Identification of Outliers in Spatial Data

Ali S Hadi
Department of Mathematics and Actuarial Science, The American University in Cairo,Egypt
ahadi@aucegypt.edu

A.H.M. RahmatullahImon*
Department of Mathematical Sciences, Ball State University, USA- rimon@bsu.edu

Abstract

In the recent years, anomalous spatial patterns or spatial outliers have received a great deal of attention in environmental statistics. Spatial outliers are those observations whose characteristics are markedly different from their spatial neighbors. Conceptually spatial outliers are very different from classical outliers. 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. A variety of outlier detection methods is available in the literature [see, e.g., Barnett and Lewis (1994), Hadi et al. (2009)], but they cannot be directly applied to spatial data in order to extract abnormal patterns. Traditional outlier detection methods mainly focuses on 'global comparisons' and identifies observations which stand apart from the remainder of the entire data set. In contrast, spatial outlier detection methods concentrate on discovering neighborhood instabilities that break the spatial continuity. Spatial *z* test is a very simple, easily understood and popular technique for the identification of spatial outliers. This test is designed to identify a single spatial outlier and hence may not accurately locate outliers when multiple outliers exist in a cluster and correlate with each other. Even the repeated use of this test may not work and in the end the genuine outliers may be left undetected (a problem known as masking) and/or some of the nonoutlying observations may be incorrectly declared as outliers (a problem known as swamping). All classical statistical methods are generally affected by the masking and swamping problems and the spatial outlier detection methods inherit this problem as well. Furthermore, the existing algorithms tend to abstract spatial objects as isolated points and do not consider their geometrical and topological properties, which may lead to inexact results. In this paper we propose a new type of spatial distances and a corresponding robust spatial z test which should be very effective in the identification of multiple spatial outlier.

Keywords: Spatial outlier; Differencing; Masking; Swamping; Spatial robust *z* test.

1. Introduction

There are numerous definitions of outliers in the statistical literatures. 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. Awareness of outliers in some form or another has existed for at least several hundred years. It is now evident that the presence of outliers can lead to wrong inferences. 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. Automated identification of outliers is tricky due to the well-known masking and swamping effects. Identification of outlying data points is often by itself the primary goal, without any intention of fitting a statistical model. The outliers themselves are points of primary interest, drawing attention to unknown aspects of the data, or especially if unexpected, leading to new discoveries.

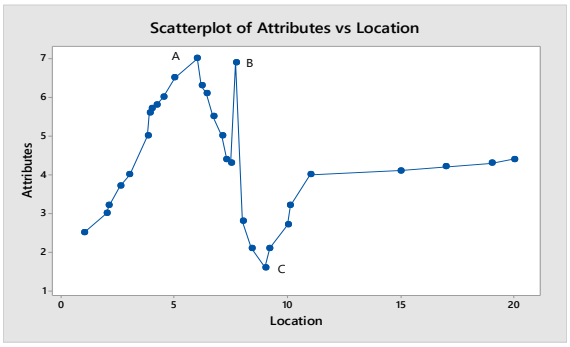
This is the time of big data and big data poses big challenges. The identification of outliers in big data poses a big challenge too. In statistical data the concept of outliers is global. Outliers are the observations

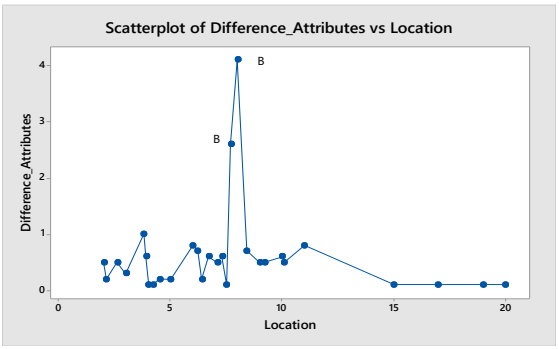
which fail to match with a global pattern (parent distribution). But most of the times big data do not obey any common global pattern and observations may come in clusters where the minority concept of outliers does not work. Spatial data is a classic example of this kind. In the recent years, anomalous spatial patterns or spatial outliers have received a great deal of attention and has become an important branch of data mining. Spatial outliers are those observations whose characteristics are markedly different from their spatial neighbors. Conceptually spatial outliers are very different from classical outliers. 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.

