariate statistical analysis to identify municipalities considered atypical when it comes to vulnerability issues based on a set of variables obtained from the 2010 census and Human Development Atlas.

In statistical terms, there is little work on variable selection using procrustes analysis, multivariate regression for compositional data and detection of multivariate outliers.

In section 2 we present a motivation for the problem, we define and characterize the variables considered in vulnerability, in materials and methods we describe statistical methods such as procrustes for variable selection, compositional data and confidence ellipse for the first two principal component scores and bagplor for two-dimensional graphics, comedian approach and Mahalanobis distance for multivariate outlier detection; in section 3 we show results and discussions; in section 4 conclusions and suggestions for future work, and, finally; in section 5 we mention bibliographic references used in this research.

2. Materials and Methods

2.1 Motivation

In order to better study social vulnerability, it is important to take into account the situation within each municipality, evaluating and studying the profiles of municipalities that are different from the others and also selecting which variables best explain its occurrence in different communities. To prepare this work, we considered a set of 20800804 interviewees who made up the sample of respondents to the Complete Questionnaire of the Demographic Census of the Brazilian Institute of Geography and Statistics (IBGE) aggregated in the 5565 Brazilian municipalities in a compositional manner together with data from the UNDP (United Nations Program for Development) that makes up the HDI (Human Development Index).

Over time, in the most diverse areas of research such as Medicine, Biology and Social Sciences, sets of data emerge that have a compositional structure with several characteristics and properties that are important for any statistical analysis. In most cases, a very common aspect in the analysis of these data is that their interpretation is made through the application of traditional techniques intended for real multivariate data after convenient transformations in the original data.

In any data set, it is important that, before applying any statistical technique, you carry out a careful analysis of its components, as there may be points that stand out regarding one or more study variables that could harm the statistical modeling of the data, such as detecting outliers so that they cannot compromise the fit.

The treatment of outliers, whether for identification, removal or both, has been exhaustively researched in the most diverse areas of knowledge such as data mining, machine learning and information theory (Barbosa, 2017).

For this work, it was proposed to use compositional data from different Brazilian municipalities, identify cases considered outliers and evaluate them by applying variable selection methods. In statistical terms, there are few published works that evaluate and classify the profiles of different municipalities.

2.2 Variable’s descriptions

The variables were obtained directly from the questionnaire applied to the sample data set that responded to the Complete Questionnaire and transformed into compositional data aggregated by municipalities (Oliveira, 2014) in the compositional data format.

In this research, variables related to: i) identification were considered: state, region, sex, age, race and zone; ii) instruction: expanded instruction level; iii) family: number of children, union nature, marital status and living with a partner; iv) work: main job type, secondary job type, home-work return, how many jobs, pension, income; v) housing conditions: water supply, form of water supply, rent, resident density per room, electricity, housing occupancy condition, sewage system, unit visited type, electricity meter, number of bathrooms, destination waste, number of rooms, bedrooms, toilet/hole, type of species, external wall material and electricity, and finally; vi) other assets: radio, television, refrigerator, washing machine, motorcycle, car, landline telephone, cell phone, cell phone with internet and number of assets, and finally; vii) UNDP: human development index as can be seen in Figure 3.

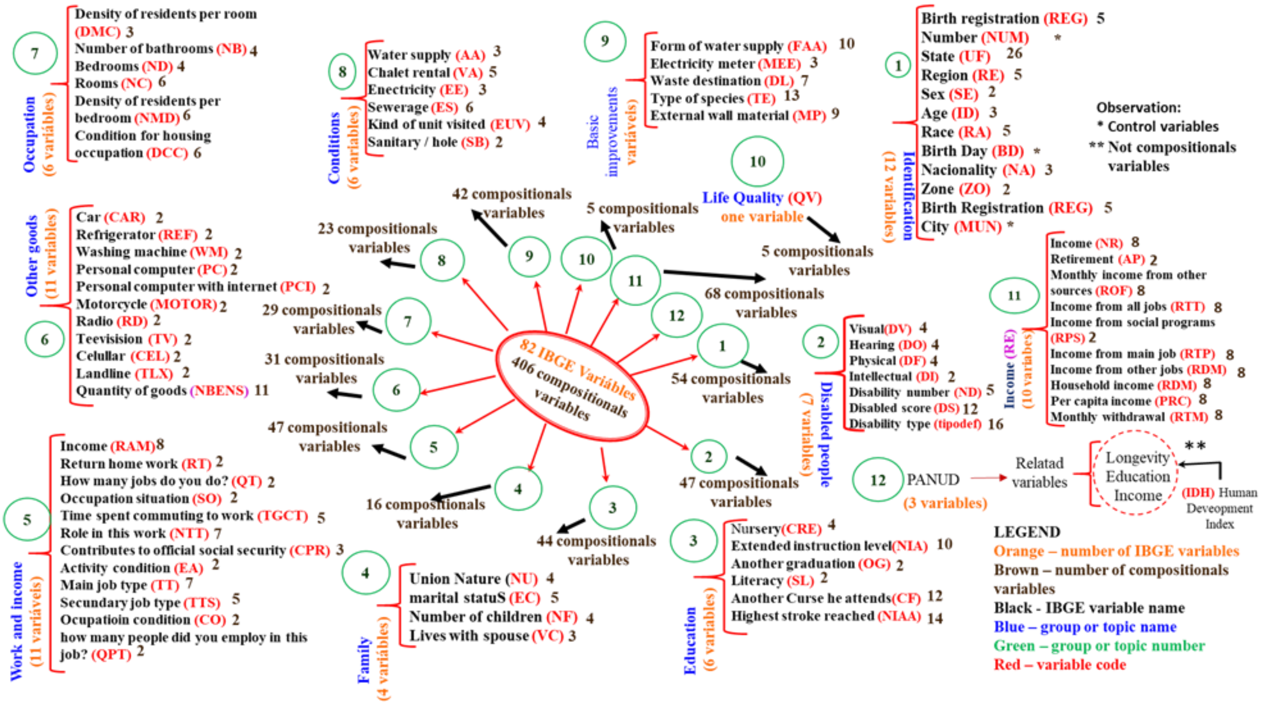

7
Occupation
(6 variábles)
Density of residents per room (DMC) 3
Number of bathrooms (NB) 4
Bedrooms (ND) 4
Rooms (NC) 6
Density of residents per bedroom (NMD) 6
Condition for housing occupation (DCC) 6
8
Conditions
(6 variables)
Water supply (AA) 3
Chalet rental (VA) 5
Enectricity (EE) 3
Sewerage (ES) 6
Kind of unit visited (EUV) 4
Sanitary / hole (SB) 2
9
Basic improvements
variáveis)
Form of water supply (FAA) 10
Electricity meter (MEE) 3
Waste destination (DL) 7
Type of species (TE) 13
External wall material (MP) 9
1
Identification
(12 variables)
Birth registration (REG) 5
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) *
Observation:
* Control variables
** Not compositionals variables
10
Life Quality (QV)
one variable
5 compositionals variables
42 compositionals variables
5 compositionals variables
23 compositionals variables
68 compositionals variables
29 compositionals variables
31 compositionals variables
47 compositionals variables
54 compositionals variables
47 compositionals variables
16 compositionals variables
44 compositionals variables
82 IBGE Variábles
406 compositionals 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
11
Income (RE)
(10 variabes)
Income (NR) 8
Retirement (AP) 2
Monthly income from other sources (ROF) 8
Income from all jobs (RTT) 8
Income from social programs (RPS) 2
Income from main job (RTP) 8
Income from other jobs (RDM) 8
Household income (RDM) 8
Per capita income (PRC) 8
Monthly withdrawal (RTM) 8
2
Disabled people
(7 variables)
Visual(DV) 4
Hearing (DO) 4
Physical (DF) 4
Intellectual (DI) 2
Disability number (ND) 5
Disabled score (DS) 12
Disability type (tipodef) 16
5
Work and income
(11 variáveis)
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
4
Family
(4 variábles)
Union Nature (NU) 4
marital statuS (EC) 5
Number of children (NF) 4
Lives with spouse (VC) 3
3
Education
(6 variables)
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
12
PANUD
(3 variables)
Relatad variables
Longevity
Education
Income
**
(IDH) Human Deveopment Index
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 3. Variable’s descriptions.

2.3. Principal component analysis

It was introduced by Pearson (1901) and independently developed by Hotelling (1933). It is a technique that linearly transforms a set of variables that explains a substantial portion of the information in the original set. The original variables (X_1, …, X_p) are transformed into p variables (Y_1, …, Y_p), called Principal components, so that Y_1 is the one that explains the largest portion of the total data variability, Y_2 explains the second largest portion and so on. The objectives of principal component analysis are: data dimensionality reduction; obtaining interpretable combinations of variables, and finally; description and understanding of their correlation structure.

The analysis is carried out with the aim of summarizing the correlation pattern between variables and, often, it is possible to arrive at sets of variables that are uncorrelated, thus leading to a grouping of them. Develop an interrelationship between the variables, that is, obtain factors common to all p variables describing their dependence structure through the construction of factors and seek latent variables that represent linear combinations of a group of variables under study that are, by in turn, related.

To perform a principal component analysis we have the following steps:

STEP 1: Code the variables X_1, X_2, …, X_p to have a mean of zero and variance of one (standardization);

STEP 2: Calculate the covariance/correlation matrix;

STEP 3: Find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_\pi$. and the corresponding eigenvectors $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots, \boldsymbol{\alpha_p}$. So that the coefficients of the *ith* principal component are then the elements of α_i, while λ_i is its variance, and finally;

STEP 4: Discard components that explain only a small proportion of the variation in the data.

In this work, the points of the scatter diagram of the first versus the second principal component located outside the 95% confidence ellipse will be considered as outliers.

2.4. Compositional data

The initial interest in compositional data arose at the end of the 19th century with Karl Pearson in the article "On a Form of Spurious Correlarion which May Arise When Indices Are used in the Measurement of Orgam" in which Pearson highlights problems with the interpretation of correlations between ratios whose Numerators and denominators have common parts influencing their study (spurious correlation in restricted data, that is, the existence of a statistical relationship between two or more variables, in which there is no logical explanation). In turn, only in 1986 through Aitchison were the fundamental concepts inherent to compositional data and a better approach to their analysis introduced.

Karl Pearson (1897) and Aitchison (1986) warned about spurious correlation in restricted data, that is, the existence of a statistical relationship between two or more variables, in which there is no logical explanation or theoretical meaning. Therefore, due to this problem, usual Multivariate Analysis methods are unable to interpret the correlation coefficients between data components. This fact frequently occurs when dealing with data in which the sum of the components is constant.

John Aitchison (1986) indicated three principles on which appropriate techniques for analyzing compositional data should be governed. When defining these principles, that author considered that in a statistical analysis of compositional data, only the proportions of the components contain relevant information. The three principles are:

- Scale invariance: When a problem is compositional, we must recognize that the absolute value of the parts that make up the samples is irrelevant, since equivalent compositions contain essentially the same information;
- Permutation invariance: The conclusions of a compositional analysis should not depend on the order of the parties involved;
- Subcompositional coherence: Analyzes on a set of parts of a composition must not depend on other uninvolved parts, meaning that the study of a subcomposition cannot lead to contradictory results with those obtained from the total composition.

