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

Plinio Santos, Lincoln Silva, Ivan Mello, Vladimir Miranda

plinio.santos@ibge.gov.br

16th Meeting of the Ottawa Group

Rio de Janeiro - Brazil

May, 2019

Background and motivation

- The National System of Costs Survey and Indices of Civil Construction (SINAPI) monthly collects huge amount of input prices for civil engineering projects.
- 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.
- Such datasets may contain outliers due to sampling and non-sampling errors. The presence of outliers may bias the estimates generating misleading results. In such scenario, outlier detection techniques are very important to guarantee good estimator properties. This work present the current methodology of outlier detection for SINAPI and a new proposal based on Mahalonobis distances.

Current methodology adopted in Sinapi (CEA)

- Boxplot thresholds
  - $LI_1 = q_1 - 1{,}5(q_3 - q_1)$
  - $LI_2 = q_1 - 3(q_3 - q_1)$
  - $LS_1 = q_3 + 1{,}5(q_3 - q_1)$
  - $LS_2 = q_3 + 3(q_3 - q_1)$
- Two variables (median deviation and relative)
  - $Dmed_{t,i,u,l} = \log(P_{t,i,u,l}/Med_{t,i,u})$
  - $R_{t,i,u,l} = \log(P_{t,i,u,l}/P_{t-1,i,u,l})$

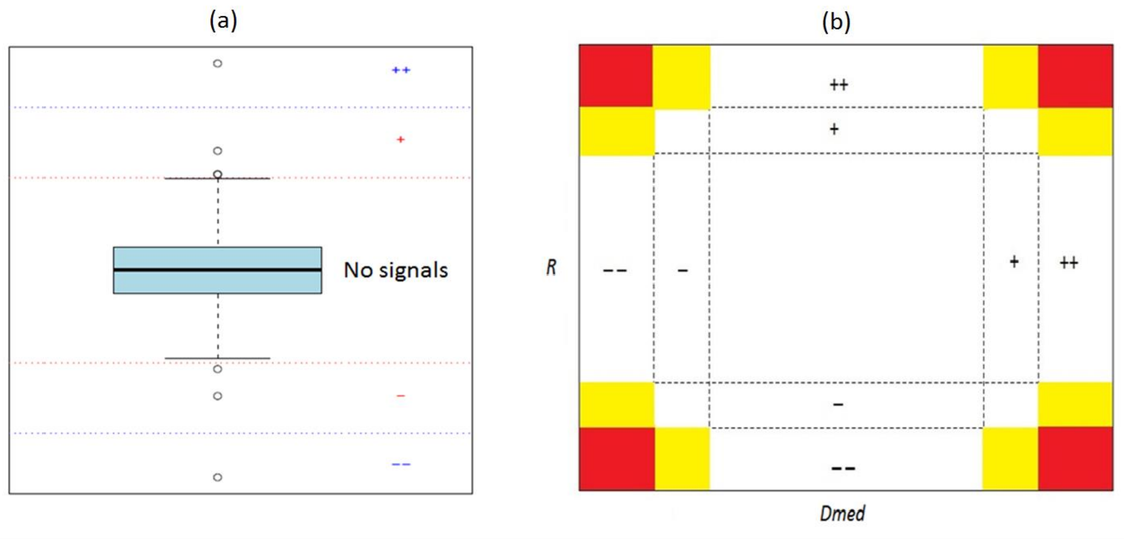

Figure 1: (a) Illustration of the use of boxplot fences to mark suspicious observations. (b) Outlier detection region of the CEA method, joining the 1-dimensional tolerance regions derived for R and Dmed.

New outlier detection approach

- Two Variables (both median deviations)

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

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

- Mahalanobis Distances

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

  - We can rely on the quantiles of the chi-square distribution to establish tolerance regions.
  - Need a robust estimation method for μ and Σ.
  - We adopt the **"Passo R"** algorithm for robust estimation, which minimize the outlier effects on the estimation of the parameters of interest by reducing the weights of the most discrepant observations.

- Challenges of implementation
  - Both Mahalanobis distances and "Passo R" are suitable for datasets that follows an approximately normal distribution.
- The log transformation solves skewness issues, but can not solve heavy-tails issues.
- The Lambert Way Transformation
  - The Lambert Way (LW) transformation can deal with problems of both skewness and kurtosis. 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). The combination of the log transformation for symmetrization of the data and the LW transformation restricted to solve heavy-tail issues is an attractive option since we can save some time with parameter estimation.

