

**Outlier detection for the National System of Costs Survey and Indices of Civil Construction (Sinapi) at The Brazilian Institute of Geography and Statistics (IBGE)**

**<sup>1</sup>Plinio Marcos Leal dos Santos, <sup>2</sup>Lincoln Teixeira da Silva, <sup>3</sup>Ivan Meirinho de Mello, <sup>4</sup>Vladimir Gonçalves Miranda**

<sup>1</sup>plinio.santos@ibge.gov.br

<sup>2</sup>lincoln.silva@ibge.gov.br

<sup>3</sup>ivanmeirinho@gmail.com

<sup>4</sup>vladimir.miranda@ibge.gov.br

Department of Price Indices / Directorate of Surveys  
Brazilian Institute of Geography and Statistics - IBGE

## Abstract

Price indices collect periodically huge amounts of data for products in different geographic locations. Such datasets may contain outliers due to sampling and non-sampling errors. The presence of outliers may bias the estimates and lead to misleading results. In such scenario, outlier detection techniques are very important to guarantee good estimator properties. The current methodology adopted in Sinapi relies on boxplot thresholds, a non-stochastic approach, of two aggregated univariate analysis to decide whether a price is pointed as outlier or not.

This work presents a new outlier detection methodology based on multivariate Mahalanobis distance. This approach takes the covariance matrix into consideration and requires that the price dataset follow approximately a multivariate normal distribution. In the approach derived here we obtain robust mean and covariances estimates adopting the “Passo R” algorithm. Furthermore, we show how to obtain normality of the price data by the use of the Lambert Way transformation, which is able to deal with skewness and kurtosis of prices distributions and provide good approximations for normality.<sup>1</sup>

# 1. Introduction

The National System of Costs Survey and Indices of Civil Construction (Sinapi) monthly collects huge amount of input prices for civil engineering projects, with the aim of producing monthly series of costs and indices.

The collected prices are used for a double purpose. First, to compile construction input price indices for different geographical areas: Each of 27 Brazilian states and a country index, derived via aggregation of the state results. The second purpose is based on a partnership between IBGE and the public bank CAIXA, where the Sinapi prices are used to generate median prices that are used to feed a system that generates costs for different building construction projects of sanitation, infrastructure, and dwelling sectors funded via public resources (IBGE, 2017).

Due this dual nature of the Sinapi, the detection and treatment of outliers deserve extra caution, since here the focus on price levels plays a role not present in consumer price indices where price variations are the main character.

An outlier data is an observation which its magnitude appears to be surprisingly different (greater or less) than most other data (Barnett, et al., 1994). This observation may be just an extreme value of the original distribution, a contamination from a different distribution supposed to fit the data (Bustos, 1988), an error coming from a sample process (selection error of a census sector, for example) or a case of error in the survey process (wrong typing, wrong reading of the questionnaire, etc.).

The existence of outlier data can contaminate estimators and generate misleading results. In the case of Sinapi, budgeting may be overpriced due to a presence of an outlier. Thus, the process of outlier detection needs to be as robust as possible in order to support and increase the efficiency of the data analysis process, guaranteeing greater accuracy to the estimates and credibility for the survey.

This work presents the current process of automated statistical outlier detection (CEA) for the Sinapi and propose an alternative approach, based on the Mahalanobis Distances (Mahalanobis, 1936), which is more refined, producing robust parameter estimation properties, as well as the possibility to do statistical tests to decide whether a data is an outlier.

The remainder of this paper is structured as follows. Section 2 presents the current methodology and introduces the alternative approach proposed. Section 3 describes a case study using the Black Annealed wire input. Finally, section 4 presents the conclusion and next steps for implementation of the new approach at IBGE.

## 2. Methodology

### 2.1 Current methodology

The current outlier detection methodology, called CEA, make use of two statistics: Price relative ( $R$ ), and median deviation ( $Dmed$ ), which are expressed as:

$$\begin{aligned} Dmed_{t,i,u,l} &= \log(P_{t,i,u,l}/Med_{t,i,u}) \\ R_{t,i,u,l} &= \log(P_{t,i,u,l}/P_{t-1,i,u,l}) \end{aligned} \quad (1)$$

Where  $P_{t,i,u,l}$  denotes the price of an input  $i$ , observed at the month  $t$ , in a given geographic aggregate  $u$  for the sample collection site  $l$ .  $Med_{t,i,u}$  is the median price of an input  $i$ , observed at the month  $t$  in a given geographic aggregate  $u$ .

While the relative measures the variation between the current level and the price of the previous month, the median deviation is used to standardize price. The log is applied for symmetric purposes. Therefore, the relative is the statistic analyzed for “control” of monthly price changes and the median deviation is used to check price levels of a given month. The CEA also adopts filters by geographic region, which amounts to only run the CEA if a minimum of nine prices is satisfied.

Acceptance (rejection) regions are derived via the use of boxplot fences. The lower ( $LI$ ) and upper ( $LS$ ) thresholds are defined as:

$$\begin{aligned} LI_1 &= q_1 - 1,5(q_3 - q_1) \\ LI_2 &= q_1 - 3(q_3 - q_1) \\ LS_1 &= q_3 + 1,5(q_3 - q_1) \\ LS_2 &= q_3 + 3(q_3 - q_1) \end{aligned} \quad (2)$$

Where in equation (2)  $q_1$  and  $q_3$  denotes the first and third quartiles, respectively, from the data distribution.