An excellent review of different aspects of spatial outliers is available in Shekhar *et al*. (2003). 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. A good number of spatial outlier detection methods are now available in the literature. These methods can be generally grouped into two categories, namely graphic approaches and quantitative tests. An excellent review of graphic approaches for the identification of spatial outliers is available in Shekar *et al*. (2002). Graphic approaches are based on visualization of spatial data which highlights spatial outliers. These methods include variogram clouds [Haslette *et al*. (1991)], pocket plots [Panatier (1996)], spatial scatter plot [Haining (1993)], and Moran scatter plot [Anselin (1996)]. The commonly used '*k*-means' clustering method may not work here. But some other distance-based methods such as DB(ε,π)-outliers and grid-based outliers, index-based outliers, nested-loop based outliers, [Knorr and Ng (1997, 1998)], *k*–nearest neighborhood approach [Ramaswamy *et al*. (2000)], resolution-based outlier factor (ROF) [Fan *et al*. (2006)] techniques can be applied there. Distance-based outlier detection models have problems with different densities. These methods cannot compare the neighborhood of points from areas of different densities. To overcome this problem some density based outlier detection methods are suggested. Among them local outlier factor (LOF) proposed by Breunig *et al*. (1999) have become very popular. Following this method Huang and Qin (2004) proposed spatial outlier factor (SOF) which is designed to identify multifactor spatial outliers. For detecting outlier with multiple attributes, traditional outlier detection approaches could not be used properly due to the sparsity of the data objects in high dimensional data space. It has been shown [Aggarwal and Yu (2001)] that the distance between any pair of data points in high dimensional space is so similar that either each data point or none data point can be viewed as an outlier if the concepts of proximity is used to define outliers. As a result, using traditional Euclidean distance function cannot effectively get outliers in high dimensional data set due to the averaging behavior of the noisy and irrelevant dimensions. The issue of robustness of spatial outlier methods in the presence of multiple outliers is discussed by Filzmoser *et al*. (2014) and Lu *et al*. (2003). These two approaches are based on Mahalanobis or robust distances computed in each neighborhood using a common estimation of the covariance matrix.

2. Identification of Outliers Using Robust Spatial z Test

In this section we propose a new test for the identification of spatial outliers. At first we present a very simple motivational example. In Figure 2.1(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 C. In contrast to that, the spatial outliers are like the spikes B as shown in Figure 2.1 (b) which are very different than their neighbors. It looks like an outlier because it violates the law of geography that the nearby things should be very similar. It is also interesting to note that the possible global outliers A and C do not look like outliers anymore and another point B emerges as more extreme than A and C. However, graphical methods are very subjective in nature. For this reason, we need to employ a quantitative test to confirm our suspicion regarding spatial outliers.





(a) (b)

Figure 2.1: Scatter plot of attribute values and their differenced values against locations

Shekar*et al*. (2003) proposed a z test on the difference of attributes that we call a spatial z test. For a set of differenced attributes (in absolute values) $D_1, D_2, ..., D_n$, under a normal assumption, a single value may be considered as an outlier if it falls outside a certain range of the standard deviation. A traditional measure of the 'outlyingness' of an observation D_i with respect to a sample is the ratio between its distance to the sample mean and the sample standard deviation

$$z_i = \frac{D_i - \overline{D}}{s_D} \qquad \text{for } i = 1, 2, \ldots, n. \tag{2.1}$$

z-values defined in (2.1) will be called spatial *z* scores since they are calculated on spatial differences. Observations with $|z_i| > 3$ are traditionally deemed as suspicious (the three-sigma rule), based on the fact that they would be very unlikely under normality, since $P(|z| > 3) = 0.003$ for a random variable *z* with a standard normal distribution.

