Australian Journal of Basic and Applied Sciences

2024; 18(1): pages 1-12
https://doi.org/10.22587/ajbas.2024.18.1.1

ORIGINAL ARTICLE AENSI Publications

Evaluation and Comparison of Outliers' Detection Methods: A Simulation Study

Jhennifer dos Santos Nascimento[1*], Jaqueline Akemi Suzuki Sediyama[1], Anderson Cristiano Neisse[2], Thaynara Aparecida de Souza Neto[1], Paulo César Emiliano[1], José Ivo Ribeiro Júnior[1], Paulo Roberto Cecon[1]

[1] Federal University of Viçosa, Department of Statistics, Minas Gerais, Brazil
[2] Independent researcher

Correspondence Author: Jhennifer dos Santos Nascimento, Federal University of Viçosa, Department of Statistics,
E-mail: jhennifer.nascimento@ufv.br.

Received date: 15 February 2024, Accepted date: 5 March 2024.

Citation: Jhennifer dos Santos Nascimento, Jaqueline Akemi Suzuki Sediyama, Anderson Cristiano Neisse, Thaynara Aparecida de Souza Neto, Paulo César Emiliano, José Ivo Ribeiro Júnior, Paulo Roberto Cecon,2024. Bootstrapping insights into attribute equivalences in ratings-based conjoint analysis of dairy products**. Australian Journal of Basic and Applied Sciences.** 18(1): pages 1-12. https://doi.org/10.22587/ajbas.2024.18.1.1

ABSTRACT: Detecting outliers is crucial for a deeper understanding of the data set. It helps select robust methods and should precede any type of statistical data analysis. In this context, the present work aimed to evaluate and compare three outlier and four leverage point detection methods using simulation. For leverage point detection, the Leverage, DFFIT, DFBETA, and Cook distance methods were considered. For outlier detection, three variations of the Mahalanobis distance were used: the classical, the minimum covariance determinant (MCD), and the minimum volume ellipsoid (MVE). The simulation study was performed in R software, in which the methods were compared by evaluating the proportion of total detection and proportion of correct detection of different groups of discrepant points inserted into a database containing ten predictor variables and one response variable, with 100 observations each. Finally, to corroborate the results obtained in the simulation study, the seven methods were used to detect discrepant points in a real database from the milk production chain. The studies showed that the classical Mahalanobis distance and the Leverage method presented the closest evaluated detection proportions to one and were considered the best detection methods.

Keywords: *multivariate outlier; leverage points; milk production chain; outlier detection; simulation study.*

INTRODUCTION

Before conducting a statistical analysis, it is crucial to thoroughly review the data set to identify errors and patterns in the observations and ensure that the assumptions of the chosen analysis are met (Quinn & Keough, 2002). Rodrigues and Paulo (2012) emphasize the importance of conducting a descriptive study to identify anomalies that may arise from errors in the recorded values and to detect scattered data points that do not align with the general trend of the rest of the data set. According to Ahn et al. (2018), outlier detection plays a vital role in gaining a deeper understanding of the data set. It is also valuable for determining the appropriate robust methods to employ. Therefore, performing outlier detection before any statistical data analysis is recommended.

When collecting observational data, points that deviate from the other data are commonly known as outliers. Depending on their discrepancy and position, outliers can be given more specific definitions within regression analysis. For Moeller et al. (2005), those distant from the estimated regression line are considered leverage points and can be classified into good leverage points when, even far from the other data, they follow the line of the regression function and bad leverage points when they do not. According to them, all these types of outliers can manifest themselves during the adjustment of a model or predictions with a previously established

model, and both have the potential to be an influential point, which in turn is an observation that unduly influences any part of a regression analysis.

A discrepant observation can be merely an extreme manifestation of the random variability inherent in the data, the result of a deviation from the prescribed experimental procedure, or an error in calculating or recording the numerical value, Grubbs (1969). According to Barnett and Lewis (1994), inherent variability is how observations vary according to the population to which they belong. Where such variation is a natural population characteristic, it is uncontrollable and reflects the distributional properties of a model that describes data generation. On the other hand, measurement error refers to inadequacies in the measuring instrument, rounding of the values obtained, or recording errors, and execution error refers to imperfect data collection.

Osborne and Overbay (2004) emphasize that discrepant points increase error variation and reduce the power of statistical tests, besides biasing and influencing model parameter estimates. However, outliers can also provide helpful information regarding the data and its population when analyzing an unusual response from a given study. Therefore, as reported by Sen and Srivastava (1990) and Seo (2006), these points must be appropriately detected in order to address them best, either to improve statistical analysis or to identify deficiencies in a model.

Quinn and Keough (2002) point out that when these values are identified as coming from an error, they should be excluded from the sample and that these outliers must be excluded only when there are a priori reasons to exclude them, as discarding observations merely because they are confusing or reduce the chance of obtaining a meaningful result is unethical. Barnett and Lewis (1994) add that corrected values can also replace these values, though in circumstances where it cannot be guaranteed that an error has occurred, the only alternative is to consider the atypical value to be random and inherent in the natural variation of the data. In the case that there is no reason to suspect that a discrepant value is an error, Quinn and Keough (2002) indicate the researcher must look for a plausible justification within the problem at hand to determine the reason for the discrepancy and use statistical techniques that are robust to discrepant values.

According to Penny and Jolliffe (2001), a univariate atypical observation is detected as discrepant and said to be an outlier when detection methods are applied to each variable individually. For multivariate data, considering multiple explanatory variables, a discrepant point is the result of a combination of occurrences, not necessarily uncommon, which, when considered together, stand out in relation to the others.

Multivariate outliers are particularly relevant when they involve more than two explanatory variables Leys et al. (2018). Procedures for identifying multiple influential observations in linear regression and evaluation of robust outlier detection methods can be seen in Nurunnabi et al. (2013) and Templ et al. (2019). Some recent references to articles in statistics on multivariate outliers are Lobato Junior and Veiga (2020), Domino (2020), López-Oriona and Vilar (2021), Aguiar et al. (2021), Amovin-Assagba et al. (2022) and specifically in agricultural sciences Gao et al. (2018), Navarro et al. (2021).