In statistics, compositional data are quantitative descriptions of the parts of a whole, which exclusively communicate information in a relative way and made up of vectors with all positive components. The most striking characteristic of this type of data is that its sum is always equal to a constant (1 for proportions and 100 for percentages). Such data are very common in research areas such as geology and soil science. Examples of compositional data are the size distribution of mineral particles (sand, saltpeter and clay) of a soil or the concentration of cations in the soil solution. In Economics for the analysis of components in household spending. In Medicine it can be applied to the composition of the body (e.g. fat, bones, muscles), in the Food Industry to the composition of foods, among others.

A data matrix of dimension $n \times p$ is compositional if the sum of its rows is constant, and sub compositional if the variables form subsets of a compositional data matrix. Let us consider an $n \times p$ data matrix fully compositional if the sum of the rows is a constant, and subcompositional if the variables are a subset of a fully defined composition data set. Such data occurs widely in archaeometry, where it is common to determine the chemical composition of glass, ceramics, metal or other artifacts using techniques such as neutron activation analysis and X-ray fluorescence (XRF) analysis, often revolving around know whether there are distinct chemical groups within the data and whether, for example, these can be associated with different origins or manufacturing technologies (Baxter, 1999). The sample space of compositional data is therefore simple space is a *D - 1* dimensional subset RD. Standard statistical methods can lead to misleading results if they are applied directly to original closed data.

Aitchison (1986) concluded that all analyzes of the parts that make up a whole could be carried out in terms of ratios of the parts of the composition. And, since the transformation of the logarithm of the ratios between variables (log-ratio) is a one-to-one correspondence in $\mathbb{R}$, the mathematical treatment of a quotient is simpler in terms of its logarithm. Thus, Aitchison (1986) proposed methodologies based on various types of log-ratio transformations. These transformations allowed the application of Multivariate Analysis procedures on the transformed data, then translating the conclusions drawn in terms of original data (Pawlowsky-Glahn et. al., 2015).

In summary, the log-contrast transformations, *alr* (transformation based on the logarithm of ratios with a single reference variable in the denominator), *ilr* (isometric transformation based on the choice of an orthonormal basis) and *clr* (isometric transformation based on the logarithm of ratios in relation to the geometric mean of the variables) must be taken into account in the analysis of compositional data. In general, the philosophy of log-contrast analysis can be summarized in five steps (Aitchison, 2005):

1. Formulation of the problem in terms of composition components;
2. Translation of this formulation in terms of log-contrast vectors of the composition;
3. Transformation of compositional data into log-contrast vectors;
4. Analysis of data expressed in log-contrasts using an appropriate usual multivariate analysis technique;
5. Interpretation of the results obtained in step 4 in terms of log-contrasts of compositions and in terms of the original variables.

In the case of this work, the *clr* transformation was considered, which is a transformation from S^D to R^D, and the result of an observation $x \in R^D$ is the transformed data $y \in R^D$, according to expression (1), with

$$y = \left(y_1, \cdots, y_D\right)' = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \cdots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right) \qquad (1)$$

Compositional data has important particular properties that assist in the application of standard statistical techniques to such concentration data. These statistical techniques are standardized for use on interval data ranging from $-\infty$ 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 obscured by spurious effects (Bucciantti, 2006).

Compositional data analysis can be divided into the following steps:

- Representation of data in log-ratios;
- Use of multivariate statistical techniques on data in transformed log-ratio coordinates, and finally;
- Interpretation of data in transformed and original coordinates.

In this work, compositional data is considered to be the sum of the proportions obtained between the levels of each variable related to vulnerability in topics such as identification, disability, family, education, work, housing and other assets (Aitchison, 2011).

More specifically, consider the problem of estimating the parts def_1, def_2, ..., def_{16} corresponding to the number of people with disabilities td_1, td_2, ..., td_{16} of a certain number of a municipality Q, appears frequently. The percentages corresponding to the types of disability td_1, td_2, ..., td_{16} in the 5565 municipalities in Brazil form a typical example. Naturally, it is of interest to analyze what these proportions would look like depending on certain contextual changes, for example, geographic location, time or availability of resources available in different municipalities to serve this population.

2.5. Procrustes analysis

It is a multivariate methodology for comparing the shape of two sets of data in an attempt to adjust to the other through transformations in one set of data that includes all the variables in the other that includes only the variables selected by the Procrustes method using one or more of the following transformations: translation that adds a common factor, rotation that rotates, reflection that reflects in a plane with a scale that multiplies by a homogeneous factor, so that the transformed data set can assume the closest form of fit to another group (Ferreira, 2004; Gower and Dijksterhuis, 2004).

The objective is to obtain a subset of variables resulting from the analysis that reproduce the original structure of the data, that is, to reduce the dimension of its set of variables without changing the structure of the data.

Initially, a data matrix is considered in which n is the number of municipalities and p is the number of variables; Suppose that the essential dimension of the data to be used in some comparison is k and that this dimension ensures that sufficient variability of the data is explained in the choice of k. Next, the score matrix is considered $Y_{n\times k}$ of the principal components that produces the best k-dimensional approximation of the original data configuration ***(q < p* e *q ≥ k)*** are sufficient to represent the same structure presented in Y. It is considered $\tilde{X}_{n\times q}$, the data matrix with the selected variables $\tilde{Z}_{n\times k}$ and ${}_k\tilde{Z}'Y_k = U\Lambda V'$, the matrix of principal component scores of the reduced data that produces the best k-dimensional approximation of the q-dimensional configuration defined in the subset of the data.

If the true dimension of the data is k, then Y can be seen as the true configuration, and $\tilde{Z}$ as the corresponding approximate configuration based only on the q variables.

The schematic diagram below shows the steps of the procedure, as shown in Figure 3.

$$\begin{tikzcd}
X_{n\times p} \arrow[r, "\text{to select}"] \arrow[d, "\text{PCA}"'] & \tilde{X}_{n\times q} \arrow[d, "\text{PCA}"'] \\
Y_{n\times k} \arrow[r, leftrightarrow, "\text{Procrustes}"] & \tilde{Z}_{n\times k}
\end{tikzcd}$$

Figure 4. Schematic diagram of the procrustes procedure.

To measure the discrepancy between the Y and $\tilde{Z}$ configurations, Procrustes Analysis (Sibson, 1978) was used, evaluating the adjustment between the two configurations by the residual sum of squares (M^2), which measures the loss of information about the structure of the data when only the q selected variables are used instead of the original p variables.

Let the configurations be: Y dimension matrix ($n \times k$) and $\tilde{Z}$ dimension matrix ($n \times k$), then, according to expression (2)

$$M^2 = traço\left\{YY' + \tilde{Z}\tilde{Z}' - 2\tilde{Z}Q'Y'\right\} \qquad (2)$$

Which can be rewritten according to expression (3) as

$$M^2 = traço\left[YY'\right] + traço\left[\tilde{Z}\tilde{Z}'\right] - 2traço\left[\Lambda\right], \qquad (3)$$

where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ and $Q = VU'$ where Q is an orthogonal matrix of dimension $k \times k$; U, Λ and V are obtained from the singular value decomposition (Golub, 1970) $\tilde{Z}'Y$ of the dimension matrix ($k \times k$), that is, ${}_k\tilde{Z}'Y_k = U\Lambda V'$.

Krzanowski (1996) shows that, under a given assumption, M^2 has a distribution proportional to chi-square with $nk - k(k-1)/2$ degrees of freedom, if the variables are not structured; the proportion is given by $\alpha\chi^2$ where $\alpha = \sigma^2\left(2p - 2k - i\right)/\left(p - k\right)$ and σ^2 is the estimate of the residual variance, obtained by the sum of squares of the elements of $U_{p-i}\,D_{p-i}\,V'_{p-k}$, divided by $(n-k-1)(p-k)$. The p-q variables are eliminated one by one, in increasing order of importance, and the elimination is done while M^2 does not exceed the critical value determined by the reference distribution.

The best subset of q variables is the subset that provides the smallest value of M^2 among all subsets of q variables.

To eliminate variables, the following procedure can be used:

a) Initially consider $q = p$, and for fixed k calculate the matrix of scores of the main components Y;

b) Use updating algorithms (Bunch *et al.*, 1978) to obtain and store the score matrix of the main components, successively excluding each variable;

c) Calculate M^2 for each score matrix and identify the X_u variable that provides the lowest M^2. Let $\tilde{Z}_{(u)}$ the corresponding matrix of scores be, and finally;

d) Eliminate the variable X_u. Do $Z = \tilde{Z}_{(u)}$ and return to step b) with $p - 1$ variables. Continue this cycle until only q variables remain.

2.6. Multivariate Regression for Fitting Compositional Data

Multivariate regression is also known as canonical correlation and allows, when faced with a number of possible dependent variables, to identify the one that is strongly explained by the set of predictor variables existing in the data set. This topic is what several authors call canonical correlation (Fiera, 2011; Fávaro and Belfiore, 2017). Canonical correlation, in turn, allows you to simultaneously consider a large number of dependent and independent variables with the advantage of allowing it to be applied when the independent variables are not known, nor is the best candidate for the dependent variable.

To work with compositional data, a first alternative presented by researchers such as Connor and Mosimann (1969) was to work with the Dirichlet distribution. However, due to the existence of some restrictive properties such as complete compositional independence and correlation structure that induces negative correlation to data with positive correlation, an alternative was to apply certain transformations proposed by Brehm *et al.* (1998) and Buccianti (2013) who make comparisons between Dirichlet models and transformations for the space of real numbers.

An alternative to this work is to follow the proposal described in Aitchison (1986) for data analysis using transformations based on the following steps:

i) formulate the problem in terms of the components of the composition;

ii) transform the compositional data into log ratios (with the appropriate choice of reference class);

iii) analyze the transformed data through multivariate statistical analysis;

iv) apply the inverse transformation to the terms of the compositions obtained in iii).

On the other hand, according to Greenacre (2019), compositional data can be considered as a set of independent variables. Given this situation, the logratio of all pairs is considered a candidate and the problem becomes one of variable selection.

2.7. Variable's selection

Selecting variables means choosing a subset that retains the most important predictor variables while excluding the others, in such a way that it seeks to avoid problems such as multicollinearity and that this subset fits as well as the model with all variables.

One variable selection procedure that can be used is stepwise, so that the logratio that best explains the maximum likelihood is selected first and fixed. Then the second best logratio is chosen and so on, until there are no more explanatory variables to be tested.

After establishing a set of explanatory variables that meaningfully explains a compositional data set, individual responses can be investigated to isolate those that are best explained or possible outliers.

The quality of the adjusted model can be verified by comparing the observed values and those predicted for the response variable.

When choosing a particular model, if on the one hand, we must try to include as many independent variables as possible to improve the forecast, on the other hand, we want to include a minimum number of variables due to cost and simplicity issues (Oliveira, 2008).

According to Draper and Smith (1998), selection of the best model is defined as the commitment to reconciling these two objectives (incorporating a certain number of variables that can improve the predictability of the model, at the same time, discarding variables that are not significant as a way of simplifying the model and reduce costs). This selection involves a dose of subjectivity and the result may be different if the procedure used for selection is changed.

2.8. Multivariate Outliers

Outliers are usually defined as data that differ drastically from the rest and their identification plays an important role in statistical analysis, as such observations can contain important information in relation to the study hypotheses. If statistical models are applied and contain data with outliers, the results can be misleading and wrong decisions can be made. (Barbosa, et. al, 2018).

First: In univariate cases, all observations that are "quite far away" from the majority of the data and that may, potentially, not follow the same model, could be outliers. With regard to graphical methodology, the most common are Q-Q plot (graphical evaluation of the fit of a given model to the data) and box-plot (visualization of data distribution).

