Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 2, 2020 (New Series), pp 155-162

Identification of Multiple Unusual Observations in Spatial Regression

A.H.M. Rahmatullah Imon[1] and Ali S Hadi[2]
[1]*Department of Mathematical Sciences, Ball State University, USA*
[2]*Department of Mathematics and Actuarial Science*, *The American University in Cairo,Egypt*

Received: 24 February 2020; Revised: 11 June 2020; Accepted: 10 July 2020

Abstract

Traditional outlier detection methods cannot be directly applied to spatial data because of its global nature. Spatial outlier detection methods concentrate on discovering neighborhood instabilities (Shekhar *et al*. 2002). However, most of the traditional detection methods may not accurately locate outliers when multiple outliers exist. Robust spatial *z* test proposed by Hadi and Imon (2018) has largely resolved this issue. But lots of unresolved issues exist in spatial regression where likewise linear or generalized linear models, the entire inferential procedure is generally affected in the presence of unusual observations called outliers (*y*-outliers) and high leverage points (*x*-outliers) or both. A large body of literature are available now for the identification of unusual observations in linear and/or generalized linear regression but this is still an unexplored area in spatial regression. In this paper we propose a new method for the identification of multiple spatial outliers and spatial high leverage points based on robust and clustering algorithms. We also propose a very simple but attractive graphical display to locate these two types of outliers in the same graph.

Keywords: Spatial outlier; Differencing; Masking; High leverage points; Clustering; GP-GSR plot.

1. Introduction

Conceptually spatial outliers are very different from classical outliers. A commonly used definition is that outliers are a minority of observations in a dataset that have different patterns from that of the majority of observations in the dataset. The assumption here is that there is a core of at least 50% of observations in a dataset that are homogeneous (that is, represented by a common pattern) and the remaining observations (hopefully few) have patterns that are inconsistent with this common pattern. Spatial outliers are those observations whose characteristics are markedly different from their spatial neighbors. The identification of spatial outliers is important because it can reveal hidden but valuable knowledge in many applications such as identifying aberrant genes or tumor cells, discovering highway traffic congestion points, locating extreme meteorological events such as tornadoes, and hurricanes *etc.*