Checking the distance between objects in space helps in detecting discrepant points in multivariate data since the distance between objects is a type of measure of similarity, where a large distance means a low similarity and a small distance means a high similarity Varmuza and Filzmoser (2016). For Rousseeuw and van Zomeren (1990), outliers in a multivariate point "cloud" can be challenging to detect, especially when the number of variables studied exceeds two and one cannot rely on the visual perception of the graph of these points.

In this context, the present work aimed to evaluate and compare seven outlier detection methods, three general and four considering an adjusted regression model. The methods were compared by evaluating the proportions of total detection and correct detection of different groups of outliers inserted into a simulated data set. The simulation study was developed in software R, R Core Team (2023), described according to the ADEM technique proposed by Morris et al. (2019), which addresses the aims, data-generating mechanisms, estimands, methods, and performance measures. The codes used in this study are available at https://github.com/aneisse/outlier-detection-paper-code.

All methods used in this study were defined in Sections 2 and 3, namely classical Mahalanobis distance, Mahalanobis distance based on minimum covariance determinant (MCD), Mahalanobis distance based on minimum volume ellipsoid (MVE), Leverage, DFFIT, DFBETA, and Cook's distance. After evaluation and comparison of the methods, to corroborate with the simulation study, the detection behavior observed in the simulation was compared with the detection behavior of the methods in a real database from the milk production chain.

2. GENERAL OUTLIER DETECTION METHODS

The methods presented in this section are general and detect outliers directly in the data set based on Mahalanobis distance variations.

2.1. Classic Mahalanobis distance

All variations of Mahalanobis distances have as their principle the evaluation of the distance between objects in space. The Mahalanobis distance, denoted by MD_i, should indicate how far the point $\boldsymbol{x_i}$ is from the center of the "cloud" of data, taking into account the shape of the cloud, Rousseeuw and van Zomeren (1990). Varmuza and Filzmoser (2016) define the Mahalanobis distance as follows.

$$MD_i = [(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \boldsymbol{C^{-1}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T]^{\frac{1}{2}}$$

Where $\bar{\boldsymbol{x}}$ is the vector of sample means of the matrix of variables $\boldsymbol{X}$ and $\boldsymbol{C}$ is the matrix of variances and covariances of $\boldsymbol{X}$.

Under the assumption of multivariate normality, the Mahalanobis distance has a chi-square distribution with p degrees of freedomχ_p^2, where p is the number of variables studied Ferreira (2018). Doulah and Islam (2018) indicate the quantile $\$\chi^2_{(p,\alpha)}$ as the cutoff point for this similarity measure, which will be the cutoff point presented in Sections 2.2 and 2.3.

Leys et al. (2018) point out that this indicator uses the multivariate sample mean and covariance matrix, which are particularly sensitive to discrepancies. For this reason, Rousseeuw and van Zomeren (1990) emphasize that it is common that the Mahalanobis distance suffers from the masking effect, a phenomenon in which a discrepant point hides another, and for this reason, multiple outliers do not necessarily have a large MD_i. According to the authors, this is because the estimators of the mean, variance, and covariance matrix are not robust and, therefore, are affected by extreme values.

2.2. Mahalanobis distance based on minimum covariance determinant (MCD)

To identify outliers based on Mahalanobis distance, it is crucial to estimate how the center of the data "cloud" and the covariance are estimated Varmuza and Filzmoser (2016). Since the classical estimators of the arithmetic mean vector and sample covariance matrix, used in Section 2.1, are sensitive to discrepant values, one can choose to use the center and covariance matrix from the minimum covariance determinant (MCD) as robust estimators for these measures of central tendency and dispersion. Denoted by MR_i the robust Mahalanobis distance is given by the following equation.

$$MR_i = \left[(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{mcd}})^T \boldsymbol{C_{mcd}^{-1}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{mcd}})^T\right]^{\frac{1}{2}}$$

Where $\bar{\boldsymbol{x}}_{\boldsymbol{mcd}}$ is the vector of sample means of observations, and $\boldsymbol{C_{mcd}^{-1}}$ is the matrix of variances and covariances based on h observations.

According to the authors, the MCD estimator chooses the subset of h observations with the most minor determinant of its covariance matrix. The ellipse formed using the center and matrix of variances and covariances based on the minimum covariance determinant is narrower than that of the classical Mahalanobis distance.

2.2. Mahalanobis distance based on minimum volume ellipsoid (MVE)

The Mahalanobis distance based on the minimum volume ellipsoid (MVE), according to Rousseeuw and van Zomeren (1990) and Lopuhaa and Rousseeuw (1991), is given by the following equation.

$$MVE_i = [(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{mve}})^T \boldsymbol{C_{mve}^{-1}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{mve}})^T]^{\frac{1}{2}}$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{mve}}$ is the vector of sample means of k observations and $\boldsymbol{C_{mve}^{-1}}$ is the variance and covariance matrix based on observations.

According to the authors, the center of the minimum volume ellipsoid of this distance is calculated based on half of the observations, and the matrix of variances and covariances is determined by the same ellipsoid multiplied by a correction factor to obtain consistency in multinormal distributions.

According to Varmuza and Filzmoser (2016), taking about half of the observations for h, seen in Section 2.2, results in the most robust version of the estimates for the vector of means and the covariance matrix. The MVE-based Mahalanobis distance is a particular case of the MCD-based Mahalanobis distance.

3. REGRESSION-BASED OUTLIER DETECTION METHODS

The methods presented in this section detect outliers by considering a fitted multiple linear regression model.

3.1. Multiple regression linear model

According to Cook (1986), statistical models are handy devices to extract and understand the essential characteristics of a data set. A beneficial model is the additive error model, which has as a particular case the linear regression model, which in turn has as specific cases the multiple linear regression model (LRM) and models of experimental designs, also known as analysis of variance (ANOVA) models (Olive, 2017).