Second: For the bivariate case, it is possible to highlight a scatter diagram together with a confidence ellipse and the bag-plot (generalization of the box-plot for bivariate data).

i) scatter diagram and confidence ellipse: Construct the scatter diagram, and then and on the same graph, the ellipse with 100(1 – α) % confidence based on the Hoteling T^2 statistic given in expression (4) by:

$$T^2 = \left(\bar{X} - \mu\right)^T S^{-1}\left(\bar{X} - \mu\right) = \frac{n(n-1)}{n(n-2)} F_{1-\alpha;2;n-2} \qquad (4)$$

where $F_{1-\alpha;2;n-2}$ is the 1 – α percentile of the *F* distribution with *2* degrees of freedom in the numerator and *n – 2* in the denominator. A given point will be considered an outlier when it is outside the 100(1 – α) % confidence region.

ii) bag-plot: generalization of the box-plot for bivariate data. (Rousseeuw, Ruts and Tukey, 1999). The interpretability, mainly visual, is considerably different from a traditional boxplot described previously. This chart is also based on the concept of depth contours. Thus, the depth contour *k* contains the observations that have a location depth greater than or equal to *k*; consists of three concentric polygons called pocket, fence and cycle. The bag is the innermost polygon and contains 50%, the fence is the polygonal line, the outer polygon that separates the points that are not outliers from the outliers and the cycle is the region that contains the outer points considered as outliers, but within the fence, and; allows you to visualize its location (median depth), dispersion (bag size), correlation (bag concentration) and outliers.

Third: In the multivariate case, detection through graphs becomes a little more complex because the analysis would have to be done observing two variables at a time, which would make the process long and unreliable, as one point may be an outlier in relation to some variables and not in relation to others, which would cause the result to be masked (Giroldo, 2008).

Compositional data is multivariate data that represents positive quantitative descriptions of the parts of a whole (proportions). In particular, they study and apply statistical techniques based on log-ratio transformations and graphical techniques on the transformed data in the detection of outlier observations which correspond to multivariate observations that, for some reason, differ from the others. In a situation like this, a proposal for detecting multivariate outliers is the use of Mahalanobis Robust Distance (MRD) and robust biplots, and considering the use of transformations such as *alr*, *ilr* and *clr*.

Among the diverse and different methods for detecting multivariate outliers, we highlight:

i) Mahalanobis distance: For each of the n samples and *p* variables, the Mahalanobis distance (D_i) is calculated by expression (5):

$$D_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})} \qquad (5)$$

for *i = 1*, …, *n*, on what $S = \sum_{i=1}^{n} (x_i - \bar{x})' (x_i - \bar{x})$ is the sample variance-covariance matrix; it is, $(x_i - \bar{x})$ is the difference vector between the concentrations measured in one group and the average of the concentrations in the other group. Each of these values is compared to the critical value that can be calculated using the Wilks lambda criterion, defined in expression (6) by:

$$\frac{p(n-1)^2 F_{p,n-1,\alpha/n}}{n\left(n-p-1+pF_{p,n-1,\alpha/n}\right)} \qquad (6)$$

where, α is the significance level; *p*, is the number of variables; *n*, is the number of municipalities; $F_{p,n-1,\alpha/n}$, is the value of the *F* statistic for *p* degrees of freedom in the numerator and *n*-1 degrees of freedom in the denominator under a significance level α/n.

When the value found by expression (5) is greater than the critical value calculated by expression (6), the sample is considered discrepant or outlier (Oliveira and Munita, 2003).

This method is more appropriate for correlated and heteroscedastic variables. In many experimental situations, this type of quadratic distance has an intuitive appeal as it contemplates the covariance structure between different variables.

The Mahalanobis distance and the leverage statistic 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 as it has a greater influence on the slope or coefficients of the regression equation. 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 can be used as a comparison technique regarding the separation between different groups, allowing to evaluate the extent and direction of differences between the average values of the variables used in discrimination. The differences between each pair of groups being compared are thus examined simultaneously through several variables, which can be correlated, so that the information provided by one of them may not be independent of that provided by the others.

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

ii) Comedian Approach: Comedian Approach is a method for detecting multivariate atypical observations that uses comedian as an alternative measure of dependence between two random variables.

The comedian obeys the following properties: symmetry, location invariant and scale invariant.

The MAD estimator is characterized by being consistent with the population parameter and tends towards asymptotic normality.

According to (Sajesh and Srinivasan, 2013) there is a three-step procedure to obtain robust estimates for location and dispersion to be calculated considering the following steps:

Step 1: Calculate the eigenvalues, λ_j, and the eigenvectors, e_j, with $j \in \{1, ..., p\}$, of $\delta(X)$ and designate by *E* the matrix whose columns are the eigenvectors and by $\Lambda = \text{diag}(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, such that $\delta(X) = E\Lambda E^T$;

Step 2: Let *P* be the diagonal matrix $diag\left(\frac{1}{MAD(X_1)},\cdots,\frac{1}{MAD(X_p)}\right)$. It is then calculated: $Q = D^{-1}E$ and $z_i = Q^{-1}x_i$ with $i \in \{1, ..., n\}$ where MAD represents the mean absolute deviation;

Step 3: The robust estimates for the location m(X), and for the dispersion S(X), are given by: $m(X) = QI$ (7) and $S(X) = Q\Gamma Q^T$ (8)

where Γ is a diagonal matrix, whose entries are given by $(\text{MAD }(z_j))^2$; $I = (\text{med}(z_1), ..., \text{med}(z_p)$, and; Z_j is the *j-th* column *Q*.

Once the robust estimators for location and dispersion are known, the interest is to verify whether, in fact, an observation is atypical (outlier) or not. One of the most used procedures is to resort to the robust Mahalanobis distance, defined as:

$$RD\left(x_i, m\right) = rd_i = \left(x_i - m\right)^T S^{-1}\left(x_i - m\right),\ \text{i}=\{1,\cdots,n\} \qquad (9)$$

with m and S defined in equations (7) and (8) respectively.

To determine whether an observation is outlier, the next step is to define a cutoff value to determine potential outliers. The cutoff value is given by:

$$cv = 1,4826 \times \frac{\chi^2_{d;1-\alpha} \times med\left(rd_1,\cdots,rd_n\right)}{\chi^2_{d;0.5}} \qquad (10)$$

where α refers to the probability of a type I error that one is willing to commit, $\chi^2_{\text{p};1-\alpha}$ and $\chi^2_{\text{p};0.50}$ refer, respectively, to quantiles 1 - α and 0.5 of the chi-square distributions with *d* degrees of freedom and 1.4826 is a correction factor that corresponds to the inverse of the value of the 0.75 quantile of the standard Normal distribution, so that the MAD don't be biased.

Thus, an observation is considered an outlier if $RD\left(x_i, m\right) > cv$

By using the cutoff value defined in (10) for the robust Mahalanobis distance as per equation (9) it can be defined as a weight function and robust estimates for location and dispersion obtained. These estimates for location and dispersion, obtained by the comedian, have a high breaking point, which helps in detecting outliers.

Note that these robust estimates use the vector of means and the population covariance matrix and then calculate the Mahalanobis distance using robust estimates.

These techniques were used for the sets considering all variables, variables selected by procrustes and stepwise.

3 Results and Discussion

For this work, using data sets from the 2010 Demographic Census and the last UNDP aggregated into municipalities in their compositional form, the following analysis steps were carried out:

Step 1: Transform this data set by calculating Neperian logarithm values;

Step 2: Application of the Procrustes procedure to select the variables that best maintain the original structure of the data;

Step 3: Detection of discrepant data by two-dimensional graphical procedures such as confidence ellipses for the first two components for the compositional variables transformed by main log Neperian and bag-plot considering all variables and also only those selected by the procrustes procedure.

Step 4: From proportional data, data detection by multivariate methods such as Mahalanobis distance, robust Mahalanobis distance and comedian approach.

Step 5: Considering the proportional data transformed by Neperian logarithm, multivariate regression adjustments were made, considering income and poverty proportions as response variables and variables related to disability, education level, identification, work, housing and other assets as explanatory variables. Figure 5 shows the scatter diagram, box plot for each component, confidence ellipse and bag-plot for the first two components of the 254 variables considered in the study.

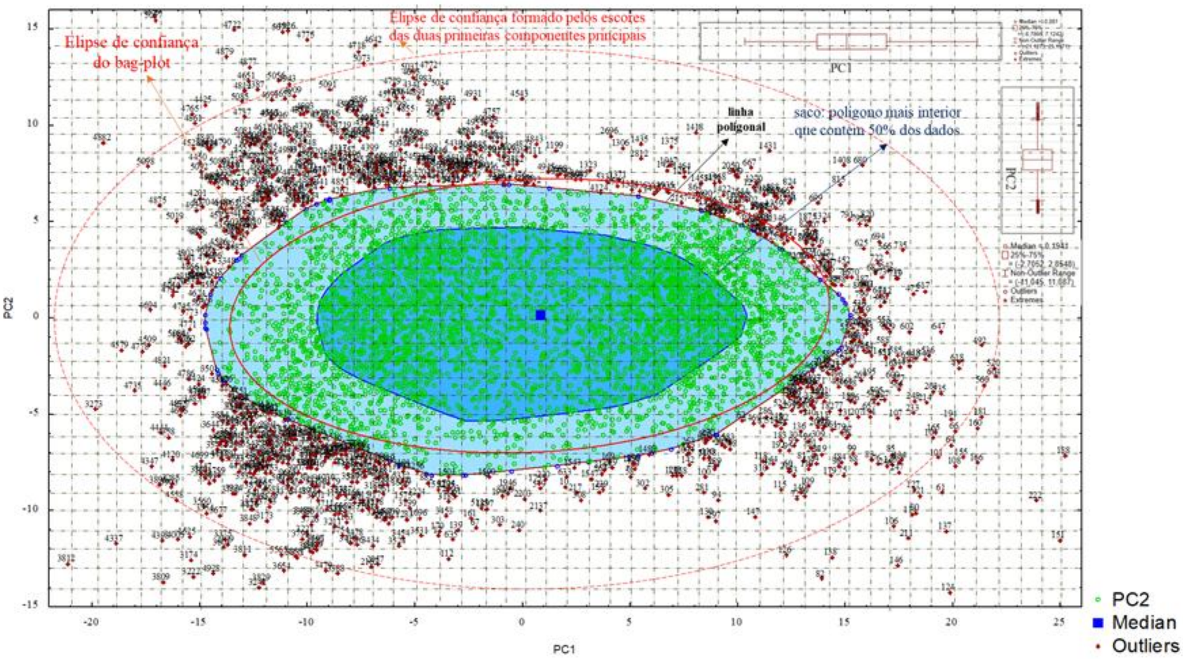


Figure 5. Scatter diagram, bag-plot, confidence ellipse and boxplot of the first two components for the 254 variables.

Analyzing Figure 5, it is possible to verify that all points located outside the confidence ellipse and outside the bag-plot are considered outliers of both, on the other hand, points located inside the confidence ellipse and outside the bag-plot are considered outliers. just for the bag-plot. According to Maltez (2020), it is estimated that approximately 50% of the points are considered outliers and for univariate box-plots it is noted that there are outliers for the second component and non-existence for the first component.

Under these conditions, the method that can detect a greater number of municipalities considered outliers was the bag-plot procedure with an estimate that can reach 50%, in the case of this work around 20%.