In this paper we propose a new measure of spatial distance defined by

$$d_i = \begin{cases} |d_{i+1} - d_i| & i = 1 \\ |d_i - d_{1-1}| + |d_i - d_{i+1}| & i = 2,3,...,n-1 \\ |d_i - d_{i-1}| & i = n \end{cases} \tag{2.2}$$

The main advantage of the proposed distances is that a distance d_i is computed for all *n* observations, not just for *n* – 1 observations. Another advantage of the use of the distance between one observation and its two spatial neighbors is that, with some human intervention or interpretation, it can be used for the detection of clusters of outliers. A cluster of contiguous outliers can be detected if between two outliers with large distances, the observations in between will have very small distances. This indicates that the two end observations (with large distances) and the ones in between (with very small distances) are all clustered outliers. This can be decided on by the analyst from the index plot of the distances. The z test is designed for the identification of a single outlier and it is now evident that they often fail to identify multiple outliers. Not only that, this test contains components like mean and standard deviation which can be severely affected in the presence of a single outlier and this distortion could be so huge that the z test may fail to identify even a single outlier. Here we replace mean and standard deviation by their robust counterparts. Sample median is a very effective robust measures of location. For the measure of dispersion, we can use the normalized median absolute deviation (NMAD). For the set of differenced attributes, d_i, we compute the *Median Absolute Deviation* (MAD) defined as

$$\text{MAD}(d_i) = \text{Median}\{|d_i - \text{Median}(d_i)|\}. \tag{2.3}$$

To make the MAD comparable to the SD in terms of efficiency, we consider the normalized MAD defined as

NMAD (d_i) = MAD (d_i) / 0.6745. (2.4)

Thus, we can introduce a robust spatial *z*-like statistic defined as

$$z_i' = \frac{d_i - Median(d_i)}{NMAD(d_i)} \quad (2.5)$$

Observations with $|d_i| > 3$ will be identified as spatial outliers.

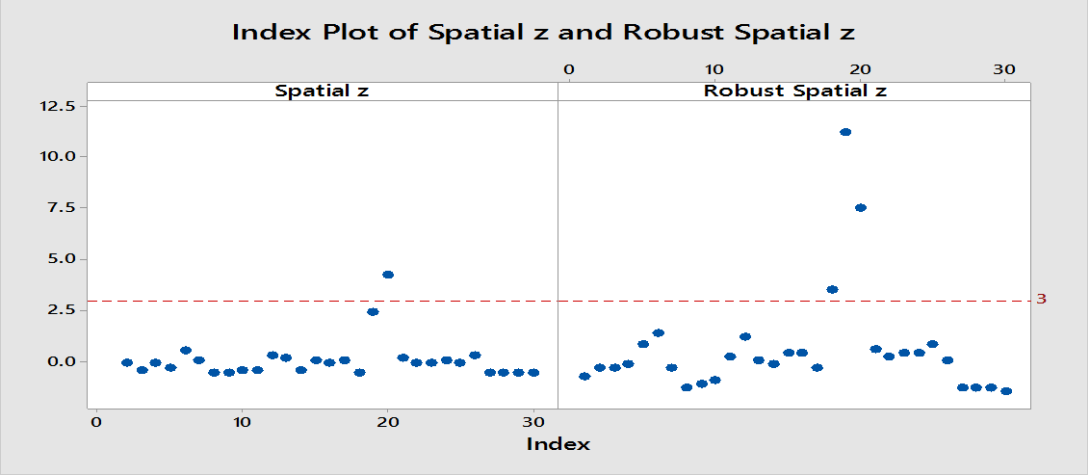
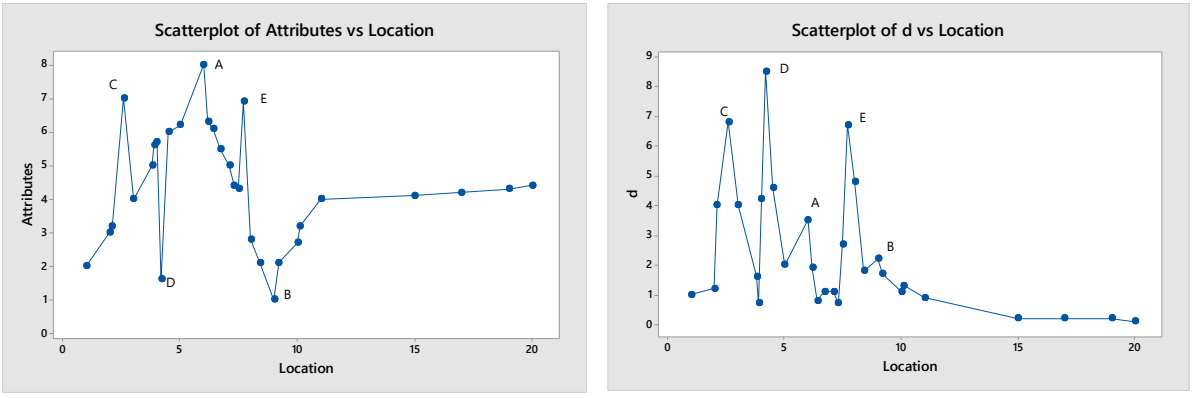