Corresponding Author: A.H.M.R. Imon
Email: rimon@bsu.edu

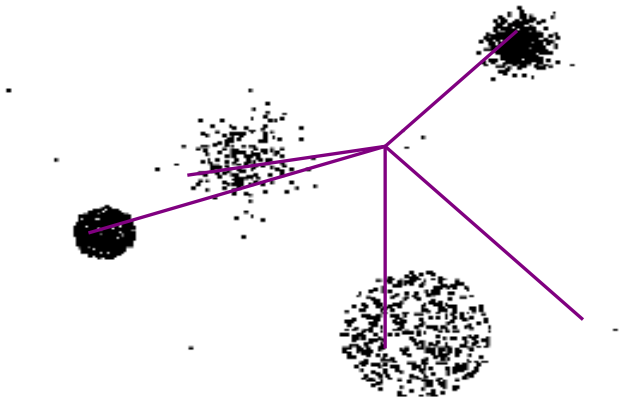


Figure 1: Outliers in data clusters

Although outliers could be easily identified in univariate, bivariate, or even trivariate data through graphical examination of the data, visual inspection does not usually work for more than three dimensions. Not only that automated identification of outliers is tricky even for a two dimensional data if the data form clusters as shown in Figure 1. Here the idea of majority minority simply does not work, bad clusters are identified as outliers (Hadi *et al*., 2009) based on classification techniques. Things could even be cumbersome in regression models where outliers can occur along the *y*-dimension, or along the *x*-dimension, or both and/or among the relationship between *x* and *y*. An excellent review of different aspects of spatial outliers is available in Shekhar *et al*. (2002) and Hadi and Imon (2018). Conceptually, spatial outliers match with outliers in big data and for this reason outlier detection techniques designed for big data are often routinely employed in spatial data. In big data the concept of outlier is local, not global so as in spatial data. The distance and/or density based methods such as *k*–nearest neighbourhood, local outlier factor (LOF), spatial outlier factor (SOF) methods have become more popular. But all these methods are designed to identify outliers along the *y*-axis and hence are not readily applicable for spatial regression. For example, temperatures and amount of rainfall of different regions may vary due to their distances from sea or mountain. Once we fit this relationship by regression we may observe not only strange temperature or rainfall pattern, the distance factor may also be unusual. Attempts have been made to identify outliers based on residuals but it only focuses on the outliers in *y*, but not in *x* or both and the whole concept is rather global than local. To overcome this problem in this paper we propose a method which not only focuses on both *x* and *y* dimensions at the same time, but also considers classification techniques to identify outliers.

2. Methodology

Let us assume that we have *n* pairs of spatial observations (u_i, v_i), $i = 1, 2, \ldots, n$. We further assume that *V* depends on *U* and we are interested to investigate their nature of relationship. In order to understand whether spatial observations are stable in their neighborhood, Shekhar *et al*. (2002) suggested considering the first order differences of the spatial observations. For both *U* and *V* we take the first order differences defined as

$$x_i = u_i - u_{i-1}, y_i = v_i - v_{i-1}; i = 2, 3, \ldots, n \quad (1)$$

Based on the differenced observations obtained in (1), let us consider a standard regression model

$$Y = X\beta + \varepsilon \quad (2)$$

where Y is a vector of observed responses of order $(n-1)$, X is an $(n-1) \times 2$ matrix of explanatory variables including the constant, β is a vector of unknown finite parameters of order 2 and ε is an n-vector of random disturbances with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 \mathrm{I}$. The traditionally used ordinary least squares (OLS) estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the vector of fitted values is $\hat{Y} = X\hat{\beta} = HY$. The matrix

$$H = X(X^T X)^{-1} X^T \quad (3)$$

is often referred to as weight or leverage matrix whose diagonal elements h_{ii} are termed leverages. The OLS residual vector $\hat{\varepsilon}$ is defined as $\hat{\varepsilon} = Y - \hat{Y}$. Observations corresponding to exceptionally large $\hat{\varepsilon}$ values are termed outliers. However, the question still remains how large is large? For this reason we often consider the standardized version of residuals. One very popular choice is deleted Studentized residuals (DSR) defined as

$$t_i = \frac{y_i - x_i^T \hat{\beta}^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}},\ i = 2, 3, \ldots, n \quad (4)$$

where $\widehat{\beta}^{(-i)}$ and $\hat{\sigma}_{(i)}$ are the OLS estimates of β and σ respectively with the i-th observation deleted. We call an observation outlier when its corresponding deleted Studentized residual value exceeds 3 in absolute value. Observations corresponding to exceptionally large h_{ii} values are termed high leverage points which are essentially outliers in the X-space. However, since residuals are also functions of leverages, it is better if we identify both outliers and high leverage points simultaneously rather than separately. Gray (1986) proposed the Leverage-Residual (L-R) plot where the leverage value h_{ii} for each observation i, is plotted against the square of a normalised form of its corresponding residual. The bulk of the cases will be associated with low leverage and small residuals so that they cluster near the origin (0, 0). The unusual cases will have either high leverages or large residual components and so will tend to be separated from the bulk of the data. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located in the area to the right.