Table 1 below shows the results of the procrustes procedure carried out on the 254 variables, selecting through this process 139 variables with their respective values of M^2 and CV and the variables that obtained $M^2 > CV$ were selected in brown together with the proportions of the variables of the municipalities that were considered discrepant by MD and MRD before and after application of the procrustes procedure.

With the application of the procrustes procedure, the results in Table 1 were obtained, which contains the list of variables that are in brown, which were included after this process and had M^2 values greater than those of CV, and this started to happen with the age proportion variable. up to 15 years old (pID1) with M^2 worth 18006.33727 and CV less than M^2 worth 17735.65993.

When analyzing Figure 6, it was possible to verify that after applying the procrustes procedure, 254 to 139 variables were reduced and that the dispersion diagram was considered quite similar when compared with that in Figure 5.

Table 1 shows the results of the procrustes procedure carried out on the 254 variables, selecting through this process 139 variables with their respective values of M^2 and CV and the variables that obtained $M^2 > CV$ were selected in brown as well as the municipalities considered most discrepant before and after application of the Procrustes procedure for MD and MRD. The main municipalities considered outliers were Huramutâ, Guajará and Porto Barreiro.

Table 2 shows the list of municipalities considered outliers together with their respective distance values from Mahalanobis in a total of 149 municipalities; ditto Table 3 for Mahalanobis Robust Distance in a total of 118 municipalities; Table 4 after procrustes procedure for Mahalanobis Distance with 101 municipalities considered discrepant, and, finally; for Table 5, Mahalanobis Distance is robust in a total of 49 municipalities considered outliers.

The next table is Table 6 which shows the discrepant results for second and third principal components (PC2 and PC3), Mahalanobis distance (MD), Mahalanobis Robusta distance (MRD) and bag-plot.

Next, including only the variables selected by Procrustes, Figure 6 shows the same graph as Figure 5 but considering only the 139 variables that were obtained by the Procrustes procedure and it should also be noted that the graph is close to the previous graph less by one horizontal rotation.

When analyzing the results in Table 6, it was possible to verify that when including all variables, 254 were considered and that after applying the procrustes procedure, they were reduced to 139 variables and with regard to the detection of multivariate outliers, when carrying out a comparative study between the inclusion of all variables and including only the variables selected by the procrustes procedure, the most stable was when the bag-plot graphical method was applied, in which there was a variation of 1.08% followed by Robust Mahalanobis Distance (MRD) with a variation of 4.27% and a greater decrease in the detection of outliers were obtained: confidence ellipse with a variation of 40.91% followed by Mahalanobis Distance with a variation of 32.21%.

Next, the regression adjustment was made considering the proportions of income and poverty as response variables and compositional variables related to identification, level of education, work, housing and other assets as independent variables. For the dependent variable income and poverty, note that ram1 means people with income between 0 and 0.125 minimum wages (below the poverty line representing 34.9% of the population); ram2, income between 0.125 and 0.25 minimum wages (at the poverty line with 3.4%); ram3, between 0.25 and 0.5 minimum wages (above the poverty line with 3.8%); ram4, between 0.5 and one minimum wage (class E with 17.8%); ram5, between one and 3 minimum wages (class D with 21.8%); ram6, between 3 and 7 minimum wages (class C with 5.6%); ram7, between 7 and 20 minimum wages (Class B with 1.76%), and, finally; above 20 minimum wages (class A with 0.8%).