Values of  $R$  (or  $Dmed$ ) between  $LI_1$  and  $LI_2$  are marked with a single minus sign (-), values below  $LI_2$  receives double minus sign (--). Values above  $LS_2$  receive a double plus sign (++). Finally, values above  $LS_1$ , but below  $LS_2$  receive a single plus sign (+). The boxplot in the Figure 1 illustrates the methodology used for signing the suspicious observations.

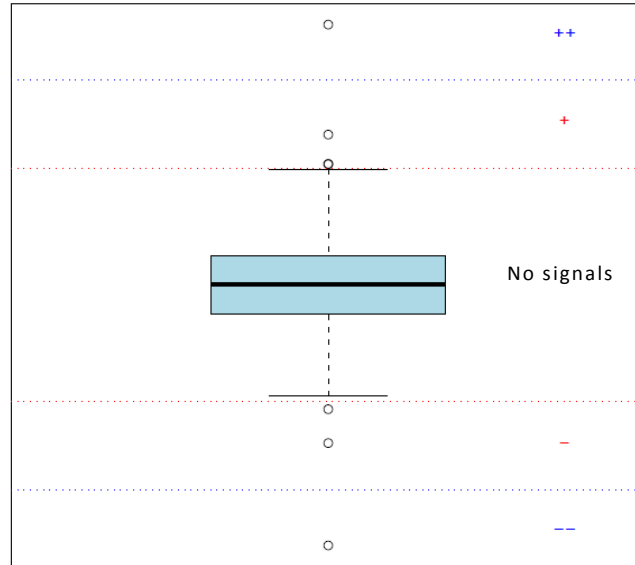


Figure 1: Illustration of the use of boxplot fences to mark suspicious observations. The dotted blue lines refer to  $LS_2$  and  $LI_2$ , while the red lines refer to  $LS_1$  and  $LI_1$

The limits derived for  $R$  and  $Dmed$  are joined in order to provide a 2-dimensional tolerance region for the determination of the tolerance region used in the CEA. In this process, the signs derived for the 1-dimensional regions, individual  $R$  and  $Dmed$ , are summed and provide the criteria to mark an observation as outlier (see Figure 2).

The yellow regions in Figure 2 represent the area where an observation receives three marks. In this case, a verification alert is signed to the analyst. While in the red regions the observation receives four marks, then the data is marked as an outlier.

Such outlier detection system has the main advantage of simplicity, from a methodological and implementation perspectives, since it is straightforward to calculate boxplot fences in any available software.

The use of the boxplot fences to determine the tolerance limits of the current method is a non-stochastic approach, without an associated distribution for parameter estimation. One limitation of this methodology is the fact that both variables under study depends on the price of the current month, which may cause a correlation bias. Also, this methodology generates too many data to be manually verified according the analysts experience, introducing an excessive element of subjectivity. Finally, use of quantiles for outlier detection assumes that the data follow a symmetric distribution. However, most of price variations is positive so the data is more sensitive to the right tail.

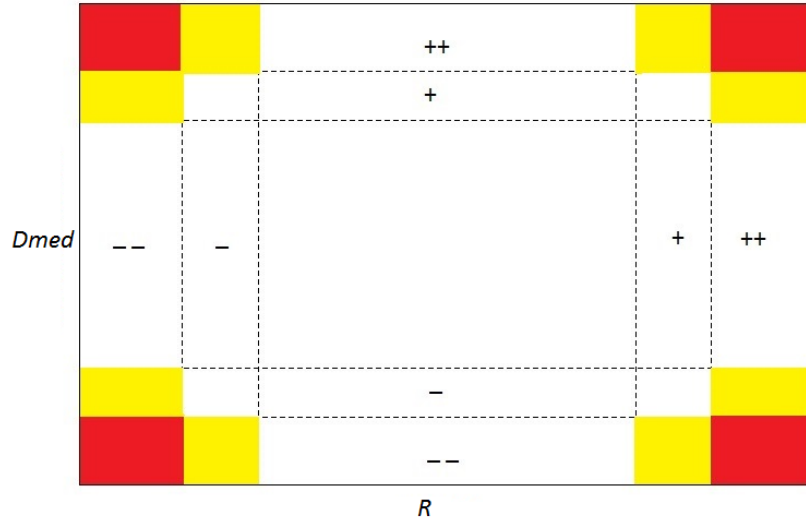


Figure 2: Outlier detection region of the CEA method. The area is constructed joining the 1-dimensional tolerance regions derived for  $R$  and  $Dmed$  using the boxplot approach in Figure 1 and summing the signs for the 2-dimensional case. Yellow area has three marks, while red area has four marks.

## 2.2 Proposal of a new outlier detection approach

In this section we discuss the proposal of the implementation of a new outlier detection system for the Sinapi survey. The main objective here is to use an outlier detection method that is good enough to identify true outliers in the data and avoid the occurrence of false positives.