The L-R plot may be effective in the identification of single outlier but it may be ineffective in the presence of multiple outliers unless we remove a group of suspect outliers prior to fitting the model. Denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Also suppose that R contains $(n-1-d)$ cases after $d<(n-1-k)$ cases in D are deleted. Without loss of generality, assume that these observations are the last d rows of X and Y so that we can partition the matrices as

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix},\quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix},\ H = \begin{bmatrix} H_R & H_{RD} \\ H_{DR} & H_D \end{bmatrix} \quad (5)$$

where $H_R = X_R(X^T X)^{-1} X_R^T$ and $H_D = X_D(X^T X)^{-1} X_D^T$ are symmetric matrices of order $(n-1-d)$ and d respectively, and $H_{RD} = X_R(X^T X)^{-1} X_D^T$ is an $(n-1-d) \times d$ matrix. However, $(X_R^T X_R)^{-1}$ can be expressed as

$$(X_R^T X_R)^{-1} = (X^T X - X_D^T X_D)^{-1} = (X^T X)^{-1} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} (X^T X)^{-1} \quad (6)$$

where I_D is an identity matrix of order d and $U_D = X_D (X_D^T X_D)^{-1} X_D^T$. Using (6), Imon (2002) defined a group deleted version of high leverage points called generalized potentials defined as

$$p_{ii}^* = \begin{cases} \frac{h_{ii}^{(-D)}}{1-h_{ii}^{(-D)}} & i \in R \\ h_{ii}^{(-D)} & i \in D \end{cases} \quad (7)$$

where $h_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i$, i= 2, 3, …, n. In other words, $h_{ii}^{(-D)}$ is the i-th diagonal element of $X(X_R^T X_R)^{-1} X^T$ matrix. The vector of estimated parameters after the deletion of d observations, denoted by $\hat{\beta}^{(-D)}$, is obtained using (6) as

$$\hat{\beta}^{(-D)} = (X_R^T X_R)^{-1} X_R^T Y_R = \hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D \quad (8)$$

where $\hat{\varepsilon}_D = X_D \hat{\beta}$. Using (6), (7) and (8), Imon (2005) introduced a group deleted version of residuals called generalized Studentized residuals (GSR) defined as

$$t_{ii}^* = \begin{cases} \frac{y_i - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)} \sqrt{1-h_{ii}^{(-D)}}} & i \in R \\ \frac{y_i - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)} \sqrt{1-h_{ii}^{(-D)}}} & i \in D \end{cases} \quad (9)$$

where $\hat{y}_i^{(-D)} = x_i^T \hat{\beta}^{(-D)}$ and $\hat{\sigma}^{(-D)}$ are the fitted values of y and the scale parameter σ respectively after the omission of the suspected outlier group indexed by D. Although the expression of generalized potentials is available for any arbitrary set of deleted cases, D, the choice of such a set is clearly important since the omission of this group determines the weights for the whole set. We call an observation outlier when its corresponding generalized Studentized residual value exceeds 3 in absolute value. No such value exists for generalized potentials. We follow Hadi (1992) to declare an observation as a high leverage point if its corresponding p_{ii}^* exceeds a threshold given as

$$p_{ii}^* > \text{Median}\,(p_{ii}^*) + 3\text{MAD}\,(p_{ii}^*). \quad (10)$$

where MAD stands for the median absolute deviation.

These above results enable us to define a simple graphical display of classifying group deleted leverages and residuals for possible identification of them. Generalized potentials are used as leverages and the generalized Studentized residuals as deletion residuals in a 'generalized potentials –generalized Studentized residuals (GP-GSR)' plot. Since the high leverage points need not to be outliers and outliers may not be points of high leverage we may expect different deletion sets D from the computation of these two quantities. Since D is the group of suspected outliers we prefer to include all observations considered to be suspect either along the y dimension or along the x dimension. We employ the blocked adaptive computationally-efficient outlier nominators (BACON) proposed by Billor *et al.* (2000) as a classifier. Another possibility could be the application of support vector regression for the same, especially when the data is big. The main advantage of the GP-GSR plot is that it is suitable for the data where masking (false negative)

and/or swamping (false positive) make single case diagnostic plots misleading. This plot, unlike the L-R plot retains the signs of residuals, which can be very important when their interpretation is concerned. Since the bulk of the cases will be associated with low leverage and small residuals, most of the pairs (t_{ii}^*, p_{ii}^*) will cluster near the origin (0, 0). The unusual cases will have either high leverages or large residual components and will tend to be separated from the bulk of the cases. High leverage cases will be located at the right corner of the plot and observations with large residuals will be located either at the upper or lower corner of the plot depending on their signs; large positive outliers will be plotted at the upper corner and large negative outliers will be located at the bottom corner of the plot.