Figure 2.2: Index plot of z scores of differenced attribute values

For this motivational example we calculate the spatial *z* scores for the difference of attributes and those are presented in Figure 2.2. Since z scores are computed on the differences of attribute values, each point appears in pairs and if it is a genuine outlier the pairs should exhibit unusual patterns. Figure 2.2 shows that neither of the suspect global outliers A nor C is identified as a spatial outlier. The spatial z test can identify only one member of the pair corresponding to the genuine outlier B. Since B stands apart from both of its neighbors, we should get three large differences around this point that what we observe from the index plot of the proposed robust z scores.

Table 2.1 presents a multiple spatial outlier data which are generated exactly in a similar way as the above motivational example. In this example, the largest and the smallest data are denoted by A and B, respectively, but neither of them is a spatial outlier. But the points, which are very different from their neighbors, are C, D and E as shown in Figure 2.3(a).



(a) (b)

Figure 2.3: Plot of original and differenced attribute values against locations for multiple outliers

Table 2.1: Multiple Spatial Outliers

Index	Location	Attributes	Traditional D_i	Proposed d_i	Spatial z_i	Robust Spatial z_i
1	1.0	2.0	*	1	*	-0.4615
2	2.0	3.0	1.0	1.2	-0.17078	-0.3195
3	2.1	3.2	0.2	4	-0.75343	1.6685
4	2.6	**7.0 C**	3.8	6.8	1.86851	**3.6565**
5	3.0	4.0	3.0	4	1.28586	1.6685
6	3.8	5.0	1.0	1.6	-0.17078	-0.0355
7	3.9	5.6	0.6	0.7	-0.46211	-0.6745
8	4.0	5.7	0.1	4.2	-0.82627	1.8105
9	4.2	**1.6 D**	4.1	8.5	2.08701	**4.8635**
10	4.5	6.0	4.4	4.6	2.30551	2.0945
11	5.0	6.2	0.2	2	-0.75343	0.2485
12	6.0	8.0 A	1.8	3.5	0.41188	1.3135
13	6.2	6.3	1.7	1.9	0.33905	0.1775
14	6.4	6.1	0.2	0.8	-0.75343	-0.6035
15	6.7	5.5	0.6	1.1	-0.46211	-0.3905
16	7.1	5.0	0.5	1.1	-0.53494	-0.3905
17	7.3	4.4	0.6	0.7	-0.46211	-0.6745
18	7.5	4.3	0.1	2.7	-0.82627	0.7455
19	7.7	**6.9 E**	2.6	6.7	0.99453	**3.5855**
20	8.0	2.8	4.1	4.8	2.08701	2.2365
21	8.4	2.1	0.7	1.8	-0.38927	0.1065
22	9.0	1.0 B	1.1	2.2	-0.09795	0.3905
23	9.2	2.1	1.1	1.7	-0.09795	0.0355
24	10.0	2.7	0.6	1.1	-0.46211	-0.3905
25	10.1	3.2	0.5	1.3	-0.53494	-0.2485
26	11.0	4.0	0.8	0.9	-0.31644	-0.5325
27	15.0	4.1	0.1	0.2	-0.82627	-1.0295
28	17.0	4.2	0.1	0.2	-0.82627	-1.0295
29	19.0	4.3	0.1	0.2	-0.82627	-1.0295
30	20.0	4.4	0.1	0.1	-0.82627	-1.1005