The consumer price index manual presents some methods for outlier detection. The first is based in quantiles like CEA, and the second is based on Tukey algorithm, another non-parametric methodology (ILO, et al., 2004). The idea in our study is propose an approach different from these methods that proposes the construction of regions based on formulas that are a function of the data that they will be applied to. The goal in our study is to propose a stochastic methodology that allows to derive statistics test. So, the approach proposed is based on the Mahalanobis distances, which has a chi-squared distribution when the assumption of normality is fulfilled. In addition, the method enables a multivariate approach, which allows to deal with correlation effects.

The relative will be separated into two variables, namely  $P_0$  and  $P_1$ , unlike the CEA, to deal with correlation issues. By doing so, the correlation bias present in the relative  $R$  is quenched. These two variables can be considered as median deviations and their formulas are as follows:

$$P_0^{t,i,u,l,p} = \frac{P_{t-1,i,u,l,p}}{Med_{t-1,i,u,p}} \quad (3)$$

$$P_1^{t,i,u,l,p} = \frac{P_{t,i,u,l,p}}{Med_{t,i,u,p}}$$

Where in equation (3)  $P_{t,i,u,l,p}$  denotes the price of an input  $i$ , observed at the month  $t$ , in a geographic aggregate  $u$  for the sample collection site  $l$  and with given informant profile  $p$ .  $Med_{t,i,u,p}$  is the median of an input  $i$ , observed at the month  $t$ , in a geographic aggregate  $u$  and with given informant profile  $p$ .

### 2.2.1 Mahalanobis distance

The Mahalanobis distance is defined as follows:

Let  $X = [P_0, P_1, \dots]_{L \times 2}$ , thus:

$$DM_l = ((X_l - \mu)^T \Sigma^{-1} (X_l - \mu))^{1/2}, \text{ for } l = 1, 2, \dots, L \quad (4)$$

Where  $\Sigma$  is the variance-covariance matrix of  $X$  and  $\mu$  its mean vector,  $l$  represents the collection site that is our sample unit, and  $L$  is the total sample size.

The Mahalanobis distance measures the distance between two or more variables from a central (or centroid) point, considering possible variance differences and linear relationships (correlation) between the data. Mahalanobis distances are also known as generalized (or weighted) distances. When the  $\Sigma$  is equal to the identity matrix, it has the exact Euclidean distance formula (Mingotti, 2005).

This methodology is suitable for situations in which data are elliptically distributed, a standard case observed for the multivariate normal distribution. Also, one of the several applications of the Mahalanobis distance is for the development of multivariate outlier detection.

The calculation of Mahalanobis distances rely on estimates of sample mean and variances, and hence are subject to the influence of outliers in the samples, since these statistics are extremely sensitive to the presence of outliers. To circumvent this problem, it is necessary to adopt a robust estimation procedure. Also, it is important to search for a mechanism that does not neglect the outliers (which may be just extreme values of the distribution) in the estimation, but provide adequate weights to outlier data. If only the outlier data are removed from the estimation, the true scale of the observations would be underestimated.

### 2.2.2 "Passo R" (Step R)

In this work the robust estimation of the mean vector and variance-covariance matrix adopts the so called "Passo R" algorithm (Silva, 1989). The "Passo R" method refers to the second step of the "ER" algorithm, which is a modification of the "EM" algorithm. The "Passo R" aims to minimize the outlier effects on the estimation of the parameters of interest reducing the weight of the most discrepant observations, by means of an iterative process.

The "E" step from "ER" algorithm refers only to missing data problems, which is not the case of Sinapi, where data is complete, so its use would become redundant. In this way, only the "Passo R" of the algorithm is used, which corresponds to the robust estimation step of the mean and variance-covariance matrix estimators. The method is expressed by the following expressions:

$$\begin{aligned}
(a) \quad \mu_q^{I+1} &= \frac{\sum_{l=1}^L W_l X_{l,q}^I}{\sum_{l=1}^L W_l} & \forall q \in \{1, 2\} \\
(b) \quad \sigma_{q,k}^{I+1} &= \frac{\sum_l^n W_l^2 \{X_{l,q}^I - \mu_q^{I+1}\} \{X_{l,k}^I - \mu_k^{I+1}\} + C_{l,q,k}^I}{\sum_l^n W_l^2 - 1} & \forall q, k \in \{1, 2\} \\
(c) \quad W_l &= \phi(d_l) / d_l \quad \forall l \in \{1, 2, \dots, L\} \\
(d) \quad d_l^2 &= [X_{lp_l}^I - \mu_{p_l}^I]^T [V_{p_l p_l}^I]^{-1} [X_{lp_l}^I - \mu_{p_l}^I] \\
(e) \quad \phi(d_l) &= \begin{cases} d_l & \text{if } d_l \leq d_0 \\ d_l^0 \exp\{-(d_l - d_l^0)^2 / 2b_2^2\} & \text{if } d_l > d_l^0 \end{cases}
\end{aligned} \tag{5}$$

Where in equation (5)  $I$  denotes the current iteration;

$d_l^0 = (p_l)^{1/2} + b_1/2$  is the distance from which the discrepant data will have its weight damped by the estimation algorithm;

$b_1$  and  $b_2$  are control parameters set by the analyst, usually as  $b_1 = 2$  and  $b_2 = 1,25$  (Little, et al., 1987);

$\phi(\cdot)$  is an influence function (Hampel, 1973);

$p_l$  is the number of variables in the analysis in the case  $l$ . In the case of Sinapi, where the data is complete,  $p_l = 2$ , which is the number of variables available for outlier detection.