Table 1. Procrustes procedure

Variables	ES1	def12	DMC3	DCC6	DCC2	DL3	MP4	CF6	def10	AA1	CPR1	NF4	TTS3	NIA6	def7	EE2	EE1	DL2	TGCT1	pREG5	pNTT4	CF4
M2	3,47	8,90	17,19	32,52	44,36	62,59	81,72	102,32	124,29	139,90	156,42	183,11	208,31	237,29	266,98	295,60	326,40	360,98	398,40	436,62	475,04	516,04
CV	23047,12	23000,93	22954,75	22908,56	22862,37	22816,19	22770,00	22723,81	22677,63	22631,44	22585,25	22539,07	22492,88	22446,69	22400,51	22354,32	22308,13	22261,95	22215,76	22169,57	22123,39	22077,20
UIRAMUTÃ	0,007	0,012	0,727	0,009	0,000	0,648	0,167	0,003	0,000	0,237	0,172	0,191	0,283	0,002	0,021	0,031	0,248	0,003	0,198	0,013	0,366	0,014
GUAJARÁ	0,001	0,007	0,116	0,003	0,000	0,398	0,000	0,043	0,000	0,240	0,071	0,194	0,209	0,014	0,019	0,039	0,639	0,227	0,285	0,000	0,484	0,038
PORTO BARREIRO	0,003	0,015	0,012	0,006	0,080	0,542	0,000	0,038	0,001	0,945	0,483	0,141	0,427	0,023	0,037	0,002	0,895	0,009	0,298	0,000	0,211	0,046
Variables	def4	EUV4	NF1	SB1	VA3	FAA10	NA2	CF9	RA1	TE8	ES4	NDO1	QT1	REG4	RA3	TE13	OG2	OG1	CF7	EUV1	pTE3	ES6
M2	560,25	598,63	644,80	686,96	734,74	786,88	848,13	909,60	976,77	1041,66	1104,25	1184,67	1241,11	1317,08	1382,39	1415,54	1496,98	1578,76	1631,71	1712,76	1802,53	1879,53
CV	22031,02	21984,83	21938,64	21892,46	21846,27	21800,08	21753,90	21707,71	21661,52	21615,34	21569,15	21522,96	21476,78	21430,59	21384,40	21338,22	21292,03	21245,84	21199,66	21153,47	21107,28	21061,10
UIRAMUTÃ	0,021	0,000	0,540	0,297	0,000	0,000	0,241	0,019	0,008	0,000	0,183	0,991	0,883	0,033	0,003	0,000	1,000	0,000	0,125	1,000	0,000	0,065
GUAJARÁ	0,011	0,000	0,461	0,952	0,000	0,000	0,000	0,033	0,164	0,001	0,102	0,892	0,910	0,025	0,001	0,000	0,854	0,146	0,084	0,991	0,000	0,580
PORTO BARREIRO	0,030	0,002	0,379	0,896	0,000	0,000	0,005	0,087	0,744	0,000	0,012	0,928	0,812	0,000	0,009	0,002	0,912	0,088	0,167	0,994	0,000	0,024
Variables	def6	NIA3	NF2	PC1	def1	CF2	ES3	ID2	def3	MP1	PC2	DMC2	CO2	TE11	def9	NMD3	MOTO1	TT7	REG3	CF3	def5	DCC1
M2	1979,31	2084,72	2183,80	2218,34	2328,87	2445,32	2563,13	2675,77	2777,21	2899,69	3031,21	3072,81	3181,45	3319,29	3462,60	3619,27	3764,83	3925,30	4083,14	4238,51	4391,28	4549,92
CV	21014,91	20968,72	20922,54	20876,35	20830,16	20783,98	20737,79	20691,60	20645,42	20599,23	20553,04	20506,86	20460,67	20414,48	20368,30	20322,11	20275,92	20229,74	20183,55	20137,36	20091,18	20044,99
UIRAMUTÃ	0,009	0,107	0,144	0,029	0,844	0,174	0,596	0,435	0,020	0,077	0,971	0,208	0,014	0,000	0,004	0,882	0,083	0,686	0,011	0,108	0,002	0,962
GUAJARÁ	0,009	0,093	0,209	0,093	0,847	0,084	0,240	0,512	0,010	0,086	0,907	0,532	0,035	0,000	0,003	0,648	0,132	0,409	0,000	0,103	0,002	0,951
PORTO BARREIRO	0,015	0,151	0,343	0,166	0,781	0,064	0,883	0,660	0,015	0,175	0,834	0,145	0,022	0,000	0,004	0,216	0,272	0,135	0,000	0,069	0,005	0,743
Variables	TE2	NIA4	NA3	VC3	EC4	QPT1	TT5	TE6	FAA1	QV5	QPT2	NB3	NA1	MP2	TE5	NDO5	NIA5	AA2	CF10	DL4	QV2	FAA2
M2	4706,59	4883,00	5063,40	5242,46	5410,32	5530,83	5621,82	5745,60	5944,09	6146,20	6343,40	6543,90	6761,29	6968,33	7194,92	7413,05	7644,52	7886,91	8133,22	8386,12	8663,55	8948,72
CV	19998,80	19952,62	19906,43	19860,24	19814,06	19767,87	19721,68	19675,50	19629,31	19583,12	19536,94	19490,75	19444,56	19398,38	19352,19	19306,00	19259,82	19213,63	19167,44	19121,26	19075,07	19028,89
UIRAMUTÃ	0,004	0,047	0,000	0,403	0,019	1,000	0,001	0,000	0,413	0,000	0,000	0,000	0,759	0,190	0,103	0,000	0,033	0,226	0,004	0,074	0,736	0,026
GUAJARÁ	0,000	0,054	0,003	0,444	0,015	0,667	0,005	0,006	0,305	0,007	0,333	0,022	0,997	0,008	0,000	0,005	0,074	0,208	0,003	0,023	0,431	0,289
PORTO BARREIRO	0,000	0,097	0,000	0,327	0,059	1,000	0,001	0,004	0,333	0,004	0,000	0,008	0,995	0,042	0,000	0,000	0,112	0,027	0,000	0,129	0,045	0,498
Variables	RT2	CF12	REG2	QV3	VC2	VA4	FAA4	MP6	EUV3	NB2	NB4	RAM4	QV1	TT2	TTS1	def2	ES2	RA5	FAA3	def8	CPR2	TT1
M2	9244,06	9522,47	9545,96	9798,86	10068,33	10384,67	10708,06	10982,75	11282,14	11581,65	11888,08	12231,48	12576,60	12690,56	12750,30	13026,51	13380,52	13722,42	14091,16	14471,34	14866,05	15275,35
CV	18982,70	18936,51	18890,33	18844,14	18797,95	18751,77	18705,58	18659,39	18613,21	18567,02	18520,83	18474,65	18428,46	18382,27	18336,09	18289,90	18243,71	18197,53	18151,34	18105,15	18058,97	18012,78
UIRAMUTÃ	0,052	0,000	0,031	0,173	0,050	0,000	0,000	0,006	0,000	0,070	0,000	0,051	0,057	0,014	0,000	0,061	0,141	0,878	0,126	0,000	0,014	0,051
GUAJARÁ	0,107	0,000	0,004	0,450	0,069	0,000	0,000	0,126	0,009	0,142	0,004	0,090	0,017	0,083	0,010	0,087	0,074	0,001	0,066	0,001	0,000	0,103
PORTO BARREIRO	0,140	0,000	0,009	0,665	0,097	0,000	0,000	0,042	0,004	0,090	0,000	0,193	0,006	0,001	0,000	0,089	0,078	0,001	0,169	0,001	0,035	0,137
Variables	ES5	NTT6	NU2	DMC1	NBens0	ID1	NC5	IDHM_L	QV4	TTS4	CF5	TE7	NTT5	NTT7	VA5	SB2	FAA7	CPR3	TV2	VC1	pNDO2	MEE2
M2	15709,99	16141,05	16599,57	17059,40	17525,89	18006,34	18458,92	18863,37	19281,07	19753,77	20250,10	20755,48	21287,80	21839,13	22397,38	22946,34	23517,04	24110,05	24732,14	25372,45	26011,08	26675,11
CV	17966,59	17920,41	17874,22	17828,03	17781,85	17735,66	17689,47	17643,29	17597,10	17550,91	17504,73	17458,54	17412,35	17366,17	17319,98	17273,79	17227,61	17181,42	17135,23	17089,05	17042,86	16996,67
UIRAMUTÃ	0,008	0,004	0,069	0,065	0,680	0,537	0,003	0,766	0,033	0,079	0,521	0,000	0,174	0,249	0,000	0,703	0,367	0,814	0,810	0,547	0,007	0,022
GUAJARÁ	0,004	0,008	0,104	0,352	0,168	0,463	0,041	0,762	0,095	0,175	0,571	0,002	0,175	0,019	0,000	0,048	0,329	0,929	0,395	0,487	0,092	0,032
PORTO BARREIRO	0,000	0,001	0,049	0,843	0,016	0,276	0,142	0,821	0,281	0,004	0,474	0,000	0,575	0,052	0,000	0,104	0,000	0,483	0,140	0,575	0,063	0,213
Variables	GEL1	RT1	TE4	NMD1	NTT2	TGCT3	SO1	NTT3	QT2	TGCT4	TE9	MP8	TLF1	ML1	DL5	CF1	TE1	MP9	NC4	TT4	]MP7	AA3
M2	27361,55	28062,60	28789,52	29529,48	30310,13	31111,62	31620,74	32246,96	32897,75	33722,82	34578,63	35383,45	36249,28	37151,54	38110,10	39097,06	40025,71	41012,57	42000,02	43034,95	44115,24	45206,95
CV	16950,49	16904,30	16858,11	16811,93	16765,74	16719,55	16673,37	16627,18	16580,99	16534,81	16488,62	16442,43	16396,25	16350,06	16303,87	16257,69	16211,50	16165,32	16119,13	16072,94	16026,76	15980,57
UIRAMUTÃ	0,157	0,948	0,000	0,018	0,004	0,169	0,529	0,042	0,117	0,061	0,000	0,005	0,016	0,049	0,149	0,016	0,893	0,002	0,012	0,055	0,013	0,537
GUAJARÁ	0,622	0,893	0,000	0,044	0,005	0,061	0,337	0,135	0,090	0,014	0,000	0,027	0,031	0,157	0,099	0,021	0,991	0,000	0,112	0,103	0,001	0,552
PORTO BARREIRO	0,834	0,860	0,000	0,188	0,000	0,158	0,751	0,001	0,188	0,042	0,000	0,027	0,045	0,165	0,009	0,000	0,994	0,000	0,287	0,497	0,000	0,028
Variables	FAA6	DCC5	RD1	NC3	RAM3	RA4	NIA9	ML2	TGCT5	EC3	DCC3	NTT1	NIA7	pNBens6	CAR1	IDHM_E	NU4	def15	EC1	pREG1	NC1	NBens2
M2	46339,37	47490,39	48612,05	49794,52	51007,18	52228,73	53465,99	54741,01	56042,33	57373,29	58734,36	59899,32	61043,25	62458,33	63933,91	65450,65	66939,43	68501,41	70071,26	71673,65	73259,19	74955,44
CV	15934,38	15888,20	15842,01	15795,82	15749,64	15703,45	15657,26	15611,08	15564,89	15518,70	15472,52	15426,33	15380,14	15333,96	15287,77	15241,58	15195,40	15149,21	15103,02	15056,84	15010,65	14964,46
UIRAMUTÃ	0,000	0,018	0,134	0,015	0,076	0,092	0,000	0,951	0,013	0,055	0,006	0,162	0,002	0,007	0,022	0,276	0,485	0,001	0,182	0,912	0,890	0,073
GUAJARÁ	0,003	0,017	0,648	0,224	0,056	0,793	0,000	0,843	0,009	0,003	0,015	0,175	0,002	0,034	0,031	0,387	0,629	0,000	0,112	0,972	0,279	0,153
PORTO BARREIRO	0,000	0,049	0,960	0,194	0,091	0,233	0,000	0,835	0,005	0,016	0,037	0,159	0,002	0,153	0,513	0,588	0,247	0,001	0,425	0,991	0,170	0,075
Variables	DL7	NB1	VA1	SE2	NIA1	CF8	def13	GEL2	EC2	TGCT2	NBens1	MEE1	NIA2	RD2	TT6	RA2	FAA9	SE1	DL1	NIA8	MP5	FAA5
M2	76724,73	78494,44	80309,17	82174,96	84099,53	86038,87	88002,70	90056,77	92168,94	94308,40	96500,29	98618,56	100867,99	103205,50	105621,33	107984,14	110489,23	113035,91	115603,92	118249,91	120930,04	123674,15
CV	14918,28	14872,09	14825,90	14779,72	14733,53	14687,34	14641,16	14594,97	14548,78	14502,60	14456,41	14410,22	14364,04	14317,85	14271,66	14225,48	14179,29	14133,10	14086,92	14040,73	13994,54	13948,36
UIRAMUTÃ	0,029	0,930	1,000	0,492	0,243	0,018	0,001	0,843	0,001	0,559	0,127	0,035	0,564	0,866	0,078	0,018	0,038	0,508	0,098	0,000	0,519	0,001
GUAJARÁ	0,001	0,832	1,000	0,478	0,278	0,019	0,000	0,378	0,001	0,631	0,193	0,944	0,483	0,352	0,011	0,042	0,000	0,522	0,211	0,000	0,000	0,000
PORTO BARREIRO	0,015	0,901	0,894	0,481	0,076	0,054	0,000	0,166	0,005	0,498	0,010	0,752	0,538	0,040	0,045	0,013	0,000	0,519	0,287	0,000	0,000	0,000
Variables	TTS5	NBens4	NU1	TE12	RAM5	DL6	NBens3	pdef16	SO2	Z1	DCC4	FAA8	NDO6	NMD2	VA2	TT3	RAM6	PCI1	NF3	NDO3	Z2	def14
M2	126405,76	129287,97	132232,86	135231,87	138405,53	141627,54	144954,63	148341,91	151379,54	154855,66	158421,96	162115,89	165879,49	169804,71	173572,23	177388,18	181491,02	185790,73	190269,80	194876,12	199536,39	204377,14
CV	13902,17	13855,98	13809,80	13763,61	13717,42	13671,24	13625,05	13578,86	13532,68	13486,49	13440,30	13394,12	13347,93	13301,74	13255,56	13209,37	13163,19	13117,00	13070,81	13024,63	12978,44	12932,25
UIRAMUTÃ	0,599	0,040	0,066	0,000	0,071	0,000	0,042	0,001	0,471	0,135	0,005	0,028	0,000	0,100	0,000	0,115	0,014	0,067	0,125	0,002	0,865	0,003
GUAJARÁ	0,510	0,134	0,086	0,000	0,089	0,041	0,175	0,002	0,663	0,556	0,014	0,008	0,000	0,307	0,000	0,286	0,019	0,737	0,136	0,007	0,444	0,001
PORTO BARREIRO	0,502	0,249	0,622	0,000	0,268	0,009	0,137	0,002	0,249	0,188	0,084	0,000	0,000	0,596	0,106	0,183	0,058	0,416	0,137	0,009	0,812	0,003
Variables	EA2	def11	NDO4	NBens8	pNU3	RAM8	TE10	TV1	TLF2	EE3	EC5	EA1	NBens5	NBens9	RAM7	MEE3	CO1	NIA10	MP3	CF11	IDHM	TTS2
M2	209335,69	214403,39	219603,46	224944,57	230359,09	236045,62	241927,11	247977,12	254302,40	260952,67	267817,24	274857,78	282437,92	290400,66	298583,66	306864,82	315785,78	325301,75	333517,98	343629,80	354155,80	365152,65
CV	12886,07	12839,88	12793,69	12747,51	12701,32	12655,13	12608,95	12562,76	12516,57	12470,39	12424,20	12378,01	12331,83	12285,64	12239,45	12193,27	12147,08	12100,89	12054,71	12008,52	11962,33	11916,15
UIRAMUTÃ	0,464	0,001	0,000	0,001	0,380	0,003	0,000	0,190	0,984	0,721	0,744	0,536	0,021	0,000	0,004	0,943	0,986	0,002	0,021	0,000	0,453	0,039
GUAJARÁ	0,651	0,001	0,004	0,011	0,181	0,003	0,000	0,605	0,969	0,322	0,869	0,349	0,087	0,019	0,005	0,023	0,965	0,002	0,752	0,000	0,532	0,096
PORTO BARREIRO	0,232	0,002	0,000	0,025	0,082	0,006	0,000	0,860	0,955	0,104	0,496	0,768	0,229	0,010	0,005	0,036	0,978	0,002	0,713	0,000	0,688	0,067
Variables	PCI2	NBens7	RAM1	MOTO2	RAM2	NC2	CEL2	ID3	IDHM_R	CEL1	CAR2	NBens10										
M2	376867,50	389219,84	402059,51	416864,99	433657,16	452328,57	473127,59	500406,84														
CV	11869,96	11823,77	11777,59	11731,40	11685,21	11639,03	11592,84	11546,65														
UIRAMUTÃ	0,933	0,010	0,683	0,917	0,066	0,080	0,902	0,028	0,439	0,098	0,978	0,000										
GUAJARÁ	0,263	0,021	0,640	0,868	0,063	0,344	0,707	0,024	0,510	0,293	0,969	0,005										
PORTO BARREIRO	0,584	0,089	0,288	0,728	0,057	0,208	0,244	0,065	0,676	0,756	0,487	0,008										

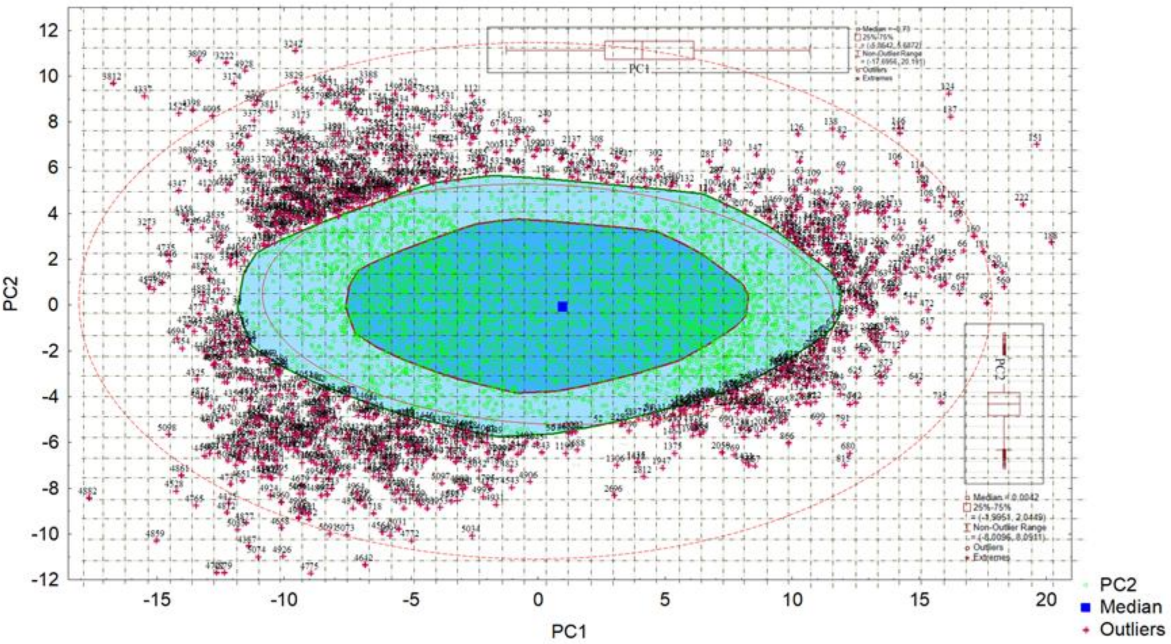


Figure 6. Scatter diagram, bag-plot, confidence ellipse and boxplot of the first two components for the 139 variables obtained from the procrustes procedure.

