

Spatial Outlier Detection Using GAMs and Geographical Information Systems

Alfonso García-Pérez and Yolanda Cabrero-Ortega

Abstract A spatial (local) outlier is a value that differs from its neighbors. The usual way in which these are detected is a complicated task, especially if the data refer to many locations. In this paper we propose a different approach to this problem that consists in considering outlying slopes in an interpolation map of the observations, as indicators of local outliers. To do this, we transfer geographical properties and tools to this task using a Geographical Information System (GIS) analysis. To start, we use two completely different techniques in the detection of possible spatial outliers: First, using the observations as *heights* in a map and, secondly, using the residuals of a robust Generalized Additive Model (GAM) fit. With this process we obtain areas of possible spatial outliers (called hotspots) reducing the set of all locations to a small and manageable set of points. Then we compute the probability of such a big slope at each of the hotspots after fitting a classical GAM to the observations. Observations with a very low probability of such slope will finally be labelled as spatial outliers.

1 Introduction. Spatial Outliers

A local or spatial outlier [3] or [6] is an observation that differs from its neighbors, i.e., $z(s_0)$, the value of the variable of interest Z at location s_0 , is a local outlier if it differs from $z(s_0 + \Delta s_0)$ where Δs_0 defines a neighborhood of location s_0 .

The usual method used to detect local outliers is somewhat complicated because, first, we have to define what is a neighborhood, i.e., what is “close”; then, we have to select some locations inside the neighborhood, to compute and compare the value of Z at these locations.

A. García-Pérez (✉)

Departamento de Estadística, I.O. y C.N., Universidad Nacional
de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040 Madrid, Spain
e-mail: agar-per@ccia.uned.es

Y. Cabrero-Ortega

C.A. UNED-Madrid, Madrid, Spain
e-mail: ycabrero@madrid.uned.es

In the first part of the paper we propose two novel techniques based on a GIS for easily and quickly detect possible local outliers. The first one, developed in Sect. 2, is based on making a geographical map where the *heights* of the ground correspond to the observations. This map of separate heights is completed by means of a Triangulated Irregular Network (TIN) interpolation. Once the geographical map has been made, local outliers are easily identified as hills with big slopes.

The second technique, developed in Sect. 3, consists in fitting a robust GAM to the observations. Then, we do the previous process (interpolation plus detection of outlying slopes) with the residuals of this robust fit.

These ideas have been previously used (with some variants) in [5, 10, 12]. Here we extend their ideas considering a more general model, a GAM one, because this is the model usually considered in a fit of spatial data.

Once identified possible local outliers, we compute, in Sect. 4, the probability of such an extreme slope according to a model fitted to the data. If, according to this model (i.e., assuming that the model is correct), the probability of such extreme slope is small, the hotspot is labelled as a local outlier.

2 Spatial Outlier Detection by Interpolation

We propose, first, to interpolate the observations $z(s_i)$ using a TIN interpolation, that is implemented in Quantum GIS (QGIS), and that essentially means to interpolate the observations with triangles. Then we use the Geographic Resources Analysis Support System (GRASS) to compute the slopes of all the triangles obtained with the previous TIN interpolation. Finally, we reclassify the slopes, using GRASS grass again, looking for outlying slopes. All locations with big slopes will be considered as hotspots, i.e., potential outliers.

Other interpolation procedures could be used, such as Inverse Distance Weighting (IDW), but TIN works well for data with some relationship to other ones across the grid, that should be the kind of data usually considered in a spatial data problem, [8].

2.1 Multivariate Spatial Outliers

If we have multivariate observations, we first transform them into the scores obtained from a Principal Component Analysis PC_1, \dots, PC_p . With this process, similar to Principal Components Regression Analysis, we can apply the previous QGIS method to each one dimensional independent variable, PC_i , obtaining so p layers of hotspots (one layer for each PC_i). The intersection of all of them will be the set of possible multivariate outliers. Moreover, in this way we also have a marginal analysis for each univariate variable.