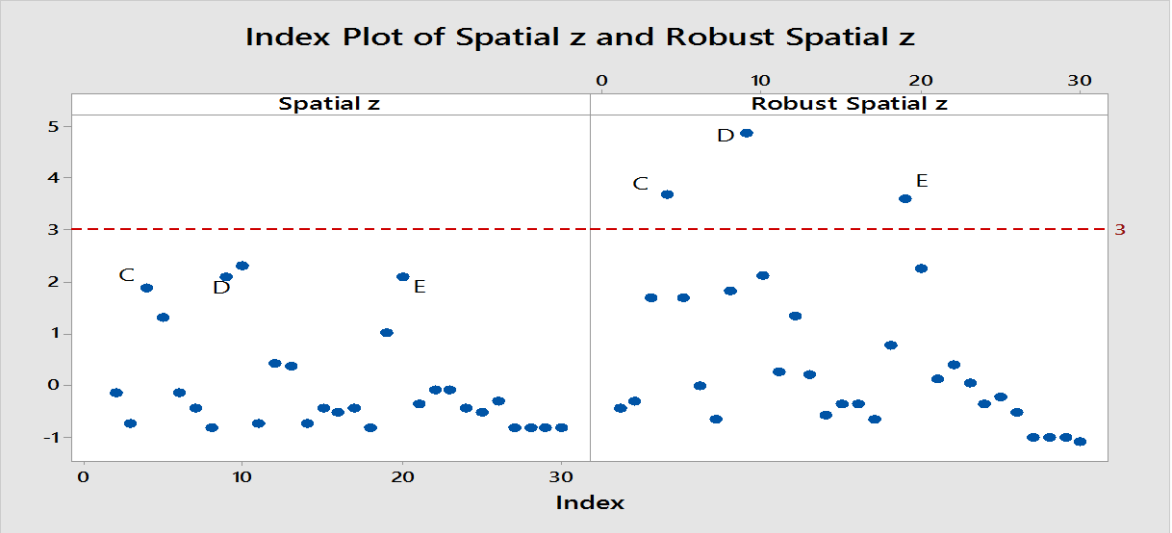


Figure 2.4: Index plot of spatial z and robust spatial z scores values for multiple outliers

The proposed d_i values are presented in column 3 of Table 3.1 and in Figure 3.1 (b). We get a clear impression that the points C, D, and E are surprising but A and B are not. Column 5 of Table 3.1 presents spatial z scores of all d_i's. We observe that all of them are substantially less than 3 in absolute terms and hence fail to identify any of the genuine spatial outliers. Similar remark applies with Figure 3.2 (a) where we display the spatial z scores for this data. We compute robust spatial z scores for this data which are shown in column 6 of Table 3.1 and also in Figure 3.2(b), we observe that points C, D and E are bigger than the cut-off point 3 (in absolute terms) and hence are easily identified as spatial outliers.

3. Conclusions

Spatial z scores have been in use for the identification of spatial outliers. Since these scores contain some components which breakdown in the presence of outliers and consequently they suffer from masking. As a remedy to this problem, we propose robust spatial *z* scores in this paper for the identification of multiple spatial outliers. A numerical example shows that the proposed method can successfully identify the multiple outliers while the existing spatial *z* scores fail to do so.

References

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data, Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 30.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association in M. Fischer, H. Scholten, and D. Unwin, eds., Spatial Analytical Perspectives on GIS, 111–125, London: Taylor and Francis.

Barnett, V., & Lewis, T. B. (1994). Outliers in statistical data, 2nd ed., New York: Wiley.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. R. (1999). OPTICS-OF: Identifying local outliers in Principles of Data Mining and Knowledge Discovery Lecture Notes in Computer Science, 262.

Fan, H., Zaïane, O. R., Foss, A., & Wu, J. (2006). A nonparametric outlier detection for efficiently discovering top-n outliers from engineering data, Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore.

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

Haining, R. (1993). Spatial data analysis in the social and environmental sciences, London: Cambridge University Press.

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

Haslett, J., Brandley, R., Craig, P., Unwin, A., & Wills, G. (1991). Dynamic graphics for exploring spatial data with applications to locating global and local anomalies, The American Statistician, 45, 234-242.

Knorr, E., & Ng, R. (1997). A unified notion of outliers: Properties and computation, Proceedings of the International Conference on Knowledge Discovery and Data Mining, 219–222.

Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets, Proceedings of 24th VLDB Conference.

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

Panatier, Y. (1996). Variowin. Software for Spatial Data Analysis in 2D, New York: Springer-Verlag.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets, Proceedings of the 2000 ACM SIGMOD Int. Conf. on Management of Data, 427–438.

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