Olive (2017) defines a response variable as the one you want to predict, and predictor variables as the variables used to indicate the response variable. Assuming a response variable Y and at least one quantitative predictor variable X_i according to Cecon et al. (2012), the statistical model of multiple linear regression describes a linear relationship between a dependent random variable Y and p independent variables X, and is given by the following equation.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + +\beta_p x_{pi} + \epsilon_i$$

with reduced matrix notation given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where, $\boldsymbol{y}$ is the vector of the response variable with dimension $n \times 1$; $\boldsymbol{X}$ is the matrix of explanatory variables with dimension $n \times (p+1)$; $\boldsymbol{\beta}$ is the vector of the multiple linear regression parameters with dimension $(p+1) \times 1$, and $\boldsymbol{\epsilon}$ is the random error vector with dimension $n \times 1$.

To perform hypothesis tests on the parameters and ensure that their least squares estimators are linear, unbiased, and of minimum variance, Cecon et al. (2012) point out that the variable must be a linear function of the independent variables, that the errors have a null expected value, that they are homoscedastic and uncorrelated with each other, that is, and $\$E[\boldsymbol{\epsilon}'\boldsymbol{\epsilon}] = \boldsymbol{I}\sigma^2$, and that they have a normal distribution.

According to Olive (2017), the vector of estimated or fitted values by the ordinary least squares method is given by the following equation.

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Hy}$$

where the matrix $\boldsymbol{H}$ is given by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \quad (1)$$

and $\boldsymbol{X}^T$ is the transposed $\boldsymbol{X}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is the inverse matrix $(\boldsymbol{X}^T\boldsymbol{X})$.

3.2. Leverage

The investigation of the diagonal elements of the least squares projection matrix, defined in Equation (1), is used to assess the influence of a discrepant observation on a fitted regression model. The element h_{ii} of the diagonal of the matrix $\boldsymbol{H}$ that exceeds $\frac{2p}{n}$ for $p > 14$ or $(n-p) > 31$ is said to be Leveraged (Belsley et al., 2004). The element can also be considered a Leverage point if it exceeds $\frac{3p}{n}$, where p is the number of model parameters, including the constant, and n is the number of observations (Kannan and Manoj, 2015).

3.3. DFFIT

The DFFIT method, difference in fit, measures the model change fit when an observation is excluded, as Belsley et al. (2004). Being,

$$DFFIT_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s\sqrt{h_i}}$$

where, $\hat{y}_i$ is the estimate of the i-th y with all observations, $\hat{y}_{i(i)}$ is the fit of y_i omitting the i-th data point, s is the estimate of the standard deviation of $\hat{y}_i$ and $h_i = \boldsymbol{x_i}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_i}^T$, where $\boldsymbol{x_i}$ is the i-th row of the $\boldsymbol{X}$ matrix, and $\boldsymbol{x_i}^T$ is the transposition of the row $\boldsymbol{x_i}$. According to Kannan and Manoj (2015) and Doulah and Islam (2018), the cut-off point for this diagnostic measure is.

3.4. DFBETA

According to Belsley et al. (2004), the DFBETA method statistic, difference in beta, accounts for the change in the parameter estimates of a linear regression when removing the i-th point in the space of the variables studied, defined by the following equation.

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\widehat{var}(\hat{\beta}_j)}$$

where, $\hat{\beta}_j$ is the estimate of the j-th parameter of the model and $\hat{\beta}_{j(i)}$ is the estimate of the j-th parameter without the i-th observation. According to the authors, the cut-off point for this measure is given by the number of observations of the model's explanatory variables.

3.5. Cook's distance

According to Cook (1979) an observation can be considered influential if important characteristics of the least squares analysis of the data, based on a full rank linear regression model, are changed substantially when the observation is excluded. With this, he indicates that the influence of the i-th observation be measured using the distance measure proposed by Cook (1977), based on the estimated parameters of a linear regression with and without a specific observation, given by

$$CD_i = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)^T \boldsymbol{X}^T\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)}{ps^2}$$

where is the estimate of the vector of parameters of the regression model with all observations, is the estimate of the vector of parameters after the -th observation is excluded, and is the transpose of the vector of residuals, is the number of parameters of the linear regression model, and is the number of explanatory variables. Under normality, the cutoff point for this distance is the median of the distribution$F_{(p,n-p)}$.

4. MATERIALS AND METHODS

4.1. Simulation study

The software R was the statistical program chosen to perform the method comparison study because it is free and open source. The study was conducted using 10000 Monte Carlo simulations to generate data sets with 10 random variables with 1000 observations each, totaling 10000 observations in each generated set.

Using the *rmnorm* function of the *lgarch* package, the random variables X_1 to X_{10} were generated from the multivariate normal distribution with mean vector $\mathbf{0}$ and variance and covariance matrix $\mathbf{\Sigma}$, such that the variables had unit variance and correlation varying randomly between $[-0.5; +0.5]$, since in practice the variables do not have zero correlation, $\boldsymbol{X} \sim NM(\mathbf{0}, \mathbf{\Sigma})$.
Table 1 shows the algebraic representation of the sample elements generated in one iteration. The samples were contaminated with groups of $3, 5, 10, 15$ and 20 outliers, such that it was possible to precisely locate the inserted outliers.

Table 1: Algebraic representation of the simulation of the sample elements generated in one iteration.

X_i	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$x_{i.1}$	$x_{1.1}$	$x_{2.1}$	$x_{3.1}$	$x_{4.1}$	$x_{5.1}$	$x_{6.1}$	$x_{7.1}$	$x_{8.1}$	$x_{9.1}$	$x_{10.1}$
$x_{i.2}$	$x_{1.2}$	$x_{2.2}$	$x_{3.2}$	$x_{4.2}$	$x_{5.2}$	$x_{6.2}$	$x_{7.2}$	$x_{8.2}$	$x_{9.2}$	$x_{10.2}$
$x_{i.3}$	$x_{1.3}$	$x_{2.3}$	$x_{3.3}$	$x_{4.3}$	$x_{5.3}$	$x_{6.3}$	$x_{7.3}$	$x_{8.3}$	$x_{9.3}$	$x_{10.3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{i.999}$	$x_{1.999}$	$x_{2.999}$	$x_{3.999}$	$x_{4.999}$	$x_{5.999}$	$x_{6.999}$	$x_{7.999}$	$x_{8.999}$	$x_{9.999}$	$x_{10.999}$
$x_{i.1000}$	$x_{1.1000}$	$x_{2.1000}$	$x_{3.1000}$	$x_{4.1000}$	$x_{5.1000}$	$x_{6.1000}$	$x_{7.1000}$	$x_{8.1000}$	$x_{9.1000}$	$x_{10.1000}$