Example 1 Let us consider Guerry data, [9], available in the R package with the same name. This data set has been analyzed in [6] and, as there, here we only use 85 departments, excluding Corsica. The two variables considered are also “population per crime against persons” (PER) and “population per crime against property” (PROP).

As we mentioned before, the descriptive process of detection of possible outliers, i.e., hotspots, consists in using QGIS, (a) incorporating first into QGIS the vectorial data, *france1.txt*, of the scores, after transforming the original observations with the two Principal Components PC_1 and PC_2 ; (b) computing a TIN interpolation for each new variable PC_1 and PC_2 ; (c) computing with GRASS the slopes from a Digital Elevation Model (DEM); (d) using again GRASS to reclassify slopes in two groups: small slopes and big slopes.

The details of the computations of all the examples in the paper are at <http://www.uned.es/pfac-estadistica-aplicada/smps.htm>.

In these computations, we obtain for PC_1 a plot (and a table) of departments with slopes higher than 30 % and, for PC_2 , slopes higher than 19 %. The intersection of both layers is showed in Fig. 1 where the outlying slopes (the unfilled circles) correspond to the departments Ain, Ardeche, Correze, Creuse, Indre, Isere, Jura, Loire, Rhone, Saone-et-Loire and Haute-Vienne.

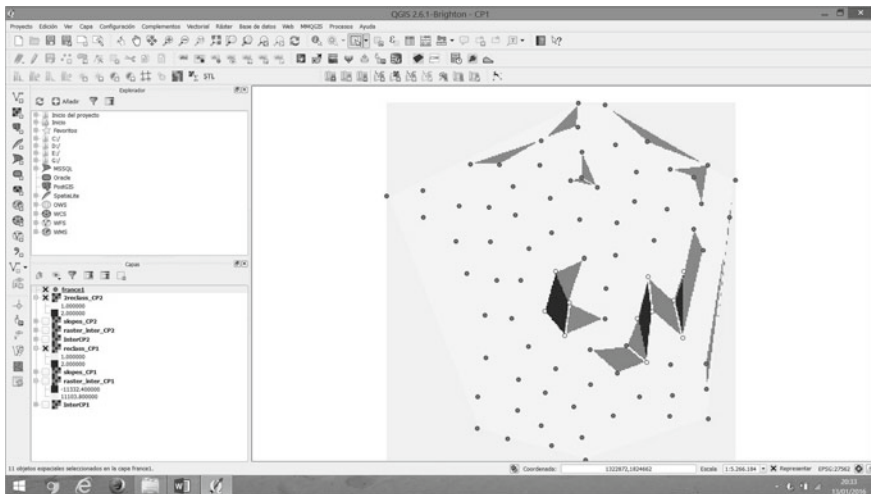


Fig. 1 Slopes reclassification (PC_1 and PC_2)

3 Spatial Outlier Detection by a Robust GAM

The method proposed in the previous section is an exploratory technique based only on a GIS. In this section we propose to fit a robust GAM to the spatial observations $z_i = Z(s_i)$. In this way, local large residuals will give us possible spatial outliers. We consider a GAM because this type of models is generally used for modeling spatial data.

With a GAM, [11], we assume that (univariate) observations are explained as

$$z_i = h(s_i) + h(u_{1i}) + \dots + h(u_{ki}) + e_i \tag{1}$$

where $s_i = (x_i, y_i)$ are the coordinates of z_i ; $u = (u_1, \dots, u_k)$ is a vector of covariates, and h is a smooth function that is expressed in terms of a basis $\{b_1, \dots, b_q\}$ as

$$h(u) = \sum_{j=1}^q b_j(u)\beta_j \tag{2}$$

for some unknown parameters β_j ([15], pp. 122). The errors e_i must be, as usual, i.i.d. $N(0, \sigma)$ random variables.

A key point in our proposal is to consider the coordinates $s_i = (x_i, y_i)$ of the observations z_i as a covariate in model (1).