A good starting point is  $W_1 = 1$ , as the data are complete. In this case the estimator is equivalent to the maximum likelihood estimator, and after the iterative process described in equation (5), the robust values will be reached.

To summarize, the algorithm calculates the mean and variance-covariance matrix at steps (a) and (b), and the steps (c), (d) and (e) are used to reweight observations for next iteration of the algorithm, until the stop criterion is reached.

The stop criterion for the "Passo R" algorithm is adjusted by the analyst, which sets the appropriate number ( $N$ ) of iterative steps. Further, the algorithm converges to a solution when the estimates of the variance-covariance matrix or mean vector are close enough as those calculated in the previous iteration step.

Thus, the stop criterion writes:

$$\text{Run While } I \leq N \text{ or } \max \left\{ \left| \frac{\hat{\mu}_I - \hat{\mu}_{I-1}}{\hat{\mu}_{I-1}} \right|, \left| \frac{\hat{\Sigma}_I - \hat{\Sigma}_{I-1}}{\hat{\Sigma}_{I-1}} \right| \right\} \geq 0.01 \tag{6}$$



Where, in equation (6), the number of iterations defined by the analyst can be anticipated if  $\hat{\mu}$  or  $\hat{\Sigma}$  estimated in iteration  $I$  is almost the same as in the previous iteration ( $I - 1$ ). Here we fix the tolerance factor as 0.01.

### 2.2.3 Challenges for implementation

The "Passo R" algorithm is suitable for robust estimations if the dataset follows an approximately normal distribution, even when subjected to contamination of atypical observations. Such restrictions have to be taken into account since prices and salaries are usually derived from asymmetric distributions. Therefore, a transformation of our prices is necessary for the implementation of the "Passo R" and the Mahalanobis distances.

The Box-Cox method (Box, et al., 1964) is a popular tool used to transform a dataset to normal distribution. The log transformation, which is a special case of the Box-Cox, is a simple transformation commonly used to symmetrize the data. The log transformation sounds as an attractive option for the Sinapi, where the "Passo R" algorithm needs to be implemented for up to 400 inputs in 27 Brazilian states, since it is easy to implement and is not CPU-intensive, an important aspect to speed-up the process.

Figure 3 presents the result of the application of the log transformation to the whole distribution of prices in the Sinapi database using the variable  $P_1$ . Notice that skewness issues of the original price distribution are solved through the use of the log transformation. However, the transformed data fits to the Laplace distribution (red curve), which is characterized by the presence of heavy tails (high kurtosis values), thus depart them from normality. Such characteristic prevents the use of "Passo R" and Mahalanobis distances.

The use of Box-Cox transformation instead of the log do not solve the heavy tail problem, since it is not suitable for heavy tails. The Box-Cox is applied to cases of high data variability and skewness, especially for standardization of residuals coming from linear models. Thus, another methodology is necessary.

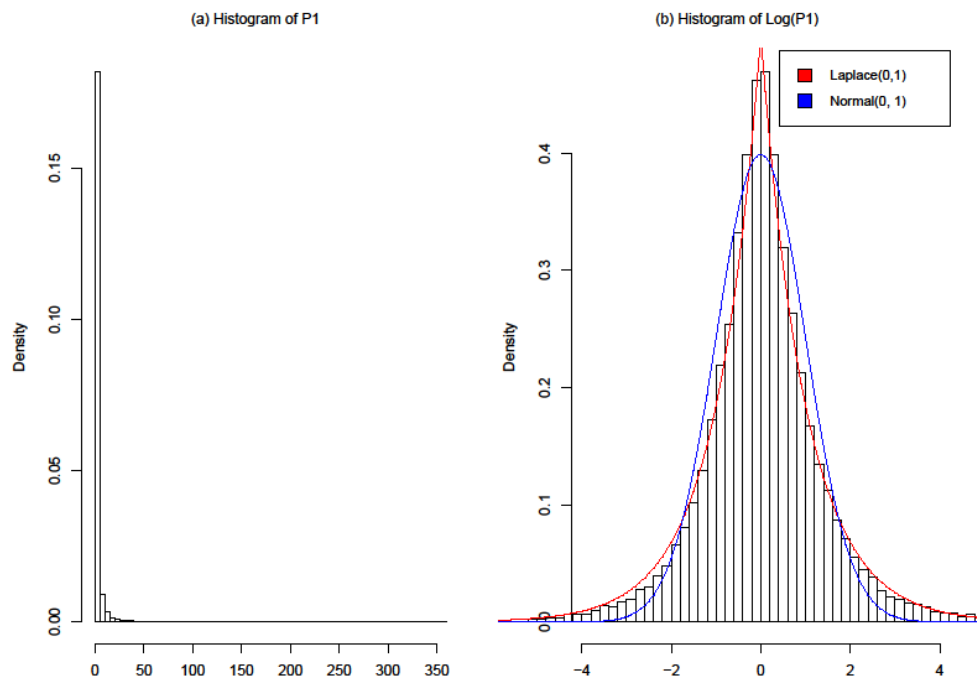


Figure 3: Distribution of input prices of Sinapi. from April 2018. (a) represents the density of the universe of prices of Sinapi. (b) shows the log transformation applied in the universe of prices of Sinapi, the red curve represents the double exponential (Laplace) distribution, while the standard normal distribution is in blue.