3. Results

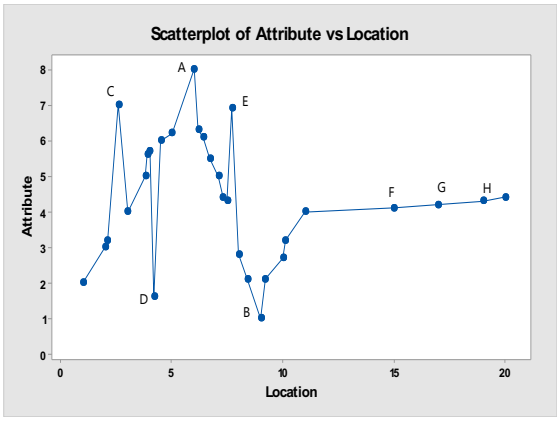
In this section we would like to present an example to demonstrate how our proposed method works in the classification of spatial regression outliers in both *x* and *y* dimensions. Here we consider a spatial outlier data given by Hadi and Imon (2018) extending the idea of Shekhar *et al*. (2002). Although this data is artificial in nature, the use of this type of data is very common in the outlier detection literature (Rousseeuw and Leroy, 1987; Hadi *et al*., 2009) because here we definitely know which observations are genuine outliers. For real data with multiple outliers due to masking and swamping there could be always lots of disagreements regarding which observations are genuine outliers or not. We present the data in Table 1 and also in Figure 2.

Table 1: Hadi and Imon (2018) spatial outlier data

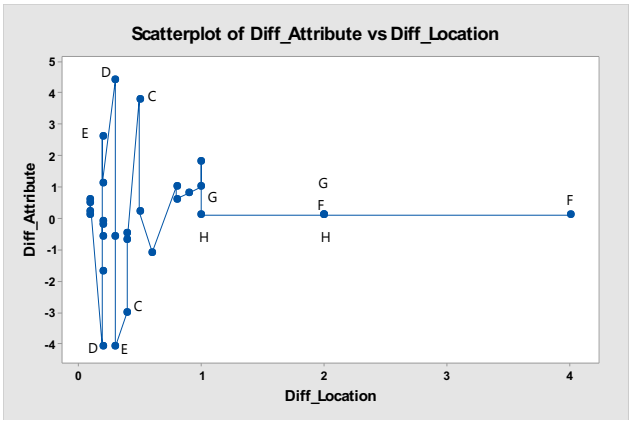
Index	Location	Attribute	Diff_Location	Diff_Attribute
1	1.0	2.0	*	*
2	2.0	3.0	1.0	1.0
3	2.1	3.2	0.1	0.2
4	2.6	**7.0 C**	0.5	**3.8 C**
5	3.0	4.0	0.4	**–3.0 C**
6	3.8	5.0	0.8	1.0
7	3.9	5.6	0.1	0.6
8	4.0	5.7	0.1	0.1
9	4.2	**1.6 D**	0.2	**–4.1 D**
10	4.5	6.0	0.3	**4.4 D**
11	5.0	6.2	0.5	0.2
12	6.0	8.0 A	1.0	1.8
13	6.2	6.3	0.2	–1.7
14	6.4	6.1	0.2	–0.2
15	6.7	5.5	0.3	–0.6
16	7.1	5.0	0.4	–0.5
17	7.3	4.4	0.2	–0.6
18	7.5	4.3	0.2	–0.1
19	7.7	**6.9 E**	0.2	**2.6 E**
20	8.0	2.8	0.3	**–4.1 E**
21	8.4	2.1	0.4	–0.7
22	9.0	1.0 B	0.6	–1.1
23	9.2	2.1	0.2	1.1
24	10.0	2.7	0.8	0.6
25	10.1	3.2	0.1	0.5