The function h could be different for each covariate and, in some cases, the coordinates covariate is split into two covariates being the model

$$z_i = h_1(x_i) + h_2(y_i) + h_3(u_{1i}) + \dots + h_{k+2}(u_{ki}) + e_i.$$

We can summarize model (1) as $z_i = H(s_i, u_{1i}, \dots, u_{ki}) + e_i$. This approach extends the ideas of [7] because they consider (pp. 52) a linear regression model. Also, some aspects of the papers [12] or [5] are extended in this way.

The robust GAM that we shall fit is the model proposed in [13, 14] although other possible robust GAMs could be the proposed in [1] or [4].

The robust M -type estimators $\hat{\beta}$ for the GAM proposed by Wong are the solution of the following system of estimating equations

$$\sum_{i=1}^n \left[w(\mu_i) \nu(z_i, \mu_i) \boldsymbol{\mu}'_i - a(\boldsymbol{\beta}) - \frac{1}{n} \mathbf{S}\boldsymbol{\beta} \right] = \mathbf{0}$$

where

$$\mu_i = E[z_i | \mathbf{u}_i]; \boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^t; \boldsymbol{\mu}'_i = \partial \mu_i / \partial \boldsymbol{\beta}; \nu(z_i, \mu_i) = (z_i - \mu_i) / V(\mu_i);$$

$$w(\mu_i) = \frac{1}{E[\varphi'_c((z_i - \mu_i) / V^{1/2}(\mu_i))]}$$

$$a(\beta) = \frac{1}{n} \sum_{i=1}^n E_{z_i | \mathbf{u}_i} [\nu(z_i, \mu_i)] w(\mu_i) \mu_i'$$

φ_c the Huber-type function with tuning constant c , and $\mathbf{S} = 2\lambda\mathbf{D}$, being λ a smoothing parameter and \mathbf{D} a pre-specified penalty matrix.

The previous system of estimating equations, hence, is formed by the robust quasi-likelihood equations introduced in [2], plus the usual penalized GAM part.

After we have a good fit, the residuals of this fit, i.e., the differences between the observed and the predicted values, will help us to detect possible spatial outliers. To do this we compute the residuals (or the scores of the residuals if $\mathbf{z}_i(s_0)$ is multivariate), we incorporate them into QGIS and we follow the same process than in the previous section: A TIN interpolation, the slopes obtained with GRASS and, finally, a reclassification with GRASS looking for outlying slopes.

Example 2 Let us consider Guerry data again, [9]. We first fit a robust GAM [13, 14] for each dependent variable, PER and PROP, and we compute the residuals for each fit. We then compute the scores of these residuals and, again with QGIS, we obtain departments with slopes both, higher than 30% for PC_1 and higher than 13% for PC_2 , Fig. 2. The hotspots obtained correspond to the departments Hautes-Alpes, Ardeche, Creuse, Indre, Loire, Rhone, Saone-et-Loire, Seine and Haute-Vienne.

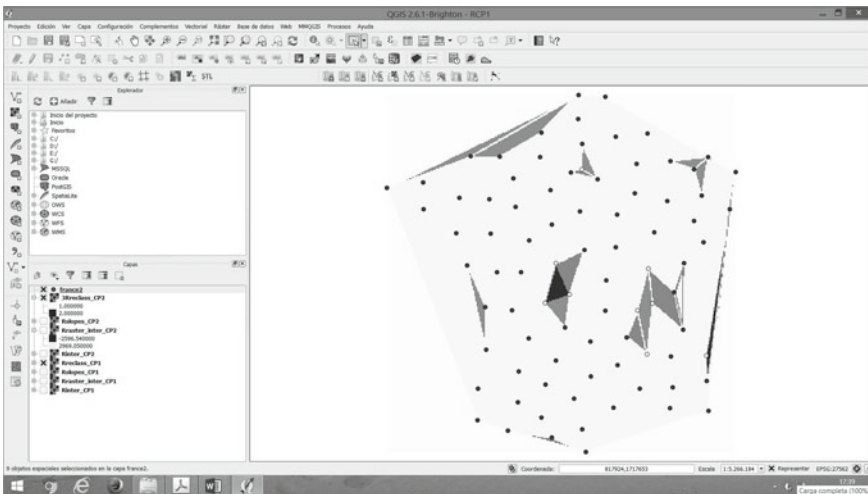


Fig. 2 Slopes reclassification of the scores of the residuals (PC_1 and PC_2)