Table 2. List of municipalities considered outliers including the 254 variables and applying MD

City	MD			City	MD	City	MD
JORDÃO	413,8614504	FEIRA NOVA DO MARANHÃO	406,8675068	VITÓRIA	370,5379685	SÃO JOSÉ DO RIO PRETO	375,4631533
MARECHAL THAUMATURGO	424,2495712	FERNANDO FALCÃO	503,7832865	NITERÓI	366,0737008	SÃO JOSÉ DOS CAMPOS	376,0583327
PORTO WALTER	435,3169272	HUMBERTO DE CAMPOS	395,8626587	ÁGUAS DE SÃO PEDRO	472,9158934	SOROCABA	362,546917
ATALAIA DO NORTE	456,3134069	ICATU	376,7705056	AMERICANA	496,4666713	VALINHOS	477,9912773
IPIXUNA	422,4039648	ITAIPAVA DO GRAJAÚ	381,2282156	ARARAQUARA	395,2863878	VINHEDO	482,1085826
ITAMARATI	418,7042952	JENIPAPO DOS VIEIRAS	422,2015607	ARARAS	361,2961837	CURITIBA	474,9424826
JURUÁ	359,25481	MARAJÁ DO SENA	540,8656924	BARIRI	358,9476264	MARINGÁ	416,6553034
JUTAÍ	422,097575	MATÕES DO NORTE	361,4025477	BARRA BONITA	399,4630986	ASCURRA	380,8623502
MARAÃ	399,3764593	MILAGRES DO MARANHÃO	392,1821284	BAURU	359,6718202	BALNEÁRIO CAMBORIÚ	427,9726401
PAUINI	417,0691013	MORROS	414,2291852	CAMPINAS	408,0207187	BLUMENAU	477,7450562
SANTA ISABEL DO RIO NEGRO	474,8158132	PAULINO NEVES	446,4321264	CATANDUVA	366,0820087	BRUSQUE	481,716554
SÃO PAULO DE OLIVENÇA	431,2939643	PEDRO DO ROSÁRIO	419,4251406	CERQUILHO	392,303632	CRICIÚMA	365,414377
AMAJARI	489,2260984	PRESIDENTE JUSCELINO	413,1180116	ILHA SOLTEIRA	369,0345849	FLORIANÓPOLIS	433,699041
NORMANDIA	402,8608037	PRESIDENTE SARNEY	415,5433783	INDAIATUBA	426,8862923	GASPAR	374,9470241
UIRAMUTÃ	668,6018167	PRESIDENTE VARGAS	380,7827142	IRACEMÁPOLIS	380,540304	INDAIAL	408,2476488
ACARÁ	387,6293627	PRIMEIRA CRUZ	507,3197229	ITATIBA	365,8840829	JARAGUÁ DO SUL	459,2465584
AFUÁ	432,6754489	SANTANA DO MARANHÃO	403,3876818	JAGUARIÚNA	372,2600918	JOAÇABA	409,879415
ANAJÁS	496,6299368	SANTO AMARO DO MARANHÃO	527,853471	JAÚ	394,9649734	JOINVILLE	402,8407105
AVEIRO	462,5464356	SÃO FÉLIX DE BALSAS	364,2441503	JUNDIAÍ	447,5516807	POMERODE	443,2350424
BAGRE	386,1853994	SERRANO DO MARANHÃO	374,9731845	LENÇÓIS PAULISTA	384,5987619	RIO DO SUL	392,0059513
CACHOEIRA DO PIRIÁ	564,3394854	TURIAÇU	370,2341797	LIMEIRA	363,2906774	SÃO JOSÉ	456,5602098
CHAVES	621,9159259	BETÂNIA DO PIAUÍ	360,1960899	MOGI GUAÇU	362,875081	TIMBÓ	431,3787908
CURRALINHO	423,530483	CAMPO LARGO DO PIAUÍ	444,7314177	MOJI MIRIM	359,0837661	TUBARÃO	398,703071
GARRAFÃO DO NORTE	408,8493914	CURRAIS	376,3911763	MONTE ALTO	373,8992222	BENTO GONÇALVES	398,5238538
GURUPÁ	428,3123527	CURRAL NOVO DO PIAUÍ	397,1318759	NOVA ODESSA	397,9331544	CARLOS BARBOSA	421,3561861
LIMOEIRO DO AJURU	410,3841594	DOM INOCÊNCIO	416,6065848	PAULÍNIA	415,7647099	CAXIAS DO SUL	418,3712509
MELGAÇO	589,9749434	JÚLIO BORGES	365,2381913	PIRACICABA	393,8290721	DOIS IRMÃOS	456,7979288
NOVA ESPERANÇA DO PIRIÁ	425,4616761	LAGOA DO BARRO DO PIAUÍ	363,1385946	PIRASSUNUNGA	376,9252489	FARROUPILHA	405,9559313
OEIRAS DO PARÁ	368,6455987	MASSAPÊ DO PIAUÍ	362,0974473	RIBEIRÃO PRETO	423,3516913	FELIZ	361,0287245
PLACAS	380,8163497	MORRO CABEÇA NO TEMPO	361,0588357	RIO CLARO	406,0486778	GARIBALDI	461,4719286
PORTO DE MOZ	363,2230114	MURICI DOS PORTELAS	379,914425	SALTINHO	431,3461291	GUAPORÉ	369,4700979
PRAINHA	412,9988846	PAU D'ARCO DO PIAUÍ	367,241301	SANTA BÁRBARA D'OESTE	443,3588257	IVOTI	361,2540922
BELÁGUA	412,2137085	SEBASTIÃO BARROS	446,8630337	SANTO ANDRÉ	419,9475392	LAJEADO	364,3393616
BURITI	392,5678574	VÁRZEA BRANCA	369,4583617	SANTOS	453,8565624	NOVA PÁDUA	383,6826467
CACHOEIRA GRANDE	437,4552756	PEDRO ALEXANDRE	416,9690225	SÃO BERNARDO DO CAMPO	387,2232801	NOVA PRATA	358,9159888
CAJARI	397,5918074	Belo Horizonte	360,7397543	SÃO CAETANO DO SUL	565,6847003	PORTO ALEGRE	380,9608327
CENTRO NOVO DO MARANHÃO	374,3221729	Poços de Caldas	386,714226	SÃO CARLOS	432,2903533	VERANÓPOLIS	406,5482734
WESTFALIA	388,7324287						

Table 3. List of municipalities considered outliers including the 254 variables and applying MRD and comedian matrix

City	MRD	City	MRD	City	MRD	City	MRD
ASSIS BRASIL	178,8858194	MARAÃ	165,5252147	ALMEIRIM	177,6359782	RURÓPOLIS	179,2070332
FEIJÓ	173,2548661	MAUÉS	174,0237348	ANAJÁS	176,5406692	SANTA MARIA DAS BARREIRAS	178,1474898
JORDÃO	175,7858801	NHAMUNDÁ	175,1573396	ANAPU	176,696619	SENADOR JOSÉ PORFÍRIO	178,0223558
MANOEL URBANO	178,9995023	NOVA OLINDA DO NORTE	173,7002267	AURORA DO PARÁ	177,1492631	PEDRA BRANCA DO AMAPARI	171,2598284
MARECHAL THAUMATURGO	173,8686767	NOVO AIRÃO	179,7306358	AVEIRO	172,8976098	MAZAGÃO	175,8789814
SANTA ROSA DO PURUS	175,4437296	NOVO ARIPUANÃ	176,6647804	BAGRE	172,8583429	TARTARUGALZINHO	178,9219092
TARAUACÁ	176,2464472	PAUINI	173,4011143	BREVES	170,4897262	BARRA DO OURO	179,8456388
AMATURÁ	168,2621591	SANTA ISABEL DO RIO NEGRO	164,659708	CACHOEIRA DO ARARI	175,5084374	CAMPOS LINDOS	179,3399743
ANAMÃ	177,0426749	SÃO GABRIEL DA CACHOEIRA	168,3532978	CAMETÁ	179,9025736	CENTENÁRIO	173,5883039
ATALAIA DO NORTE	165,2633307	SÃO PAULO DE OLIVENÇA	172,01375	CHAVES	177,2494714	CHAPADA DA NATIVIDADE	178,8426544
AUTAZES	170,1737123	SILVES	178,7837533	CUMARU DO NORTE	171,1094068	GOIATINS	174,6192619
BARCELOS	161,5213656	TABATINGA	177,969107	CURRALINHO	174,7985746	PALMEIRANTE	178,6088941
BARREIRINHA	171,6386405	TAPAUÁ	172,1023702	GURUPÁ	178,7820527	RECURSOLÂNDIA	170,2531577
BENJAMIN CONSTANT	175,4721896	TONANTINS	175,3869745	IGARAPÉ-MIRI	172,250112	CACHOEIRA GRANDE	178,4389627
BERURI	175,0119887	URUCURITUBA	178,2090994	IPIXUNA DO PARÁ	179,3000134	CENTRO NOVO DO MARANHÃO	172,3906829
BOA VISTA DO RAMOS	175,5253191	AMAJARI	144,2434221	JACAREACANGA	157,7662533	HUMBERTO DE CAMPOS	178,7526901
BORBA	175,0703335	ALTO ALEGRE	155,0094545	JURUTI	172,0006927	MARAJÁ DO SENA	179,6852942
CAAPIRANGA	178,0619686	BONFIM	165,4375323	LIMOEIRO DO AJURU	179,342952	PAULINO NEVES	177,8583691
COARI	177,8865684	CANTÁ	172,8034687	MELGAÇO	174,160698	PRESIDENTE JUSCELINO	177,2593691
EIRUNEPÉ	179,2863279	CARACARAÍ	179,8493358	MOJU	177,3513011	PRIMEIRA CRUZ	174,483686
ENVIRA	177,4447878	CAROEBE	175,8689839	MUANÁ	177,0140936	SANTO AMARO DO MARANHÃO	173,8283109
FONTE BOA	175,8609739	IRACEMA	159,3850991	NOVO REPARTIMENTO	180,1568965	CURRAL NOVO DO PIAUÍ	179,9741473
GUAJARÁ	180,3086942	NORMANDIA	158,6969538	ÓBIDOS	178,7922891	MORRO CABEÇA NO TEMPO	174,7765954
IPIXUNA	176,6016004	PACARAIMA	171,4861	OEIRAS DO PARÁ	179,575677	PARNAGUÁ	179,3107408
JAPURÁ	177,185122	RORAINÓPOLIS	180,2454057	ORIXIMINÁ	179,7451551	SANTA FILOMENA	177,3976323
JURUÁ	174,1066317	UIRAMUTÃ	163,6395943	PACAJÁ	176,8320125	SÃO JOÃO DA SERRA	176,9806516
JUTAÍ	167,4911209	ACARÁ	175,4250031	PORTEL	172,624438	PILÃO ARCADO	179,9859668
LÁBREA	174,8973034	AFUÁ	170,4391701	PORTO DE MOZ	172,6669244	JAPORÃ	166,552874
MANICORÉ	178,3247103	ALENQUER	177,098227	PRAINHA	169,477174	PARANHOS	178,1349069
CAMPINÁPOLIS	175,2079955	NOVA NAZARÉ	179,3199836				