26	11.0	4.0	0.9	0.8
27	**15.0 F**	4.1	**4.0 F**	0.1
28	17.0	4.2	2.0	0.1
29	19.0	4.3	2.0	0.1
30	20.0	4.4	1.0	0.1

This example gives a clear distinction between classical outlier and spatial outlier. In Figure 2(a) attribute values are plotted against their locations. For global outliers, traditional statistics will essentially look at the attribute values in the *y* axis and if we do that we observe that the points which are very high such as A or very low such as B. In contrast to that, the spatial outliers are like the spikes C, D and E. They look like spatial outliers because they violate the law of geography that the nearby things should be very similar. When we take the first order difference of the attributes as shown in Figure 2(b) clearly C, D and E look very different than their neighbors. It is also interesting to note that the possible global outliers A and B do not look like outliers anymore. In general, we do not search for outliers along the *x*-axis. But when we carefully look at Figure 2(a), we observe that the point F has a marked difference from its neighbors. Points G and H look unusual too. This difference is visible more clearly when we look at the first order difference of the locations as shown in Figure 2(b). Point F now clearly looks like a high leverage point or an outlier along the *x*-space. Points G and H look more extreme as well.



2(a). The original data



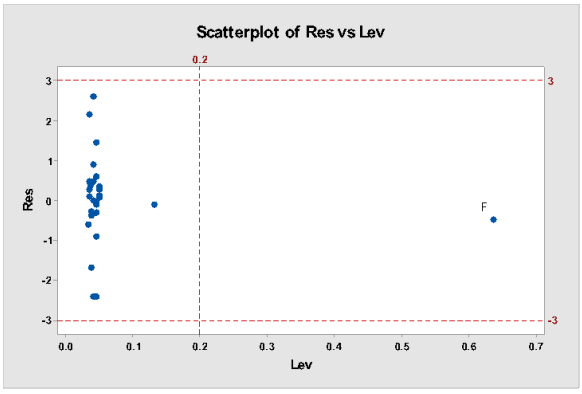
2(b) The first order differenced data

Figure 2: Scatter plot of the original and the first order differenced data

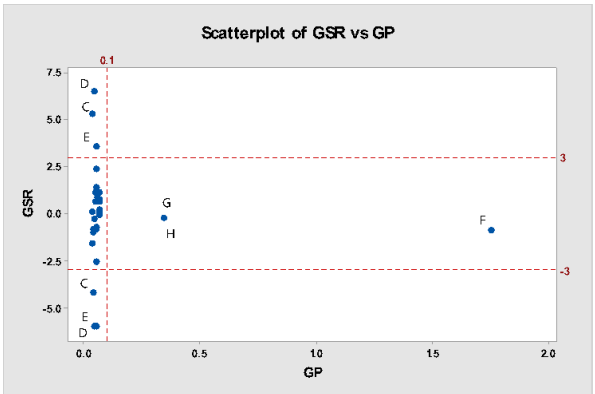
Now we run a spatial regression of attributes on locations. Since our interest is to understand the neighbourhood instability we consider the first order difference of attributes and locations as given in columns 4 and 5 of Table 1. We then run a regression of differences in attributes on differences in location and the resulting deleted Studentized residuals and leverages are given in columns 2 and 3 of Table 2. Although DSR is very popular outlier measure it fails to identify even a single observation as an outlier. Here the cut-off for the leverage is 0.2 and it can identify F as a high leverage point. We see exactly the same picture in the L-R plot as shown in Fig 3(a). Now we compute GSR and GP and the results are presented in columns 4 and 5 of Table 2. We use BACON classifier to obtain the D set first and then compute GSR and GP as outlined in equations (7) and (8). It is worth mentioning that the cut-off value for GP is 0.1 based on equation (10). We also present the GP-GSR plot for this data in Figure 3(b). These results clearly show the merit of our proposed method. It can successfully identify 3 spatial outliers (C, D and E) and 3 spatial high leverage points(F, G, H).

