Outlier detection for compositional data by municipality and disabled people in Brazil

Paulo Tadeu Meira e Silva de Oliveira

EESC-USP. São Paulo, Brasil

Abstract: Outliers are observations with a unique combination of characteristics identifiable as being different from the others. Disabled people are considered to be those who have physical, hearing, intellectual or sensory disability, factors which, in interactions with various barriers, can obstruct their full and effective participation in society like other people. According to the IBGE in the 2010 Demographic Census, there were 45.6 million disabilities people in Brazil distributed in different municipalities. Data were considered by municipality justified by the fact that the level of service provided by these people varies according to the infrastructure and availability of existing resources in the most diverse locations. The data sets of the 2010 Demographic Census, aggregated by municipality on topics related to disabled people, identification, education, family, work and income, housing occupation, housing conditions, housing basic improvements, other goods, life quality and all for each of the 645 municipalities of the São Paulo state. Compositional data are those that establish the relative information, they are parts of a whole, the sum of these data in each line is a constant, in such a way that it represents 100%. In this work, for compositional data with logarithmic transformation were applied in conjunction with the variable selection procedure in which the variables for each model then, they were applied for detection method Distance from: Mahalanobis, Euclidean, standardized Euclidean, and, score of the first two main components using the corresponding data from the municipalities of the State of São Paulo, and finally, a comparative study will be carried out between the results obtained by these different methods and topics. The objective was the comparative study between these methods for detecting outliers. In this work it was verified that the outliers are more concentrated in work and income between the topics and Pearson's distance between the methods.

Key words: disability people, compositional data, regression analysis, outlier’s data.