For each group of inserted outliers, 1000 repetitions were made, and each method was evaluated using two types of detection proportions: the total detection proportion given by Equation (2) and the correct detection proportion given by Equation (3), where the number of correctly detected points corresponds to the detection of the exact outliers that were inserted.

$$\mathrm{P_t} = \frac{\text{total number of points detected}}{\text{number of outliers entered}} \tag{2}$$

$$\mathrm{P_c} = \frac{\text{number of points detected correctly}}{\text{number of outliers entered}} \tag{3}$$

The reference value for both proportions is 1, because the methods are expected to detect precisely the entered outliers.
To analyze the results obtained in each method, boxplots were created to allow a visual analysis of the measures of position, dispersion, and symmetry of the average proportions of total and the correct detection of outliers and Leverage points $\mathrm{P_c}$. The graphs were arranged side by side for better visibility of the behavior of the detections of the two analyzed proportions.

4.1.1 General outlier detection methods

For the methods in Sections 2.1, 2.2 and 2.3, the samples were contaminated with groups of $3, 5, 10, 15$ and 20 outliers inserted by drawing lots for rows and columns, adding 3 standard deviations to the observation located at the drawn position ij, i-th row and j-th column of $\boldsymbol{X}$, the data matrix.
This generated contaminated samples for each group of outliers inserted, so the detection proportions could be calculated with a high repetition number.

4.1.2 Regression-based outlier detection methods

For the leverage point detection methods, from Sections 3.2 to 3.5, it was necessary to create a response variable to fit the multiple linear regression model, since the methods require a fitted model to perform detection. The response variable Y was created with the following beta parameter vector, $\boldsymbol{\beta} = (2;\ 3;\ 1.5;\ 9;\ 3;\ 2;\ 1;\ 6;\ 4;\ 3.75)^T$.
For these methods, the samples were contaminated with groups of $3, 5, 10, 15$ and 20 outliers by drawing rows, and 3 standards deviations were added to all observations located in the i-th row of the data matrix $\boldsymbol{X}$, causing the y observations to stand out from the others and be distant from the straight line of the fitted regression model.

4.2. Real data application

The discrepant point detection methods were applied to data from the milk production chain obtained from the Educampo Leite platform, an initiative of Sebrae-MG. For this study, data from 2020 was used, which accounted for producers, partner agribusinesses, operations in municipalities, and about million liters of milk produced.
The variables X_1 - area used for livestock farming (ha); X_2 - total number of animals (heads/month); X_3 - total working hours per employee per day (hours); X_4 - somatic cell count (count); X_5 - total bacterial count (count); X_6 - lactating cows per livestock area (heads/ha); X_7 - capital stock per lactating cow (R$/head); X_8 - expenses with bulks in relation to gross income (R$); X_9 - expenses with concentrates in relation to gross income (R$); X_{10} - expenses with medicines per liter of milk (R$/liter); X_{11} - remuneration of family labor in relation to total remuneration with labor (R$), were used to explain milk production Y (1000 liters).

Initially, a descriptive analysis of all variables was performed, using the *Desc* function from the *DescTools* package of the R software, which returns a summary with descriptive statistics, such as: sample size, the number of missing values, number of zeros, mean, median, standard deviation, coefficient of variation, interquartile ranges, skewness coefficient, kurtosis coefficient, among others, in addition to graphs such as histogram and box-plot of each variable.

As the variables presented symmetry deviation and/or variance heterogeneity, the Box-Cox transformation was used in the data, as Ruppert (2001) recommended. According to him, this transformation is recommended to stabilize variances. The R function used for the procedure is *BoxCox* from *DescTools* package. Then the *stepwise* procedure was used for variable selection, which, according to Ferreira (2013), is important to avoid including correlated variables in a regression model. The *stepAIC* function of the *MASS* package of R automatically selects the variables that obtain the model with the lowest index of Akaike's information criterion Emiliano et al. (2014).

From the model with the transformed and selected variables, the assumptions of normality and homogeneity of the residuals were checked, using the Kolmogorov-Smirnov and Breusch-Pagan tests, whose functions in R, are respectively *ks.test* of *stats* package and *bptest* of *lmtest* package.
All the methods considered in the simulation study were used to detect the discrepant points present in the milk production chain data set, so it was finally possible to compare the detection patterns of the methods in the real data with those in the simulation study.

5. RESULTS AND DISCUSSION

5.1 Simulation study

5.1.1 General outlier detection methods

Based on Figures 1 and 2, the simulation study showed that the classical Mahalanobis distance was the distance that presented the total detection proportion closest to one and therefore the method that best detected the inserted outliers, since the boxplots for the correct detection proportion of the three distances were accurate and precisely at one. Thus, indicating that all distances detected the outliers that were inserted, with the distances based on MCD and MVE detecting points beyond these. Note that in Figure 1, the box plot with the green legend corresponds to the proportion of correct detection, which in turn is not visible in the graph due to the very low variability in P_c detection of the methods. So the green box plot is just dashed.