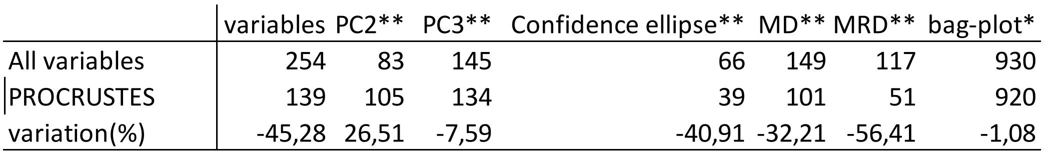
Table 4. List of municipalities considered outliers including the 139 variables after the procrustes procedure

City	MD	City	MD	City	MD		
JORDÃO	197,0588661	ITAIPAVA DO GRAJAÚ	176,3420068	PAULÍNIA	172,0783623	TUBARÃO	176,9650156
MARECHAL THAUMATURGO	187,4910906	JENIPAPO DOS VIEIRAS	193,2560551	RIBEIRÃO PRETO	174,0727158	BENTO GON	179,5059656
PORTO WALTER	181,2353315	MARAJÁ DO SENA	240,4099988	SALTINHO	206,4836146	CARLOS BA	197,3592674
ATALAIA DO NORTE	198,342314	MILAGRES DO MARANHÃO	169,1254591	SANTA BÁRBARA D'OEST	182,4414636	CAXIAS DO	175,8711484
IPIXUNA	180,6595433	PAULINO NEVES	183,8512897	SANTO ANDRÉ	175,2713311	DOIS IRMÃO	185,4123144
ITAMARATI	183,0866793	PEDRO DO ROSÁRIO	169,0538626	SANTOS	177,5938136	FARROUPIL	181,7522032
JURUÁ	169,2282356	PRESIDENTE JUSCELINO	186,113448	SÃO CAETANO DO SUL	231,6361818	FELIZ	175,1941837
JUTAÍ	174,0368307	PRESIDENTE SARNEY	180,9023408	SÃO CARLOS	173,890566	FLORES DA	169,8097868
MARAÃ	178,1617478	PRIMEIRA CRUZ	220,9958207	VALINHOS	206,0223644	GARIBALDI	219,2524321
PAUINI	171,6673661	SANTANA DO MARANHÃO	200,5404379	VINHEDO	207,2066676	GRAMADO	169,7395513
SANTA ISABEL DO RIO NEGRO	188,2060914	SANTO AMARO DO MARANHÃO	206,1926418	CURITIBA	195,9898028	IVOTI	169,9702892
SÃO PAULO DE OLIVENÇA	167,6955381	SÃO ROBERTO	168,5028184	MARINGÁ	180,8425396	NOVA PÁDU	218,2078023
AMAJARI	233,0844845	SERRANO DO MARANHÃO	171,0593844	ANTÔNIO CARLOS	178,4117074	NOVA PETR	175,1864706
NORMANDIA	169,017749	CAMPO LARGO DO PIAUÍ	184,7138368	BALNEÁRIO CAMBORIÚ	174,0519105	NOVA PRAT	173,3976303
UIRAMUTÃ	286,3188554	CARAÚBAS DO PIAUÍ	176,6415698	BLUMENAU	222,503495	VALE REAL	172,7582837
ANAJÁS	200,1086254	CURRAL NOVO DO PIAUÍ	171,3371901	BRUSQUE	225,1431853	VERANÓPOL	177,5666742
AVEIRO	187,8473877	DOM INOCÊNCIO	177,0278257	COCAL DO SUL	168,2062824	WESTFALIA	170,5196897
CACHOEIRA DO PIRIÁ	239,2040364	GUARIBAS	168,4807964	FLORIANÓPOLIS	191,4281682	FEIRA NOVA	186,4231896
CHAVES	264,1607256	MURICI DOS PORTELAS	184,1373572	GASPAR	182,1980864	FERNANDO	204,8009906
GARRAFÃO DO NORTE	168,4477695	SEBASTIÃO BARROS	202,1706247	GUABIRUBA	175,1918695	JUNDIAÍ	183,2672875
GURUPÁ	172,2738002	PEDRO ALEXANDRE	178,8767243	INDAIAL	186,3095465	NOVA ODES	176,4420309
MELGAÇO	245,4415104	ÁGUAS DE SÃO PEDRO	204,5500743	JARAGUÁ DO SUL	208,6354641	SÃO JOSÉ	199,6284598
NOVA ESPERANÇA DO PIRIÁ	177,86154	AMERICANA	205,8561994	JOAÇABA	191,1321242	TIMBÓ	226,5495682
PRAINHA	172,2584854	CAMPINAS	170,8011959	JOINVILLE	190,6696512		
CACHOEIRA GRANDE	185,3643532	CERQUILHO	178,1316266	POMERODE	216,5269322		
CAJARI	168,2252851	INDAIATUBA	178,1970858	RIO DO SUL	186,2939876		

Table 5. List of municipalities considered outliers including the 139 variables after selection using the procrustes procedure, applying MRD and comedian matrix

City	MRD	City	MRD	City	MRD		
BARCELOS	33,66427257	CÂNDIDO DE ABREU	34,64319967	IPUAÇU	33,33731668	JAPORÃ	28,16207734
MAUÉS	34,60763	DIAMANTE D'OESTE	33,96348964	ENGENHO VELHO	32,5427324	JUTI	34,02186879
AMAJARI	31,3665781	ESPIGÃO ALTO DO IGUAÇU	33,7467432	REDENTORA	33,478456	LAGUNA CARAPÃ	30,12210135
ALTO ALEGRE	30,17274132	HONÓRIO SERPA	34,56429771	ROQUE GONZALES	34,67899104	MIRANDA	34,02754839
BONFIM	32,45978396	MANGUEIRINHA	33,93923639	AMAMBAI	31,11803194	NIOAQUE	33,99717114
IRACEMA	32,96455636	MANOEL RIBAS	32,49382929	ANTÔNIO JOÃO	34,59926023	PARANHOS	30,43596181
PACARAIMA	32,60954824	NOVA LARANJEIRAS	30,76252882	ARAL MOREIRA	33,97624686	PORTO MURTINHO	34,45327045
JACAREACANGA	32,01443564	ORTIGUEIRA	34,39155126	CAARAPÓ	32,26801368	TACURU	31,10806555
SANTANA DO ARAGUAIA	34,62922062	PORTO BARREIRO	34,65511747	CORONEL SAPUCAIA	33,75111668	RIBEIRÃO CASCALHEIRA	34,12932513
PEDRA BRANCA DO AMAPARI	33,94621961	SÃO JERÔNIMO DA SERRA	34,63842857	DOIS IRMÃOS DO BURITI	32,53686249	RONDOLÂNDIA	34,36794567
SÃO JOSÉ DO VALE DO RIO PRETO	34,40494527	TAMARANA	32,17384286	DOURADINA	32,10023509		
PEDRA BELA	34,40157361	TURVO	34,28832297	ITAPORÃ	31,24951777		
BOA VENTURA DE SÃO ROQUE	33,87011811	ENTRE RIOS	34,21055179	ITAQUIRAÍ	34,54069899		

Table 6. Number of municipalities outlying before and after carrying out the procrustes procedure



	variables	PC2**	PC3**	Confidence ellipse**	MD**	MRD**	bag-plot*
All variables	254	83	145	66	149	117	930
PROCRUSTES	139	105	134	39	101	51	920
variation(%)	-45,28	26,51	-7,59	-40,91	-32,21	-56,41	-1,08

* Quantity estimated by the ranges of the main components considering all variables and only the variables selected by the procruste procedure.

** These numbers refer to the number of municipalities considered outliers

Table 7. Assessment measures for the different models

Model	S	R2	R2(aj)	R2(pred)
1	0,0969164	88,04%	87,64%	86,61%
2	0,261572	91,04%	90,75%	90,29%
3	0,202351	87,63%	87,22%	86,61%
4	0,108372	83,98%	83,45%	82,06%
5	0,120642	95,19%	95,03%	94,78%
6	0,201647	94,21%	94,02%	93,66%
7	0,727885	68,42%	67,37%	65,48%
8	1,28917	60,22%	58,90%	56,87%

Further on, Table 7 is shown for the different amounts of income and poverty proportions, it was possible to verify that they obtained a higher value of S and lower values of S, R^2, adjusted R^2 and predicted R^2 was for model 8, while the lowest value of S and the highest values of R^2, adjusted R^2 and predicted R^2 were for model 1.

The next to be shown is Table 8, which shows the frequency of times that each variable appears for each response.

Table 8. Frequency of each independent variable in each response

vartiable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total	variable	total
pZ1	7	pREG4	3	pNF4	6	pTT7	5	pTE10	6	pMP1	6	pDMC2	1	pSB1	3	pFAA8	3	pMEE1	5	pPC1	7	pdef3	5
pZ2	2	pREG5	2	pCPR1	4	pTTS1	3	pTE11	2	pMP2	4	pDMC3	2	pSB2	3	pFAA9	5	pMEE2	4	pPC2	6	pdef4	5
pSE1	7	pNIA1	4	pCPR2	5	pTTS2	4	pTE12	4	pMP3	5	pNDO1	2	pES1	5	pAA1	5	pMEE3	4	pPCI1	4	pdef5	1
pSE2	7	pNIA2	5	pCPR3	5	pTTS3	8	pTE13	1	pMP4	2	pNDO2	2	pES2	2	pAA2	5	pRD1	5	pPCI2	3	pdef6	2
pID1	6	pNIA3	4	pTGCT1	6	pTTS4	2	pDCC1	5	pMP5	3	pNDO3	3	pES3	2	pAA3	2	pRD2	6	pMOTO1	5	pdef7	2
pID2	5	pNIA4	3	pTGCT2	3	pTTS5	5	pDCC2	5	pMP6	1	pNDO4	3	pES4	2	pDL1	4	pTV1	5	pMOTO2	6	pdef8	5
pID3	6	pNIA5	5	pTGCT3	4	pTE1	4	pDCC3	5	pMP7	4	pNDO5	1	pES5	2	pDL2	3	pTV2	4	pCAR1	3	pdef9	3
pRA1	7	pNIA6	6	pTGCT4	1	pTE2	0	pDCC4	7	pMP8	3	pNDO6	2	pES6	5	pDL3	4	pML1	6	pCAR2	8	pdef10	0
pRA2	4	pNIA7	3	pTGCT5	4	pTE3	3	pDCC5	4	pMP9	2	pNMD1	2	pFAA1	7	pDL4	3	pML2	5	pQV1	3	pdef11	3
pRA3	4	pNIA8	1	pTT1	5	pTE4	5	pDCC6	1	pNC1	0	pNMD2	2	pFAA2	5	pDL5	3	pGEL1	6	pQV2	4	pdef12	4
pRA4	4	pNIA9	6	pTT2	3	pTE5	4	pVA1	6	pNC2	4	pNMD3	4	pFAA3	3	pDL6	1	pGEL2	7	pQV3	4	pdef13	1
pRA5	3	pNIA10	3	pTT3	5	pTE6	3	pVA2	5	pNC3	4	pNB1	6	pFAA4	1	pDL7	1	pCEL1	4	pQV4	8	pdef14	0
pREG1	1	pNF1	6	pTT4	6	pTE7	4	pVA3	4	pNC4	4	pNB2	5	pFAA5	4	pEE1	7	pCEL2	5	pQV5	7	pdef15	0
pREG2	3	pNF2	4	pTT5	5	pTE8	3	pVA4	4	pNC5	4	pNB3	4	pFAA6	0	pEE2	4	pTLF1	7	pdef1	4	pdef16	2
pREG3	2	pNF3	3	pTT6	4	pTE9	2	pVA5	3	pDMC1	5	pNB4	4	pFAA7	3	pEE3	2	pTLF2	5	pdef2	4		

Studying Table 8, it is possible to notice that the following independent variables pTTS3, pCAR2 and pQV4 appear in the 8 response variables and those that do not appear in any of them were pTE2, pNC1, pFAA6, pdef10, pdef14 and pdef15.