#### 2.2.4 Lambert Way Transformation

The Lambert Way (LW) transformation can deal with problems of both skewness and kurtosis, generating approximately normal distributions (Goerg, 2014). The LW function provides an explicit inverse distribution which, estimated via maximum likelihood, can remove heavy tails from a distribution and still provide explicit expressions of cumulative distribution function (cdf) and probability density function (pdf).

In summary, this methodology consists in estimating a parameter ( $\delta$ ), by the use of a maximum likelihood method, that governs the tail behavior. The LW method guarantees a bijective transformation, independent support of parameter values, and uniqueness when  $\delta > 0$ . Thus, for simplicity we must assume  $\delta > 0$  (Goerg, 2014).

The LW transformation can be easily implemented via the LambertW package available in the R software (R Core Team, 2013)

The major limitation of this methodology relies on the fact that it cannot be applied to joint distributions, hence if applied to marginal distributions it is not guaranteed that a set of variables will be jointly normalized. The method depends on whether these marginal transformations maintain the previously existing correlations in order to fit to a normal multivariate distribution.

Figure 4 illustrates the result of the data obtained after the use of the LW transformation to solve the heavy tails problems. Note that the data fits nicely to an approximately normal distribution (blue curve), while the Laplace distribution (red curve) is no longer suitable to the data.

Once we guarantee that our data fits a normal distribution, we can rely on the Mahalanobis distances and the “Passo R” to derive robust and stochastic-based tolerance regions.

As mentioned before, is necessary a method that is fast enough to be put into practice in Sinapi. Thus, the combination of the log transformation for symmetrization of the data and the LW transformation restricted to solve problems of heavy-tail is an attractive option since we can save some time with parameter estimation.

If we can ascertain that our dataset is approximately normal we can benefit of theoretical results for the Mahalanobis distances, that under such circumstances can be approximated to a chi-square distribution. This result is very suitable for the definition of the criteria for detecting outliers, since we can rely on the quantiles of the chi-square distribution to establish tolerance regions.

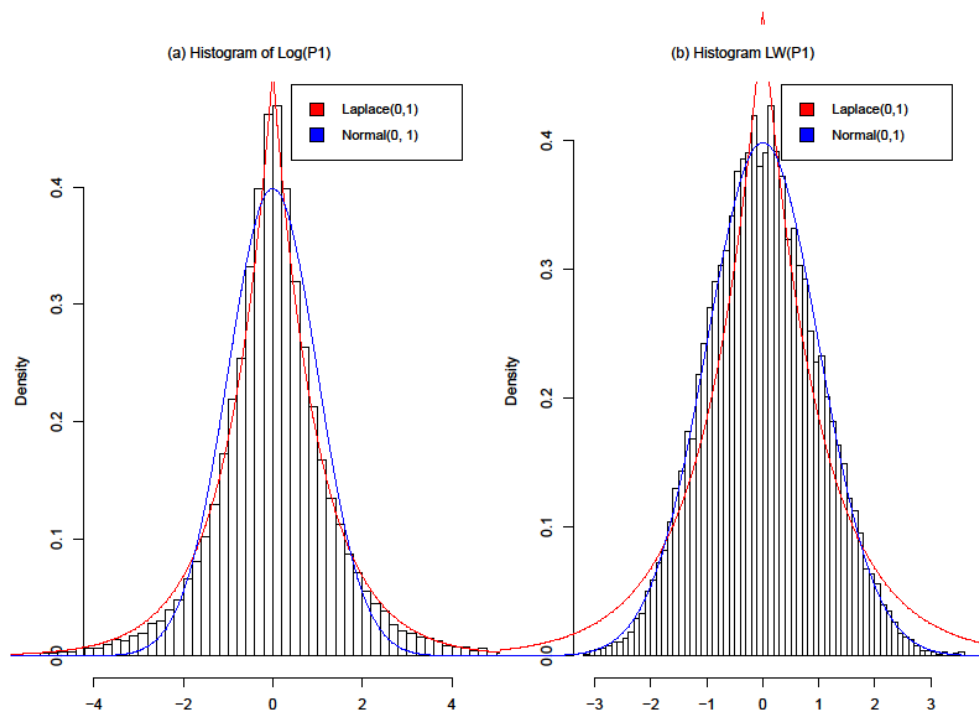


Figure 4: Distribution of transformed input prices from April 2018. (a) log transformation applied in the universe of prices of Sinapi. (b) shows the Lambert Way transformation applied after the log transformation in universe of prices of Sinapi. The red curve represents the double exponential (Laplace) distribution, while the standard normal distribution is in blue.

### 3. Case study – Black annealed wire

In this section we compare the CEA with the method proposed in section 2.2. The analysis is employed to the input Black Annealed Wire (BAW), present in the Sinapi's database. The data has sample size equal to 303 and represents the reference period of April 2018.