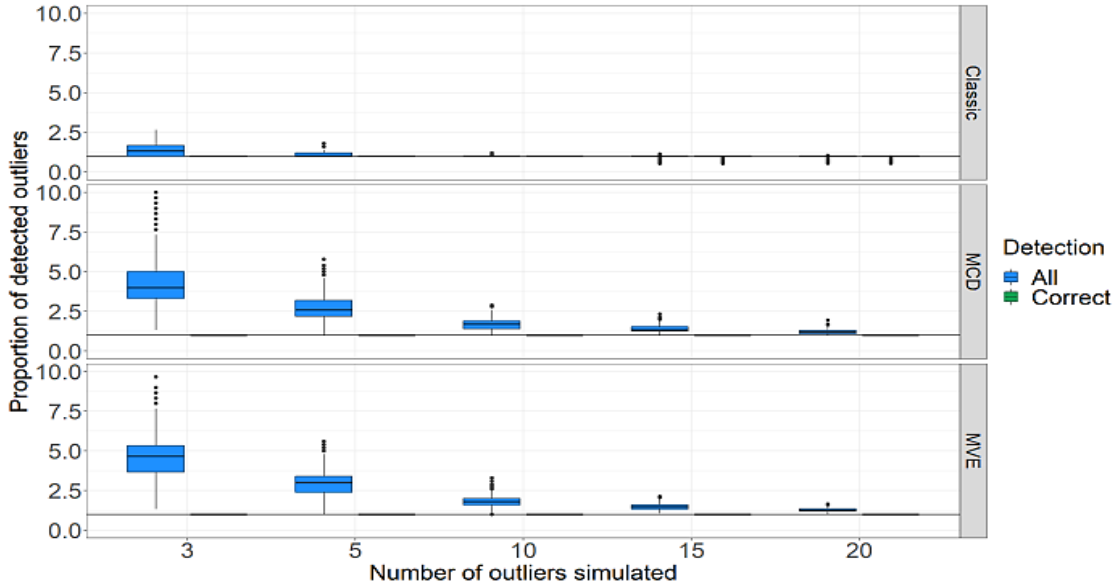


Figure 1: Boxplots for total and correct detection proportions for classical, MCD and MVE Mahalanobis distances.

By observing the boxplots for the total detection proportions, for groups of three outliers, the distances based on MCD and MVE were detected up to ten times more than the number of outliers inserted, where in 50% of the simulations, about four times more than expected was detected. As the size of the inserted outlier groups increased, the total detection proportions of these two methods became closer and closer to one, and all three methods experienced a reduction in detection variability, as seen in the blue boxplots in Figure 1.

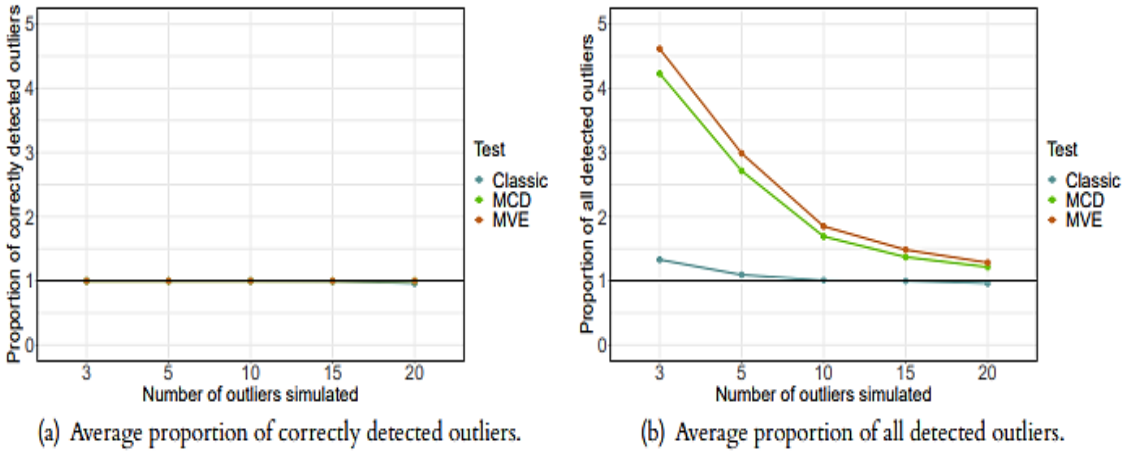


Figure 2: Average detection proportions based on 1000 Monte Carlo simulations for each outlier group for the Classic Mahalanobis Distance (classic), MCD-based Mahalanobis Distance (MCD), and MVE-based Mahalanobis Distance (MVE) methods.

Another way to present the detection proportions is using the average detection of each method, based on the 1000 simulations. In Figure 2 (a), the lines representing the average proportion of correct detection are almost wholly superimposed on the constant straight-line $f(x) = 1$, which means that, on average, the inserted outliers were detected by the three analyzed methods.

In Figure 2 (b), referring to the average proportion of total detection, the MCD and MVE distances are above the constant straight line at one, indicating that these two methods detect outliers in addition to the inserted ones. For example, for the group of three inserted outliers, the MCD-based Mahalanobis distance detected, on average, four times more than it should have detected.

5.1.2 Regression-based outlier detection method

Figure 3 shows the boxplots with the simulation results for the leverage point detection methods. The simulation study showed that the Leverage method (Figure 3a) had total and correct detection proportions close to one, indicating that the technique adequately detected the inserted discrepant points. The boxplots of the DFFIT method (Figure 3b) showed that for the groups of three and five inserted discrepant points, the total detection proportion was above one, while the proportion of correct detection was below one, indicating that the method did not detect the inserted points and ended up detecting other points besides these. For the other groups of inserted Leverage points $(10, 15, 20)$, both detection proportions were below one. Therefore, the points inserted were not detected, a negative behavior of the evaluated measures.

