Robust detection techniques for multivariate spatial data

Marie Ernst and Gentiane Haesbroeck

Department of Mathematics, University of Liege, Belgium

m.ernst@ulg.ac.be

Abstract

Spatial data are characterized by statistical units, with known geographical positions, on which non spatial attributes are measured. Spatial data are frequently observed in many fields (geography, social sciences,...) and may contain two types of atypical observations: global and/or local outliers (see [2] for more details).

The attribute values of a global outlier are outlying with respect to the values taken by the majority of the data points while the attribute values of a local outlier are extreme when compared to those of its neighbors.

Classical outlier detection techniques may be used to find global outliers as the geographical positions of the data is not taken into account in this search. The detection of local outliers is more complex especially when there are more than one non spatial attribute.

The beginning of this research project focused on two existing techniques respectively presented by Filzmoser *et al.* [1] and Lu *et al.* [3]. These two approaches are based on Mahalanobis or robust distances computed in each neighborhood using a common estimation of the covariance matrix. Two improvements of these techniques are suggested.

Firstly these techniques can classify an observation as local outlier even if its neighbors do not have a similar behavior. A proposal is then to use the local outlier detection only for observations with spatially autocorrelated neighborhood. Therefore some (partial) results will be presented about spatial autocorrelation calculated in a multivariate and robust way.

Secondly imposing the same structure to each neighborhood seems to be quite unrealistic. The choice of a common and global estimation of the covariance matrix in the local distances was justified by the fact that the sample size might be small with respect to the number of variables when the data are restricted to the neighborhood. A natural way to overcome this limitation is to estimate the covariance matrix using a regularized estimator. This yields the second new proposal for detecting local outliers.

In addition to these two ideas based on existing techniques, a non parametric detection technique is proposed. This technique uses depth functions to measure outlyingness. Therefore too small depth values will determine global outliers and local outliers will be detected as the observations whose majority of neighbors are lying in "far" depth contours.

References

[1] Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C., Identification of local multivariate outliers, *Statistical Papers*, (2013), 1–19.

[2] Haslett, J., Brandley, R., Craig, P., Unwin, A. et Wills, G., Dynamic Graphics for Exploring Spatial Data With Applications to Locating Global and Local Anomalies, *The American Statistician*, **45**, (1991), 234-242.

[3] Lu, C.T., Chen, D. and Kou, Y., Multivariate spatial outlier detection, *International Journal on Artificial Intelligence Tools*, **13**, (2004), 801–812.