Figure 5 shows the application of CEA to the BAW input. One can note that only one discrepant point was detected, receiving four plus signs, while no point received three signals, would require verification by the analyst according to CEA. We note that in order to calculate the tolerance limits for the CEA all observations with price relative equal to unity are excluded. A process similar to the employed in the Tukey algorithm (ILO, et al., 2004).

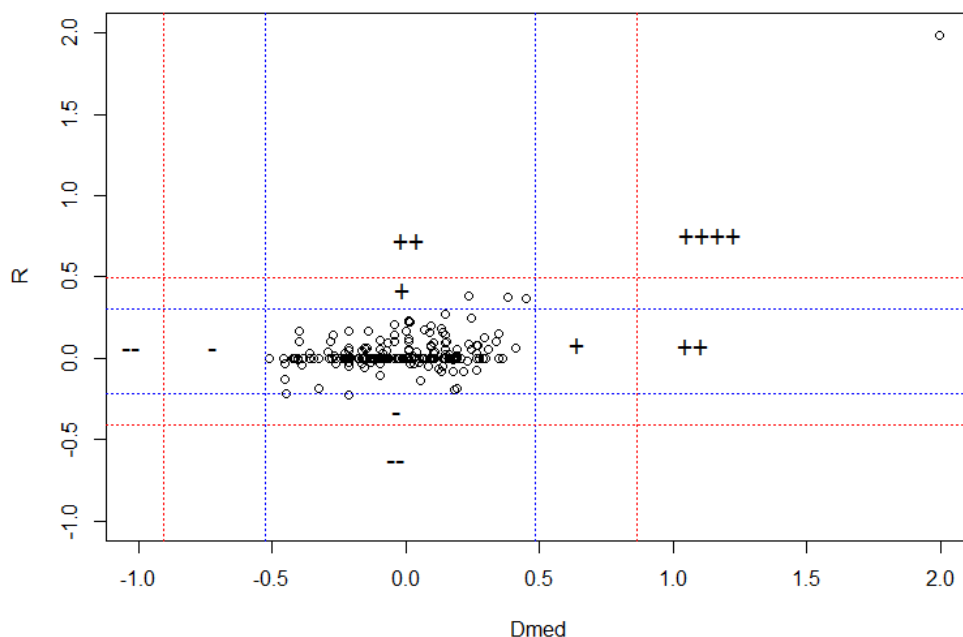


Figure 5: Tolerance regions according CEA. Red lines represent the area that a point receives two marks (plus or minus sign), blue lines represent the area where a point receive a single sign.

Figure 6 shows the histograms of the log transformed standard prices  $P_0$  e  $P_1$  for the BAW input and after the LW transformation. Figure 6 (a) reveals the presence of extreme values, justifying the use of “Passo R” to calculate robust statistics to be used in the calculation of Mahalanobis distances, reducing the influence of the extreme values on the estimates. In addition, the p-values of the Shapiro-Wilk normality test are shown for the comparison between the BAW log price and the transformed version of the data via the LW method. One can note that the data transformed via LW can be considered normal at a significance level of 1%. The transformation for the case of  $P_0$  was not necessary, since the log was already enough to normalize the data, even so the LW transformation managed to improve its results.

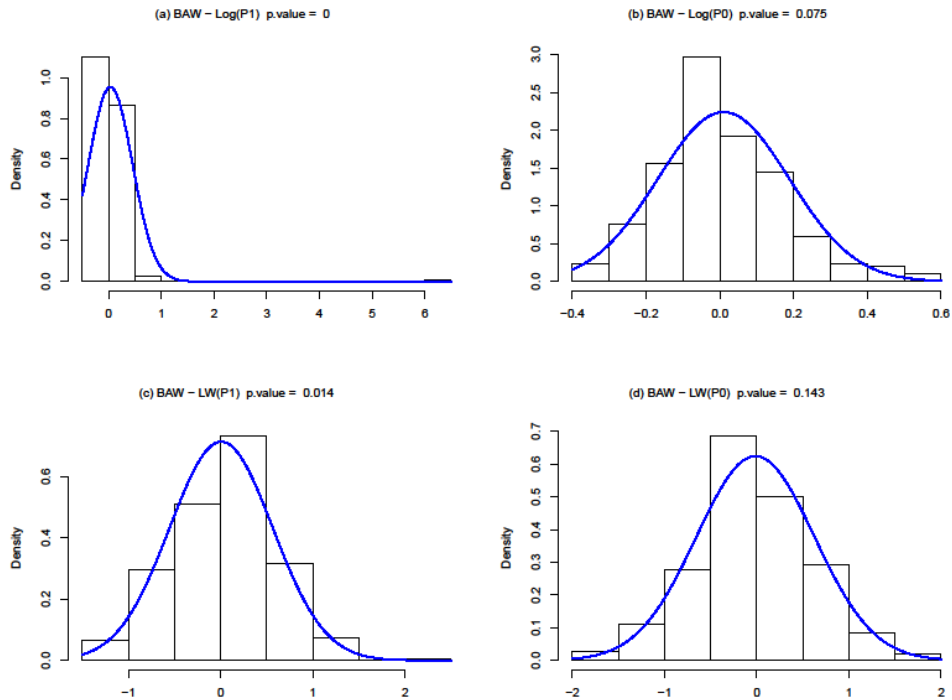


Figure 6: BAW Histograms using log transformation (top) and using the LW transformation (bottom). The blue line represents the theoretical quantiles of the standardized normal distribution.

