[1]Plinio Leal dos Santos, Lincoln Teixeira da Silva

[1]plinio.santos@ibge.gov.br

Department of Price Indexes / Directorate of Surveys

Brazilian Institute of Geography and Statistics - IBGE

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

Price indexes 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 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.