Table 2: Residuals and leverages for the spatial outlier data

Index	Del St. Residual	Leverage	GSR	GP
1	*	*	*	*
2	0.45678	0.040885	1.09925	0.06658
3	0.11424	0.051079	0.20420	0.06290
4	2.15139	0.035779	**5.31765 C**	0.03590
5	–1.69711	0.037989	**–4.20445 C**	0.03835
6	0.47421	0.035612	1.13209	0.04571
7	0.32834	0.051079	0.72348	0.06290
8	0.06085	0.051079	0.07645	0.06290
9	–2.41622	0.045639	**–6.03394 D**	0.05185
10	2.61223	0.041275	**6.49375 D**	0.04367
11	0.07639	0.035779	0.07645	0.03590
12	0.89298	0.040885	0.13783	0.06658
13	–0.92108	0.045639	2.34566	0.05185
14	–0.10826	0.045639	–2.56908	0.05185
15	–0.33030	0.041275	–0.32208	0.04367
16	–0.28573	0.037989	–0.85865	0.03835
17	–0.32172	0.045639	–0.74085	0.05185
18	–0.05503	0.045639	–0.84428	0.05185
19	1.43538	0.045639	**3.58453 E**	0.05185
20	–2.42202	0.041275	**–6.02091 E**	0.04367
21	–0.39244	0.037989	–1.00843	0.03835
22	–0.62533	0.034647	–1.61273	0.03631
23	0.58737	0.045639	1.39415	0.05185
24	0.26077	0.035612	0.59882	0.04571
25	0.27470	0.051079	0.59176	0.06290
26	0.35834	0.037710	0.84471	0.05471
27	–0.49040	**0.636897 F**	–0.91806	**1.75404 F**
28	–0.12121	0.131865	–0.24012	**0.34271 G**
29	–0.12121	0.131865	–0.24012	**0.34271 H**
30	–0.02276	0.040885	–0.06854	0.06658



3(a). L-R plot



3(b). GP-GSR plot

Figure 3: Diagnostic plots for the spatial regression data

4. Discussion and Conclusion

The main objective of our research was to develop a method for the joint identification of outliers and high leverage points for spatial regression. In Section 2 we develop a new method to identify both of them and propose a new graphical display called GP-GSR plot to locate both of them in the same graph. In spatial statistics literature observations with neighbourhood instability are diagnosed as outliers. For this reason we employ our method on the first order difference of x and y. A numerical example clearly shows the advantage of using our proposed method. It clearly shows that the proposed method can successfully identify outliers and high leverage points simultaneously while the existing methods fail to do so.

Acknowledgements

The authors express their thanks and gratitude to the anonymous reviewers for giving some useful suggestions that led to considerable improvement in the methodology and presentation of the results. The authors also express their thanks to the Chief Editor for his suggestions on restructuring the contents for better expression.

References

Billor, N., Hadi, A. S., and Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators.*Computational Statistics and Data Analysis*, **34**, 279-298.

Gray, J. B. (1984). A simple graphic for assessing influence in regression. *Journal of Statistical Computation and Simulation*, **24**, 121-134.

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*, **14**, 1-27.

Hadi, A. S. and Imon, A. H. M. R. (2018). Identification of multiple outliers in spatial data. *International Journal of Statistical Science*, **16**, 87-96.

Hadi, A. S., Imon, A. H. M. R. and Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, **1**, 57–70.

Imon, A. H. M. R**.** (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, **22**, 207–218.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*, **32**, 929–946.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Shekhar, S., Lu, C., and Zhang, P. (2002). Detecting graph-based spatial outlier. *Intelligent Data Analysis*, **6**, 451-468.