The LW method does not guarantee the multivariate normality, necessary for the implementation of Mahalanobis distances, although the normality assumptions of the independent variables can be accepted. However, Figure 7 presents the joint distribution of  $P_0$  and  $P_1$  and reveals that the transformed data fits to a multivariate normal distribution. It is interesting to note the proximity to a bell form in Figure 7 (a), as expected for the normal multivariate density. The contour plot presents an elliptical shape in Figure 7 (b), a typical situation in which the use of the Mahalanobis distances for outlier detection applies. In addition, the Royston test is presented, which is a multivariate extension of the Shapiro-Wilk test (Royston, 1983) and can be run by the use of MVN package in R (Korkmaz, et al., 2014). At a significance level of 1%, the multivariate normality hypothesis between the variables cannot be rejected.

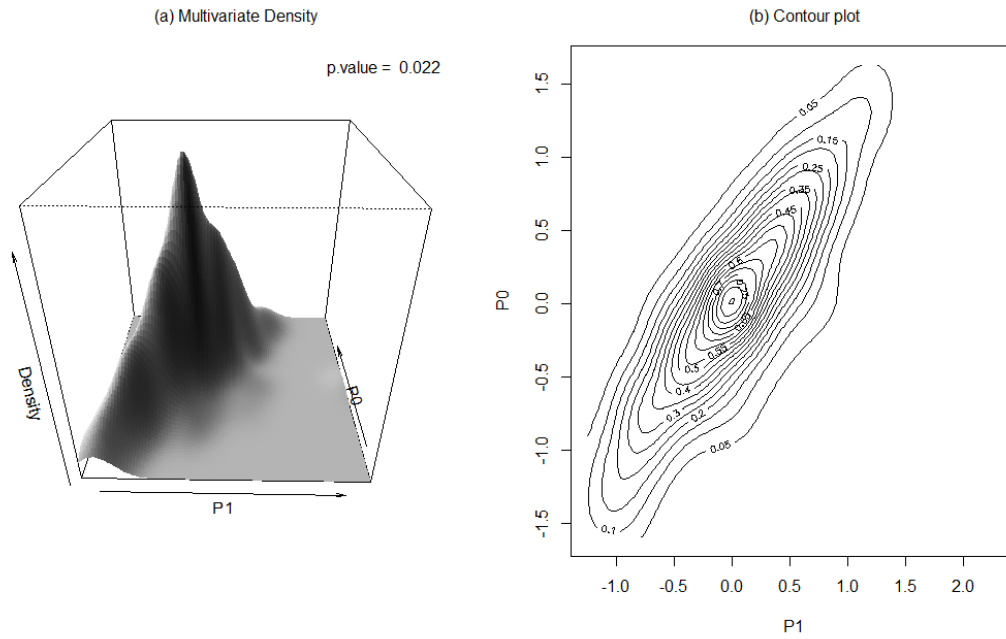


Figure 7: (a) 3-dimensional perspective multivariate density plot, where two dimensions refer to the values of the  $P_0$  and  $P_1$  and the third dimension is the multivariate normal probability density function. (b) Contour plot represents a projection of the multivariate density plot in a 2-dimensional space. Also, the p-value of the Royston Multivariate normality test is presented.

Once the normality assumption is fulfilled, we can calculate the Mahalanobis distances using the “Passo R” methodology for the robust estimation of the means and variances.

Figure 8 presents the results obtained, where tolerance regions are derived for two levels of significance,  $\alpha = 1\%$  and  $\alpha = 5\%$ . We note the presence of a discrepant observation (red dot in Figure 8) which departs from the bulk of the data and lies outside the tolerance region even for  $\alpha = 1\%$ . This is the same observation that was detected by the CEA. One also notes the presence of three observations (blue dots) lying outside the tighter tolerance region set by  $\alpha = 5\%$ .

By setting the criteria that values lying outside the outer thresholds are outliers and these that lying between the two tolerance regions as extreme values to be verified, the new method identified one outlier and three observations that needs to be verified by the analyst, while the CEA identified just one discrepant observation.