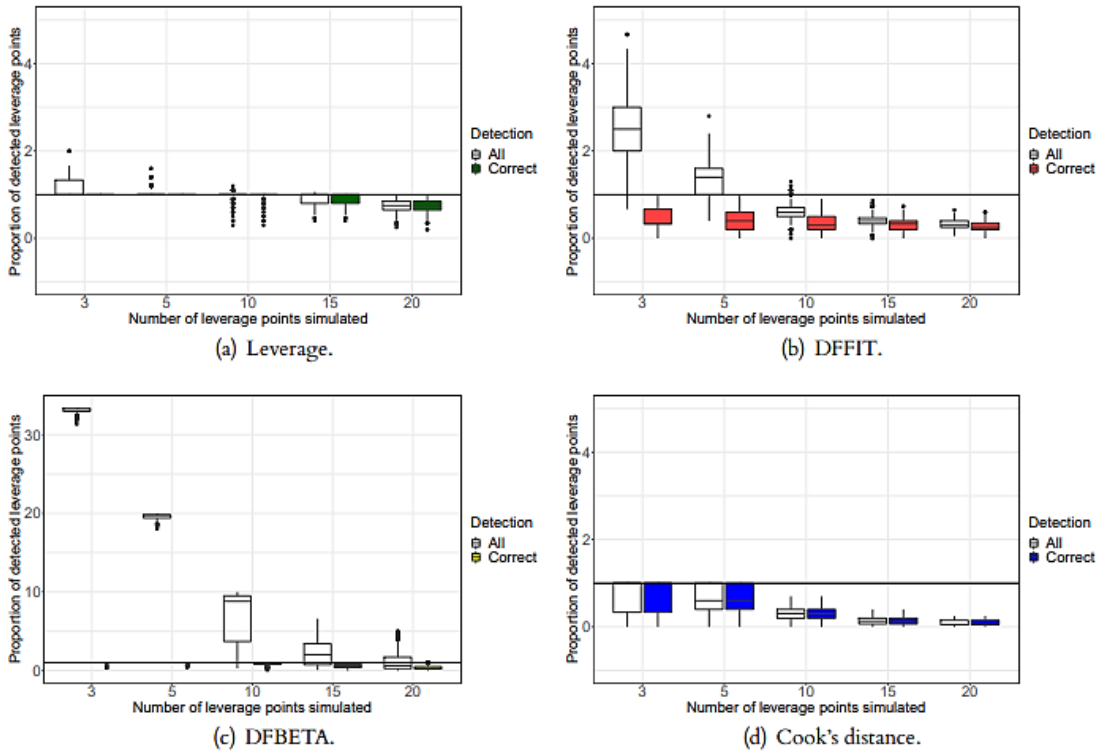
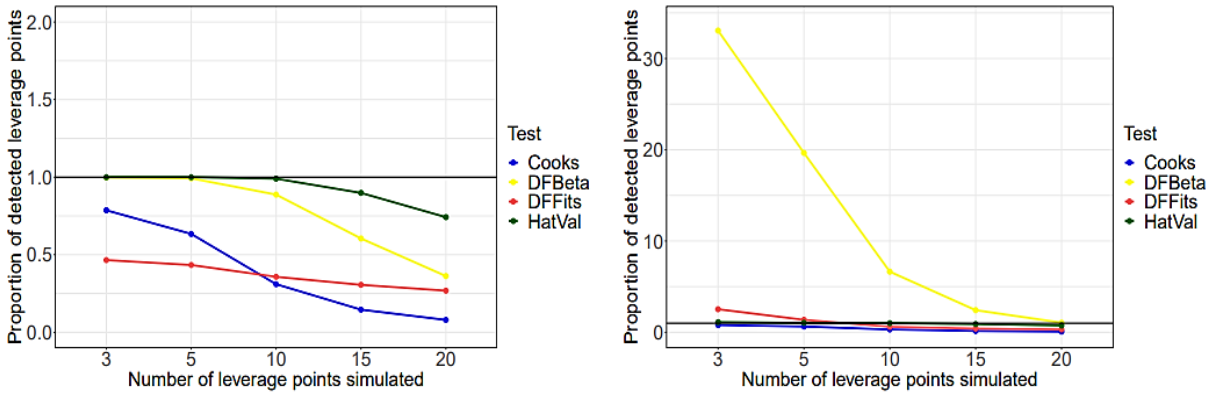


Figure 3: Boxplots for the detection proportions of the Leverage, DFFIT, DFBETA, and Cook's distance methods.

The total detection proportion of the DFBETA method [Figure 3(c)] was more significant than thirty for the three inserted discrepant points group and around twenty for the group of five inserted points. Although the proportion of correct detection remained at one in these groups, many points were detected in addition to the correct ones. For the group of ten points inserted, the total detection proportion of this method was up to five times more than it should be. Still, in the simulations the detection proportion was less than one, indicating that when it detects too many, these are not the points that were inserted. For the remaining groups, with fifteen and twenty points inserted, the two detection proportions were below one and close to zero.
Cook's distance method [Figure 3(d)] had total and correct detection proportions of one in the simulations, for the groups of three and five inserted discrepant points. As the size of the group of inserted points increased, the detection decreased, and for the group of twenty discrepant points, both detection proportions were around zero, i.e., no leverage point was detected. This method is therefore greatly affected by the masking effect, where one contrasting point affects the detection of another.

The average detection of the two evaluated proportions can be seen in Figure 4. For the proportion of correct detection [Figure 4(a)], it can be observed that the Leverage method maintained an average of one or close to one for all sizes of groups of Leverage points inserted, i.e., among the evaluated methods, it presented the most adequate detection in all groups. On the other methods had average proportions of correct detection consistently below one, apart from DFBETA, which remained at one for the groups of three and five leverage points inserted.



(a) Average proportion of correctly detected outliers. (b) Average proportion of all detected outliers.

Figure 4: Average proportions of correct detection and total detection for the Leverage (HatVal), DFFIT (DFFits), DFBETA (DFBeta), and Cook's Distance (Cooks) methods.
In the total detection proportion [Figure 4(b)], the Leverage method also remained close to one, indicating that the detected points were the same as the inserted points. The DFBETA and DFFIT methods, for the groups of three and five points inserted, showed a total detection proportion above one. Since the proportion of correct detection for this method was below one for some groups, they do not detect the points that were inserted, but other points besides these. Finally, Cook's distance method decreased the average proportion of total detection, reaching zero for twenty inserted discrepant points.

5.2. Real data application

The descriptive analysis showed that they all had the right skewness and discrepancy points. Table 2 shows some of the statistics analyzed. For this, the Box-Cox transformation was used with a parameter equal to, respectively, for the variables andX_{11}.

Table 2: Descriptive statistics of the variables. Includes sample length, mean, median, standard deviation, coefficient of variation, interquartile range (IQR), skewness, and kurtosis.

Variables	Sample length	Mean	Median	Standard deviation	Coefficient of variation	IQR	Skewness	Kurtosis
Y	625	596021.74	382300.50	688893.91	1.16	471412.50	3.32	14.36
X_1	625	89.69	62.71	101.62	1.13	74.00	5.31	48.47
X_2	625	210.69	158.96	179.53	0.85	158.88	2.55	8.78
X_3	625	3.99	2.97	3.09	0.77	2.56	2.53	8.66
X_4	625	444.95	369.56	258.89	0.58	254.25	1.88	4.77
X_5	625	54.04	19.14	154.53	2.86	33.38	11.48	179.62
X_6	625	1.37	1.03	2.04	1.49	0.88	11.77	180.41
X_7	625	36870.34	31197.36	29095.42	0.79	17685.93	5.32	41.71
X_8	625	0.21	0.22	0.12	0.53	0.11	3.57	42.94
X_9	625	0.06	0.05	0.03	0.55	0.03	2.19	9.64
X_{10}	625	0.66	0.66	0.16	0.24	0.18	0.91	6.76
X_{11}	625	0.17	0.17	0.14	0.82	0.14	3.31	24.49

After fitting the multiple linear regression model with all variables transformed, the stepwise automatic variable selection system was used, which alternately excluded and included variables in the regression model until the model with variables X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10} and X_{11}, was obtained, which was the model that obtained the lowest AIC. This model had an adjusted coefficient of determination $R^2 = 86.41\%$ and had the assumptions of normality and homogeneity of variances of the residuals met by the indicated tests.