Continuing, Table 9 shows the number of outliers considering each of the different responses.

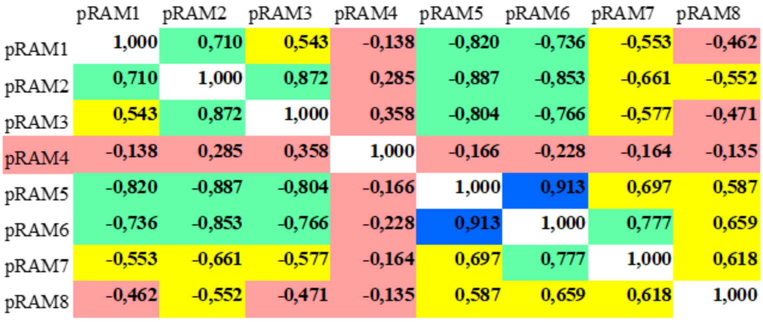
Table 9. Number of outliers and variables selected by income and poverty proportion level

responses	outliers	variables
ram1	359	99
ram2	392	90
ram3	365	92
ram4	372	103
ram5	375	99
ram6	384	89
ram7	228	59
ram8	503	57

From Table 9 it was possible to verify that the one that presented the most was 503 outliers for response ram8 and the one that presented the least was ram7 with 228 variables, while the one that presented the greatest number of variables for the adjustment was ram4 with 103 variables and the one that needed least was ram8 with 57 variables.

Next, we show Table 10 with the correlation matrix between the different proportions of income and poverty in the different municipalities in Brazil. Note that the values painted in blue are in modules greater than 0.9; in green between 0.7 and 0.9; in yellow between 0.5 and 0.7; and finally; in pink values lower than 0.5.

Table 10. Correlation matrix between the proportions of income, transformed by log ratio and poverty response variables



	pRAM1	pRAM2	pRAM3	pRAM4	pRAM5	pRAM6	pRAM7	pRAM8
pRAM1	1,000	0,710	0,543	-0,138	-0,820	-0,736	-0,553	-0,462
pRAM2	0,710	1,000	0,872	0,285	-0,887	-0,853	-0,661	-0,552
pRAM3	0,543	0,872	1,000	0,358	-0,804	-0,766	-0,577	-0,471
pRAM4	-0,138	0,285	0,358	1,000	-0,166	-0,228	-0,164	-0,135
pRAM5	-0,820	-0,887	-0,804	-0,166	1,000	0,913	0,697	0,587
pRAM6	-0,736	-0,853	-0,766	-0,228	0,913	1,000	0,777	0,659
pRAM7	-0,553	-0,661	-0,577	-0,164	0,697	0,777	1,000	0,618
pRAM8	-0,462	-0,552	-0,471	-0,135	0,587	0,659	0,618	1,000

Studying Table 10 it was possible to verify that the majority are in green followed by yellow, which corresponds to medium and high correlations and there is also a case of very high correlation between the variables ram5 (class D) and ram6 (Class C) with a value of 0.913 and the weakest was -0.138 between ram1 (below the poverty line) and ram4 (class E).

4 Conclusions

Using all variables meant the use of 254 variables, while, after the procrustes procedure, 139 variables were considered.

The scatterplot graph in Figure 5 after the data is passed through the procruste procedure maintains, in a very approximate way, the original structure of the data obtained in the same diagram in Figure 4 using all variables.

This produces an important contribution to Applied Statistics in areas such as Archaeometry and Sensometry, among others, in which the set of variables selected by Procruste closely approximates the original structure of the data when all variables are used.

Regarding outlier detection, it was noted that:

I increase their number when using only the variables selected by the procrustes procedure considering box plot of the scores of the second principal component and decrease when considering procedures such as box-plot of the third principal component, confidence ellipse, Mahalanobis distance, robust Mahalanobis distance and bag -plot

The best adjustment was for the response variable income and poverty between 0.5 and one minimum wage (model 5) and the worst adjustment was for income and poverty above 20 minimum wages (model 8).

The adjustment with the largest number of variables was for the income and poverty response between 0.5 and one minimum wage (model 4) which includes 103 variables and the one with the smallest number of variables includes 57 variables for the income and poverty response variable greater than 20 minimum wages (model 8).

The adjustment that presented the lowest number of detected outliers was for the response variable income and poverty between 7 and 20 minimum wages (model 7) with 228 municipalities considered outliers and the largest was for the response variable income and poverty above 20 minimum wages (model 8) with 503 municipalities considered outliers.

Regarding variable selection procedures, it shows great differences when using a variable selection procedure seeking to preserve their structure (procrustes) and just reducing the number of variables in order to preserve the fit (stepwise in multivariate regression).

Brazil is represented by 64.7% with an income of at most one minimum wage, and 29.3% of the population with an income of less than 0.125 minimum wage is below the poverty line.

Future Work

Comparative study between variable selection methods by adjustments using techniques such as Multivariate Regression, PLS and Canonical Correlation.

Acknowledgments

By professor Dr. Júlia Pavan for indicating the topic and to IBGE for the availability of data.

Conflicts ofInterest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: OLIVEIRA, P. T. M. S. **Data curation:** OLIVEIRA, P. T. M. S. **Formal analysis:** OLIVEIRA, P. T. M. S. **Funding acquisition:** OLIVEIRA, P. T. M. S. **Investigation:** OLIVEIRA, P. T. M. S. **Methodology:** OLIVEIRA, P. T. M. S. **Project administration:** OLIVEIRA, P. T. M. S. **Software:** OLIVEIRA, P. T. M. S. **Resources:** OLIVEIRA, P. T. M. S. **Supervision:** OLIVEIRA, P. T. M. S. **Validation:** OLIVEIRA, P. T. M. S. **Visualization:** OLIVEIRA, P. T. M. S. **Writing - original draft:** OLIVEIRA, P. T. M. S. **Writing - review and editing:** OLIVEIRA, P. T. M. S.

References

1. Aitchison, J.. *The Statistical Analysis of Compositional Data*. Chapman Hall, The Blackburn Press (2011).
2. Barbosa, J. J.; Pereira, T. M.; Oliveira, F. P. Uma proposta para identificação de outliers multivariados. *Ciência e Natura*. (2018). https://www.repositorio.ufop.br/bitstream/123456789/11454/1/ARTIGO_PropostaIdentifica%c3%a7%c3%a3oOutliers.pdf
3. Barnett, V.; Lewis, T. *Outliers in statistical data*. Wiley & Sons, New York (1994).
4. Buccianti, A.; Mateu-Figueras, G.; Pawlowsky-Glahn, V. *Compositional Data Analysis in the Geosciences from Theory to Practice*. Geological Society Special 264 (2006).
5. Bunch, J. R., Nielsen, C. P.; Sorensen, D. C. Rank one modification of the symmetric eigenproblem. *Numerische Mathematik*, **31**, 31-48 (1978).
6. Carmo, M. E.; Guizardi, F. L. O conceito de vulnerabilidade e seus sentidos para as políticas públicas de saúde e assistência social. *Cadernos de Saúde Pública* **3** (2018). doi:10.1590/0102-311x00101417.
7. Costa, M. C. R. *Qualidade de vida em adolescentes: Um estudo no terceiro ciclo do ensino básico.* 2012. 377 f. Tese. Universidade de Salamanca, Salamanca, 2012.
8. Gower, C. J.; Dijksterhuis, G. B. Procrustes Problems. Oxford Statistical Series, **30.** Oxford, England (2004).
9. Ferreira, E. B. *Análise generalizada de procrustes via R: uma aplicação em laticínios.* Dissertação de mestrado em Agronomia ULFA, Lavras-MG (2004).
10. Giroldo, F. R. S. *Alguns métodos robustos para detectar outliers multivariados.* Dissertação de Mestrado, IME-USP, São Paulo, São Paulo-SP (2008).
11. Golub, G. H.; Reinsch, C Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403-420 (1970)..
12. Jolliffe, I. J. Discarding variables in principal component analysis. I: artificial data. *Applied Statistics*, **21**, 160-173 (1972).
13. Jolliffe, I. J. Discarding variables in principal component analysis. II: real data. *Applied Statistics*, **22**, 21-31 (1973)..
14. Kranowski, W. J. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics*, **38,** 139-147 (1989).
15. Krzanowski, W. J. A stopping rule for structure preserving variable selection. *Statistics and Computing*, **6**, 51-56 (1996).
16. Leite, C. C. *Técnicas exploratórias na detecção de outliers em dados composicionais.* Dissertação de Mestrado em Matemática e Aplicações. Universidade de Aveiro, Portugal (2019).
17. Maltez, M. L. S. *Novas abordagens na detecção de outliers em dados composicionais.* Dissertação de Mestrado em Matemática e Aplicações. Universidade de Aveiro, Portugal (2020).
18. Nunes, E. L. G.; Andrade, A. G. Adolescentes em situação de rua: prostituição, drogas e HIV/AIDS em Santo André, Brasil. *Psicologia e Sociedade*, **21** (1), p.45-54 (2009).

19. Oliveira, P. T. M. S. *Pessoas com deficiência: o que encontramos por trás da inclusão*. In: XXI SINAPE, ABE, Natal-RN (2014).
20. Oliveira, P. T. M. S. *Pessoas com deficiência: questão de risco sob aplicação de regressão logística politômica e sob visão epidemiológica.* In: XV Escola de Modelos de Regressão, no período entre 2 a 5 de março de 2015. Centro de Convenções UNICAMP, Campinas-SP, Brasil, 2015.
21. Pessalacia, J. D. R.; Menezes, E. S.; Massuia, D. A vulnerabilidade do adolescente numa perspectiva das políticas de saúde pública. *Revista Bioethikos*, **4(**4), 423-430 (2010).
22. Pawlowsky-Glahn, V.; Egozcue, J. J.; Tolosana-Delgado, R. Modeling and analysis of compositional data. John Wiley & Sons, USA (2015).
23. Sibson, R. Studies in the robustness of multidimensional scaling. *Journal of the Royal Statistical Society*, B, **40**, 234-238 (1978).
24. Sajesh, T. A.; Srinivasan, M. R. An Overview of Multiple Outliers in Multidimensional Data. *Sri Lankan Journal of Applied Statistics*, **14**(2), (2013).
25. Sousa, R. C. A. *Análise estatística de dados composicionais.* Dissertação de Mestrado em Matemática e Aplicações. Universidade de Aveiro, Portugal (2016).
26. Van Den Boogaart, K. G.; Tolosana-Delgado, R. *Analyzing Compositional data with R* (2013).