On Robustness for Spatial Data

A. García-Pérez[1*] and Y. Cabrero-Ortega[2]

[1] *Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040-Madrid, Spain; agar-per@ccia.uned.es*
[2] *C.A. UNED-Madrid, Spain; ycabrero@madrid.uned.es*
[*] *Presenting author*

Keywords. *Robustness;Spatial outliers;GAMs;GIS;Spatial Influence Function.*

1 Identification of local outliers using Robust GAMs and Geographical Information Systems

In the first part of the paper we propose two different methodologies for detecting possible local outliers, that we call *hotspots*. The first one is based in using Geographical Information Systems (GIS) considering a map for the observations where the *heights* of the ground have been replaced by the data $z(s) = z(x, y)$ of the observed variable. We do a TIN interpolation and then, with GRASS through QGIS, we compute the slopes of the triangles thus obtained, concluding with the detection of outlying slopes with GRASS again. Areas with big slopes will indicate zones of possible outliers. This method works with a Big amount of Data. These ideas have been used (with some variants) by Felicísmo [1994].

The second technique consists in fitting a robust Generalized Additive Model (GAM) to the observations. Then we do the previous process (interpolation plus detection of outlying slopes) to the residuals of this robust fit where the Longitude, x and the Latitude, y, are used as covariates in the model. We use QGIS again because the statistical package R can be run inside, obtaining so, layers that can overlapped on other pre-existing ones. This second detecting method has been used in Liu et al. [2001]. Here we extend their ideas considering a more general model, a GAM one, because this is the common model considered in a spatial data fit.

After we have obtained a reduced set of *hotspots*, we compute the probability of obtaining such outlying slopes according to a classical GAM. Those hotspots for which we obtain a small probability, will be labeled as local outliers. We apply these techniques to Guerry data, as Filzmoser et al. [2014] did, obtaining the same

conclusions than they.

2 A Spatial Influence Function

We know that the Hampel's Influence Function IF of a functional T at a model F, is a very useful tool to measure the effect of an outlier z on an estimator (on a functional, really). For instance, the Hampel's influence function of the Mean is the function of z, $IF(z; \text{Mean}, F) = z - \mu$, where $\mu = \int u\, dF(u)$. With this function we see that the outlier z influences the Mean linearly and in an unbounded way. Nevertheless, in this standard definition of the IF, the coordinates of the outlying observation z are omitted and, as we have seen, these could be very important.

Using the three-dimensional notation of this paper, we can rewrite the IF for the Mean as $IF(z(s_0); T, F) = IF(h(x_0, y_0); T, F) = z(s_0) - \mu$, where h is a smooth function used in the GAM fit that is expressed in terms of a basis.

If $z(s)$ is a local outlier and not a global one, it will affect the estimator because, al least, it is a none locally expected value but, because it is not outside of the bulk of the data in the OZ axis, it will pass unnoticed for the Hampel's Influence Function IF, i.e, the IF does not measure the influence of local outliers that are not global, on estimators (or functionals) because these local outliers act on the x-y plane and not on the OZ axis. Nevertheless, they are influential observations and they do affect the value of the estimator.

We need, then, to modify (extend) the IF to take into account both local and global outliers because, all of them affect the estimator. This new influence function will be called *Spatial Influence Function*, SIF. It is defined using the ideas of the previous section and its main properties studied.

References

Felicísmo, A.M. (1994). Parametric statistical method for error detection in digital elevation models. *ISPRS Journal of Photogrammetry and Remote Sensing*, **49**, 29–33.

Filzmoser P., Ruiz-Gazen, A. & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, **55**, 29–47.

Liu, H., Jezek, K.C. & O'Kelly, M.E. (2001). Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. *International Journal of Geographical Information Science*, **15**, 721–741.