The next step was to use all the methods considered in the simulation study to detect outliers, as presented in Table 3. This table shows that Mahalanobis distances based on MCD and MVE estimators detected more than twice as many outliers as the classical Mahalanobis distance. The MVE-based Mahalanobis distance was detected slightly more than the MCD-based one.

Table 3: Outlier selection of classical Mahalanobis distance, Mahalanobis distance MCD, Mahalanobis distance MVE, Cook's distance, DFBETA, DFFIT, and Leverage methods.

Outlier Detection Methods	Outliers detected	Leverage points Detection Methods	Leverage points detected
Mahalanobis classic	63	Cook distance	0
Mahalanobis MCD	145	DFBETA	20
Mahalanobis MVE	169	DFFIT	23
-	-	Leverage	49

In the simulation study, the Mahalanobis distance based on the minimum covariance determinant and the minimum volume ellipsoid incorrectly detected, on average, about four times more outliers than the classical Mahalanobis distance. In the real data, a similar behavior of the methods occurred, where the distances based on the MCD and MVE estimators detected . In contrast,of outliers than the classical Mahalanobis distance. In contrast, the MVE-based distance detected more outliers than the MCD-based one.

Cook's distance did not detect any outliers, and Leverage detected more than double the number of points detected by the DFFIT and DFBETA methods. According to the simulation, the DFFIT, DFBETA and Cook's distance methods failed to detect Leverage points for the largest groups of discrepant points inserted. In the actual data, the same pattern was observed, the Leverage method was the method that detected the most leverage points. The DFFIT and DFBETA methods detected about half of the points Leverage detected. And Cook's distance did not detect any Leverage points, just as in the simulation, where both detection proportions were around zero for the group of twenty points entered.

Figure 5 shows the Venn diagram with the points detected by each method, whose objective was to relate the points detected by each method through the intersections. The Mahalanobis distances based on MCD and MVE detected all the points that were detected using classical Mahalanobis distance [Figure 5(a)], and there was also an intersection of 80 points detected by these two methods.

(a) Classic Mahalanobis distance, MCD-based Mahalanobis distance and MVE-based Mahalanobis distance.

(b) DFFIT, DFBETA and Leverage methods

Figure 5: Venn diagrams with the number of discrepant points detected by each method in the milk production chain dataset.

The Leverage, DFFIT, and DFBETA detection methods [Figure 5(b)] showed a higher diversity of detected points. Since Cook's distance did not detect any points, it is not included in the diagram. It can be seen in this figure that the three methods detected in common only discrepant points, which represent the points detected by DFBETA, what was detected by DFFIT, and Leverage.

Since the classical Mahalanobis distance and the Leverage method presented the best detection proportions evaluated in the simulation study, it was interesting to observe if both would detect the same points. In this sense, Figure 6 shows the Venn diagram containing the number of points they detected. Although not comparable, all the points detected by Leverage were detected by the classical Mahalanobis distance.

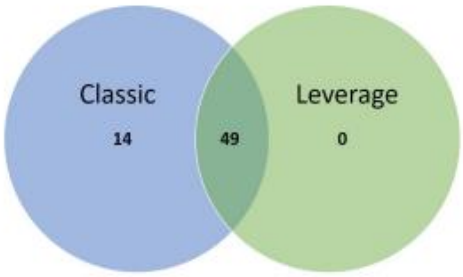


Figure 6: Venn diagram with the discrepant points detected with the Leverage and classical Mahalanobis distance methods.

6. CONCLUSIONS

Based on the simulation study, among the general methods, the classical Mahalanobis distance was considered the best method of outlier detection because it was the one that most detected the inserted outliers without detecting other points besides these. Among the methods that take regression analysis into account, the Leverage method was considered the best method for detecting leverage points for the same reason, since just like the classical Mahalanobis distance, this method had both detection proportions evaluated at close to one when compared to the other respective methods evaluated.

The methods that detected the most outliers in the simulation study were those that detected the most outliers in the real data set, and there was a relationship in the frequency of detection. The simulation study was important since it revealed who the outliers are, allowing us to determine which methods are the best. The study with real data was important to verify that the methods with the highest detection frequency would detect outliers with a similar frequency in the real data set.

Conflict Of Interest
The authors declare no conflict of interest.

Competing Interests
There are no competing interests.

Ethics Committee
None.

Funding Information
This study was partly financed by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001.

Authorship Contributions
The authors of this publication develop it under the following main aspects:
Jhennifer dos Santos Nascimento: Conceived the idea for the study, responsible for the analysis and interpretation, also responsible for writing, reviewing and translate the manuscript.
Jaqueline Akemi Suzuki Sediyama: Conceived the idea for the study, contributed to the data interpretation, responsible for writing and reviewing the manuscript.
Anderson Cristiano Neisse: Responsible for programing the simulation study.
Thaynara Aparecida de Souza Neto: Responsible for reviewing the theoretical framework.
José Ivo Ribeiro Júnior: Responsible for reviewing the theoretical framework, and final approval of the manuscript.
Paulo Roberto Cecon: Contributed to the critical review of the content, and final approval of the manuscript.
Paulo César Emiliano: Responsible for critical review of the content, and final approval of the manuscript.

REFERENCES

Aguiar, P., A. Cunha, M. Bakon, A. M. Ruiz-Armenteros and J. J. Sousa, 2021. Multivariate outlier detection in postprocessing of multi-temporal PS-InSAR results using deep learning. Procedia Computer Science, 181(5): 1146–1153. DOI: 10.1016/j.procs.2021.01.326.