4 Identification of Spatial Outliers

With the procedures considered in the two previous sections we obtain a set of possible local outliers. In this section we compute, mathematically, if the behavior around a hotspot is very unlikely or not to label it as an actual spatial outlier, computing the probability of obtaining an slope as big as the one obtained at a given location s_0 . Considering the framework of the last section, a large (positive or negative) slope, i.e., a large *derivative* of function H (h in fact) at s_0 will give us a good idea if $z(s_0)$ is a local outlier or not.

To compute the probabilities of large slopes at the hotspots previously identified, we first fit a classical GAM. We consider now a classical GAM fit instead of a robust one to magnify their slopes because the classical model will be more sensitive than the robust and the slopes less soft. Also because we know the (asymptotic) distribution of the estimators of the parameters in a classical GAM but not in the robust one.

From a mathematical point of view, the slope at a point s_0 in the direction v is stated as the directional derivative along v (unit vector) at s_0 .

If we represent, as usual, by $D_v h(s_0)$ the collection of directional derivatives of function h (assuming that it is differentiable) along all directions v (unit vectors) at s_0 and by MS the *maximum slope*, i.e., $MS(s_0) = \sup_v |D_v h(s_0)|$, we compute the probability of obtaining the observed maximum slope $ms(s_0)$, i.e., $P\{MS(s_0) \geq ms(s_0)\}$. If this probability is low (for instance lower than 0.05), we shall label $z(s_0)$ as a local outlier (more formally, we could say that we are rejecting the hypothesis of being zero the slope at s_0 , i.e., that $z(s_0)$ is not a local outlier) and, as the smaller the probability, the greater should be considered $z(s_0)$ as a local outlier.

Because we have assumed that the smooth function h has a representation in terms of a basis, (2), the slope will depend on the estimators of the parameters β_j , estimators that are approximately normal distributed ([15], pp. 189) if the z_i are normal.

From vector calculus, we known that the largest value for the slope at a location s_0 is gradient norm, i.e.,

$$MS(s_0) = \sup_v |D_v h(s_0)| = \|\nabla h(s_0)\| = \sqrt{\left(\frac{\partial}{\partial x} h(x, y)\Big|_{s_0}\right)^2 + \left(\frac{\partial}{\partial y} h(x, y)\Big|_{s_0}\right)^2}$$

and because h is expressed in term of a basis, the probability that we have to compute refers to the random variable

$$\sqrt{\left(\sum_{j=1}^q \frac{\partial}{\partial x} b_j(s_0) \cdot \hat{\beta}_j\right)^2 + \left(\sum_{j=1}^q \frac{\partial}{\partial y} b_j(s_0) \cdot \hat{\beta}_j\right)^2} \tag{3}$$

If this is low, $z(s_0)$ will be labelled as a local outlier.

4.1 Cubic Regression Splines

We shall use a cubic regression splines to explain function h in the fit of a GAM to the observations z_j . For this aim we shall use the R function `gam` of the R package `mgcv`. The cubic spline function, with k knots v_1, \dots, v_k , that we fit ([15], pp. 149–150) is $(v_j \leq v \leq v_{j+1})$

$$P(v) = \frac{v_{j+1} - v}{h_j} \beta_j + \frac{v - v_j}{h_j} \beta_{j+1} + \left[\frac{(v_{j+1} - v)^3}{h_j} - h_j (v_{j+1} - v) \right] \frac{\delta_j}{6} + \left[\frac{(v - v_j)^3}{h_j} - h_j (v - v_j) \right] \frac{\delta_{j+1}}{6}$$

where $h_j = v_{j+1} - v_j, j = 1, \dots, k - 1$ and $\delta_j = P''(v_j)$.