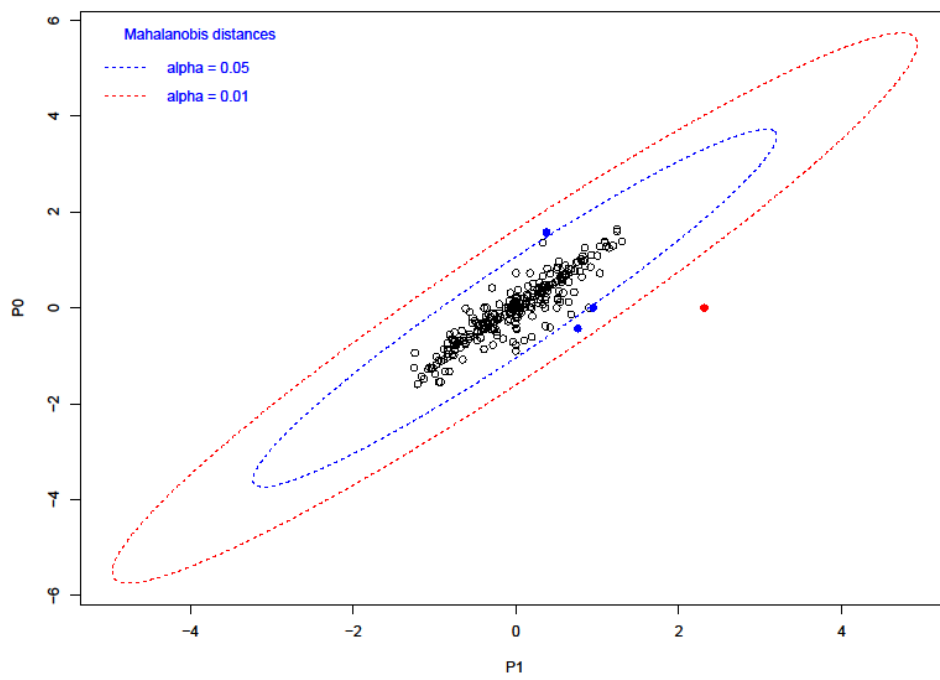


Figure 8: Tolerance regions using Mahalanobis distances. Red curve represents the tolerance limits at  $\alpha = 1\%$  while blue curve represents limits at  $\alpha = 5\%$ . Sample size = 303.

## 4. Conclusion

This work presented the current methodology of outlier detection of the Sinapi survey and proposed another method, based on the Mahalanobis distances, that surpasses the limitations of the current one.

We discuss how the robust estimation of mean vector and variance-covariance matrix in the calculation of Mahalanobis distances are taken into account by means of the "Passo R" algorithm. The assumption of normality, which is commonly neglected in these types of analyses, was considered too. The combination of Log and LW method proved to be able to circumvent problems with skewness and kurtosis, respectively, guaranteeing the normality necessary for the application of "Passo R" and the Mahalanobis distances.

A case study was presented for the BAW input used for the comparison between CEA and Mahalanobis distances. Both methods were able to detect the most discrepant point of the distribution. We noted that the CEA is more permissive than the Mahalanobis distances, missing three possible extreme values detected by the Mahalanobis distances tolerance region.

The proposed fences  $\alpha = 1\%$  and  $\alpha = 5\%$  are suitable for the BAW input. However, a fine-tuning need to be performed considering various inputs of the Sinapi, to decide as to whether these fences really are the most appropriate, and if different values should be adopted for different inputs.

Several products in the Sinapi input database have a large price variation due to their region of commercialization, but products geographically characterized were not considered. This issue will be addressed in future developments of the method.



## References

- Barnett, V e Lewis, T. 1994.** Outliers in statistical data. *John Wiley & Sons*. 1994.
- Beguin, Cedric e Hulliger, Beat. 2003.** Robust Multivariate Outlier Detection and Imputation with Incomplete. *EUREDIT Deliverable D4/5.2.1/2 Part C*. 2003.
- Box, G. E. P. e Cox, D. R. 1964.** An analysis of transformations. *Journal of the Royal Statistical Society*. Series B, 1964, Vol. 26.
- Bustos, Oscar H. 1988.** *Outliers y robustez*. Rio de Janeiro : Relatório Técnico - IMPA, 1988.
- Goerg, Georg M. 2016.** LambertW: An R package for Lambert W x F Random Variables. *R package version 0.6.4*. 2016.
- Goerg, Georg M. 2014.** The Lambert Way to Gaussianize heavy tailed data with the inverse of Tukey's h transformation as a special case. *The Scientific World Journal*. 2014.
- Hampel, F. R. 1973.** Robust estimation: a condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*. 27, 1973, pp. 87-704.
- IBGE. 2017.** *Sistema nacional de pesquisa de custos e índices da construção civil: métodos de cálculo*. Rio de Janeiro : Coordenação de Índices de Preços, 2017. Vol. 43.
- ILO, et al. 2004.** Consumer Price Index Manual: Theory and Practice. *International Labour Office*. 2004.
- Korkmaz, S, Goksuluk, D e Zararsiz, G. 2014.** MVN: An R Package for Assessing Multivariate Normality. *The R Journal*. 6, 2014, Vol. 2, pp. 151-162.
- Little, R J A e Smith, P J. 1987.** Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*. 82, 1987, pp. 58-68.
- Mahalanobis, P C. 1936.** On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*,. 1936, Vol. 2.
- Mingotti, S A. 2005.** *Análise de dados através de métodos de estatística multivariada*. Belo Horizonte : Universidade Federal de Minas Gerais - UFMG, 2005.
- R Core Team. 2013.** R: A language and environment for statistical. *R Foundation for Statistical Computing*. 2013.
- Rais, Saad . 2008.** Outlier detection for the Consumer Price Index. *SSC Annual Meeting*. 2008.
- Royston, J P. 1983.** Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W. *Applied Statistics*. 1983, Vol. 32, 2, pp. 121-133.
- Silva, Pedro Luís do Nascimento. 1989.** *Critica e imputação de dados quantitativos utilizando o SAS*. Rio de Janeiro : Instituto de Matemática Pura e Aplicada - IMPA, 1989.