Ahn, J., M. H. Lee and J. A. Lee, 2018. Distance-based outlier detection for high dimension, low sample size data. Journal of Applied Statistics, 46 (1): 13–29. DOI: 10.1080/02664763.2018.1452901.

Amovin-Assagba, M., I. Gannaz and J. Jacques, 2022. Outlier detection in multivariate functional data through a contaminated mixture model. Computational Statistics & Data Analysis, 174(3): 107496. DOI: 10.1016/j.csda.2022.107496.

Barnett, V., and T. Lewis, 1994. Outliers in statistical data. 3rd ed. John Wiley & Sons.

Belsley, D. A., K. Edwin and R. E. Welsch, 2004. Regression diagnostics: identifying influential data and sources of collinearity. 3rd ed. John Wiley & Sons.

Cecon, P. R., A. R. Silva, M. Nascimento and A. Ferreira, 2012. Métodos estatísticos. UFV.

Cook, R. D. 1986. Assessment of local influence. Journal of the Royal Statistical Society, 48(2): 133–169. DOI: 10.1111/j.2517-6161.1986.tb01398.x.

Cook, R. D. 1977. Detection of influential observation in linear regression. Technometrics, 19(1): 15–18. DOI: 10.2307/1268249

Cook, R. D. 1979. Influential observations in linear regression. Journal of the American Statistical Association, 74(365): 169–174. DOI: 10.2307/2286747.

Domino, K. 2020. Multivariate cumulants in outlier detection for financial data analysis. Physica: A Statistical Mechanics and its Applications, 558(3B): 124995. DOI: 10.1016/j.physa.2020.124995.

Doulah, M. S., and M. H. Islam, 2018. Alternative robust methods of multivariate outlier detection. Journal of Mathematical and Statistical Analysis, 1(2): 1–9.

Emiliano, P. C., M. J. Vivanco and F. S. de Menezes, 2014. Information criteria: How do they behave in different models? Computational Statistics & Data Analysis, 69(1): 141–153.

Ferreira, D. F., 2013. Recursos computacionais utilizando R. UFLA.

Ferreira, D. F., 2013. Recursos computacionais utilizando R. UFLA.

Gao, H., P. Madsen, J. Pösö, G. Aamand, M. Lidauer and J. Jensen, 2018. Short communication: Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population. Journal of Dairy Science, 101(12): 11159–11164. DOI: 10.3168/jds.2018-15123.

Grubbs, F. E., 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics, 11(2): 1–21. DOI: 10.1080/00401706.1969.10490657.

Kannan, K. S., and K. Manoj, 2015. Outlier detection in multivariate data. Applied Mathematical Sciences, 9(47): 2317–2324. DOI: 10.12988/ams.2015.53213.

Leys, C., O. Klein, Y. Dominicy and C. Ley, 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. Journal of Experimental Social Psychology, 74 (3): 150–156. DOI: 10.1016/j.jesp.2017.09.011.

Lobato Junior, D., and R. D. Veiga, 2020. Análise de diagnóstico em modelos de regressão normal e logístico. Revista Brasileira de Biometria, 38(4): 449–482. DOI: 10.28951/rbb.v38i4.461.

Lopuhaa, H. P., and P. J. Rousseeuw, 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, 19(1): 229–248. DOI: 10.1214/aos/1176347978.

López-Oriona, A., and J. A. Vilar, J. 2021. Outlier detection for multivariate time series: A functional data approach. Knowledge-Based Systems, 233 (12): 107527. DOI: 10.1016/j.knosys.2021.107527.

Moeller, S. F., J. von Frese and R. Bro, 2005. Robust methods for multivariate data analysis. Journal of Chemometrics, 19(10): 549–563. DOI: 10.1002/cem.962.

Morris, T. P., I. R. White, and M. J. Crowther, 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11): 2074–2102. DOI: 10.1002/sim.8086.

Navarro, J., I. M. de Diego, P. C. Pérez and F. Ortega, 2021. Outlier detection in animal multivariate trajectories. Computers and Electronics in Agriculture 190(11): 106401. DOI: 10.1016/j.compag.2021.106401.

Nurunnabi, A., A. S. Hadi, and A. Imon, 2013. Procedures for the identification of multiple influential observations in linear regression. Journal of Applied Statistics 41(6): 1315–1331. DOI: 10.1080/02664763.2013.868418.

Olive, D. J., 2017. Linear regression. Springer Nature.

Osborne, J. W., and A. Overbay, 2004. The power of outliers (and why researchers should always check for them). Practical Assessment, Research and Evaluation, 9(6): 1-8. DOI: 10.7275/qf69-7k43.

Penny, K. I., and I. T. Jolliffe, 2001. A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. Journal of the Royal Statistical Society: Series D (The Statistician) 50(3): 295–307. DOI: 10.1111/1467-9884.00279.

Quinn, G. P., and M. J. Keough, 2002. Experimental design and data analysis for biologists. Cambridge University Press.

R Core Team. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing (Vienna, Austria, 2023). https://www.R-project.org/.

Rodrigues, A., and E. Paulo. 2012. Introdução à análise multivariada. In Análise multivariada: para uso em cursos de administração, ciências contábeis e economia, Atlas, pp. 1–72.

Rousseeuw, P. J., and B. C. van Zomeren, 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411): 633–639. DOI: 10.1080/01621459.1990.10474920.

Ruppert, D., 2001. Statistical analysis, special problems of transformations of data. Social & Behavioral Sciences, 26: 15007–15014. DOI:10.1016/B0-08-043076-7/00513-1.

Sen, A., and M. Srivastava, 1990. Regression analysis: theory, methods, and applications. Springer-Verlag.

Seo, S., 2006. A review and comparison of methods for detecting outliers in univariate data sets. MA thesis, University of Pittsburgh.

Templ, M., J. Gussenbauer, and P. Filzmoser, 2019. Evaluation of robust outlier detection methods for zero-inflated complex data. Journal of Applied Statistics, 47(7): 1144–1167. DOI: 10.1080/02664763.2019.1671961.

Varmuza, K., and Filzmoser, P. 2016. Introduction to multivariate statistical analysis in chemometrics. CRC Press.