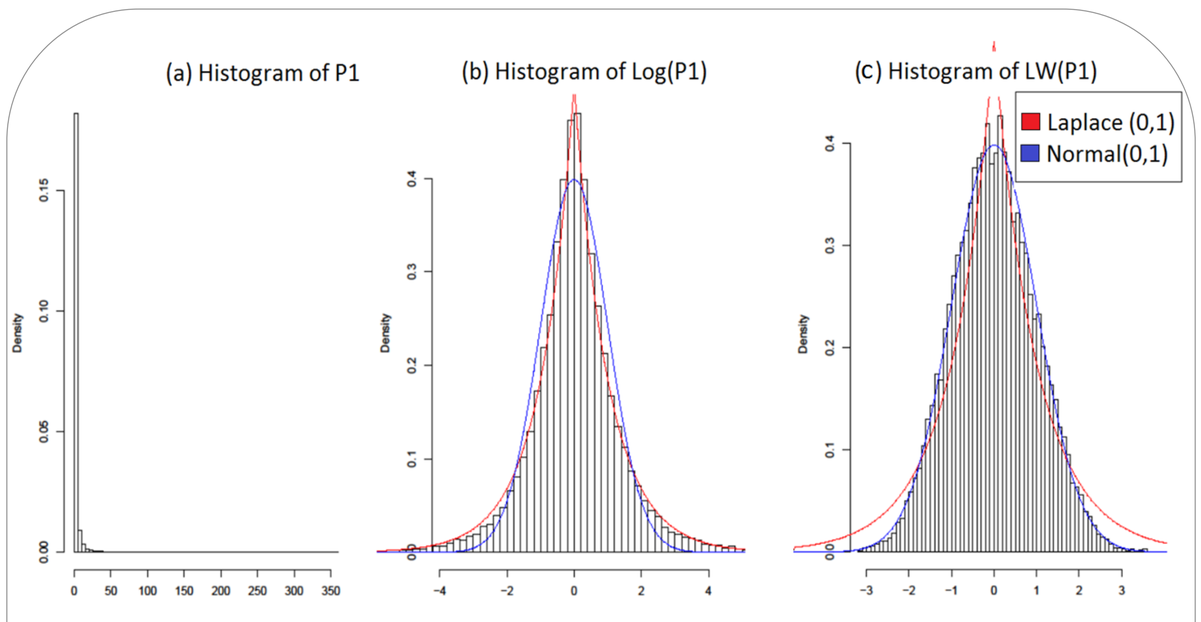

Figure 2: Distribution of input prices of SINAPI for April 2018. (a) represents the density of the universe of prices. (b) shows the log transformation applied in the universe of prices. (c) Lambert Way transformation applied after the log transformation

Case study – Black annealed wire

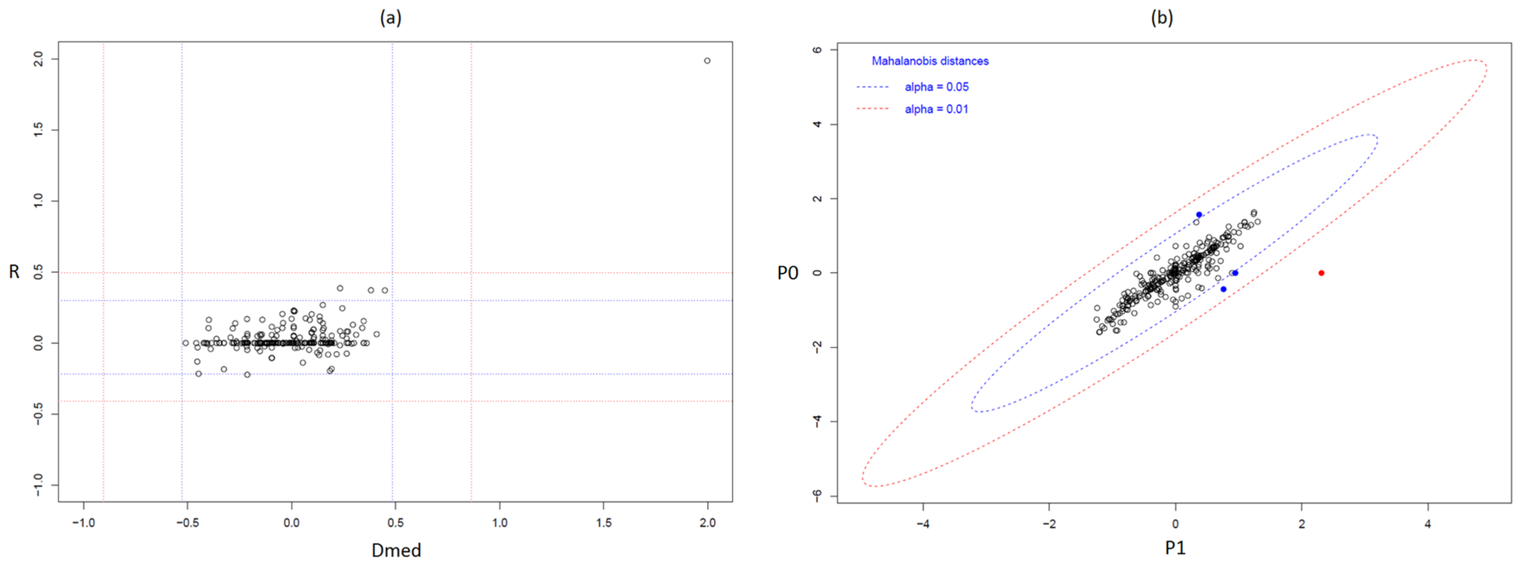

Figure 3: (a) Tolerance regions according the CEA. (b) Tolerance regions given by Mahalanobis distances. The red curve represents the tolerance limits at α=1% while the blue curve represents limits at α=5%. Sample size = 303.

Conclusions

- **The combination of Log and LW** methods proved to be able to overcome problems with skewness and kurtosis.
- **Both methods used** (CEA and Mahalanobis distances) were able to detect the most discrepant point of the distribution.
- **The CEA is more permissive than the Mahalanobis distances**, missing three possible extreme values.
- **Future developments** of the method includes:
  - **A fine-tuning study** need to be performed considering various inputs of the SINAPI, to decide as to whether these fences (α=5%, α=1%, etc) are really the most appropriate.
  - Considerations about products that are geographically characterized.