The first derivative of P (partial derivative in formula (3)) is

$$P'(v) = \frac{\beta_{j+1} - \beta_j}{h_j} + \left[-\frac{3(v_{j+1} - v)^2}{h_j} + h_j \right] \frac{\delta_j}{6} + \left[\frac{3(v - v_j)^2}{h_j} - h_j \right] \frac{\delta_{j+1}}{6}$$

and considering as knots the locations, v_j ,

$$P'(v_j) = \frac{\beta_{j+1} - \beta_j}{h_j} - \frac{\delta_j h_j}{3}.$$

If the term $\delta_j h_j / 3$ is negligible, we have to compute the probabilities,

$$P \left\{ (\hat{\beta}_{j+1} - \hat{\beta}_j) / h_j > \text{observed slope} \right\}$$

based on a normal model because ([15], pp. 189) $\hat{\beta}_j$ is approximately normal distributed with mean β_j .

Table 1 Probability of a big slope for both variables

		Probability	
Dept	Department	PER	PROP
5	Hautes-Alpes	0.08677979	0.734663
1	Ain	0.7796545	0.9039119
7	Ardeche	0.08590459	0.5845837
19	Correze	0.8543756	0.968079
23	Creuse	0.3344432	0.8536806
36	Indre	0.8043197	0.9364876
38	Iserre	0.2926037	0.7874324
39	Jura	0	0.0062001
42	Loire	0.5497284	0.8805521
69	Rhone	0	0.365532
71	Saone-et-Loire	0.45913	0.8109866
75	Seine	0	0
87	Haute-Vienne	0.01982465	0.6981038

Example 3 Let us consider Guerry data again. The set of all departments detected as possible outliers for, at least, one of the two methods explained in Sects. 2 and 3, together with the probabilities of such slopes (i.e., the p-values of the bilateral test of the null hypothesis $H_0 : \beta_{j+1} - \beta_j = 0$), are in Table 1.

Hence, we can label as spatial outliers the observations at Jura, Rhone and Seine. As is remarked in [6], Seine (together with Ain, Haute-Loire and Creuse) is a global outlier and a local one.

Hence, if we do not consider the Department of Seine (because is a global outlier) we have two departments that can be considered as spatial outliers: Jura and Rhone, two departments in what is called the Rhône-Alpes area, i.e., the same result than in [6].

Acknowledgment This work is partially supported by Grant MTM2012-33740.

References

1. Alimadad A, Salibian-Barrera M (2011) An outlier-robust fit for generalized additive models with applications to disease outbreak detection. *J Am Stat Assoc* 106:719–731
2. Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. *J Am Stat Assoc* 96:1022–1030
3. Cressie NAC (1993) *Statistics for spatial data*. Wiley, New York
4. Croux C, Gijbels I, Prosdocimi I (2012) Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics* 68:31–44
5. Felicísimo AM (1994) Parametric statistical method for error detection in digital elevation models. *ISPRS J Photogramm Remote Sens* 49:29–33
6. Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C (2014) Identification of local multivariate outliers. *Stat Papers* 55:29–47
7. Fotheringham AS, Brunson C, Charlton M (2002) *Geographically weighted regression. The analysis of spatially varying relationships*. Wiley, New York
8. Franke R (1982) Scattered data interpolation: tests of some methods. *Math Comput* 38:181–200
9. Guerry A-M (1833) *Essai sur la statistique morale de la France*. Crochard, Paris. English translation: HP Whitt and VW Reinking, Edwin Mellen Press, Lewiston, 2002
10. Hannah MJ (1981) Error detection and correction in digital terrain models. *Photogramm Eng Remote Sens* 47:63–69
11. Hastie T, Tibshirani R (1990) *Generalized additive models*. Chapman & Hall, London
12. Liu H, Jezek KC, OKelly ME (2001) Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. *Int J Geogr Inf Sci* 15:721–741
13. Wong KW (2010) *Robust Estimation for Generalized Additive Models*. MA Thesis, Department of Statistics, The Chinese University of Hong Kong
14. Wong RKW, Yao F, Lee TCM (2014) Robust estimation for generalized additive models. *J Comput Graph Stat* 23:270–289
15. Wood SN (2006) *Generalized additive models: an introduction with R*. Chapman & Hall/CRC Press