Spatial Outlier Detection of Gaussian Shapes

Boleslo E. Romero[1]

[1] Department of Geography, University of California, Santa Barbara, USA 93106-4060
bo_romero@geog.ucsb.edu

Abstract

Spatial outliers exist at locations where attribute characteristics are quite different compared to their local neighborhood. The interplay between spatial dependence and anomalies at various scales has historically led to abundant research. Such anomalies may represent hazardous situations or rare places with special characteristics in need of preservation. Numerous computational methods have been developed for finding spatial outliers and have often been tested with real data. In contrast, this study evaluated three established techniques against a controlled Gaussian distribution to discern breakdown characteristics of the classifiers. It builds a foundation for further analysis involving the detection of such anomalies. Though the established methods generally work well, the results also reveal challenges for detecting spatial outliers accurately.

Keywords: spatial outlier, spatial dependence, scale, raster, simulation.

1. Introduction

Outliers typically have extremely different values compared to an entire data set (Barnett and Lewis, 1994). Spatial outliers, however, are objects in spatially-referenced systems that have significantly different attribute values in a local neighborhood (Shekhar *et al*., 2003). Spatial outliers are often removed to model general trends. The statistical structure of spatial data is often easier to discern without such extreme attribute values, however, accounting for outliers and finding reasons why they occur remains a current research goal (Cressie and Wilke, 2011).

The concept of spatial dependence (Tobler, 1970) presented statistical challenges because the assumption of independence is violated to some degree. This issue led to the development of quantitative measures of spatial association to better understand the general character of spatial phenomena (Moran, 1950; Getis and Ord, 1992). Eventually, theories began to include the idea of local instabilities and analytical methods provided means for finding spatial outliers (Anselin, 1995; Anselin, 1993).

Even though outliers seem to be discontinuities compared to their spatial neighborhood, they may still exhibit smoothness at finer scales. Consider that if a series of position measurements are assumed to be independent and affected only by random error, an estimate of the true location may exhibit qualities of a Gaussian distribution. Likewise, if an attribute is measured as being quite different at a location compared to the surroundings, a discontinuity may occur and errors may again lead to representation with Gaussian qualities. Because of the local smoothness of the Gaussian function, the spatial outlier may or may not be detected, depending upon the spatial resolution of analysis, the height or width of the anomaly, or possibly the assignment of discrete data values as an aggregate of properties across space.

Spatial outliers are important for understanding nature and might represent rare situations of interest. A better understanding of their nature begins with identifying their location and characteristics. Numerous detection methods have been

developed and most are 'field-tested' with complex real data (Aggarwal, 2013; Chandola *et al*., 2009). Alternatively, this is a controlled study of an outlier with Gaussian properties and the performance of three spatial outlier detection methods: the non-iterative 'Median' algorithm of Lu, *et al*. (2003), the 'Z-test' algorithm of Shekhar, *et al*. (2003), and the 'SLOM' algorithm of Chawla and Sun (2006). This is part of continuing research to further evaluate spatial outlier detection.

1.1. Discrete value assignment

The following analyses are constrained to the field representation of spatial processes as discrete-valued grids. It is important to consider how values of the underlying process are assigned to the cells. Three assignment operators, the minimum, maximum, and mean of the cell area were tested since their use could result in considerably different results. It is expected that using the maximum will boost the outlier detection.

1.2. Scale and spatial resolution

If an outlier's spatial structure matches the resolution of the data, the signal can be clearly observed. However, at coarser or finer resolutions, the signal can be masked or swamped. For example, if the resolution is much finer than the scale of the anomaly, then the similarity of neighboring data values may result in outliers failing to be detected. This leads to a major concern about determining what resolution an outlier detection method breaks down.

1.3. Spatial data distribution

The Gaussian model was selected since spatial processes may exhibit such variability either directly or through transformation. Measurement errors and the Central Limit Theorem are also related to the Gaussian form. The two dimensions, varying height and width of the Gaussian form, were tested.

2. Methods

Two open-source programs were used. The statistical software R was used for the simulations and analysis (R Development Core Team, 2013); the ELKI software was used to perform the spatial outlier detection (Achtert *et al.*, 2013).

As shown in Figure 1, the Gaussian parameters were varied independently. To begin, the height and width were both set to one unit. The parameters were incremented up to a maximum factor of nine, first raising the function to a high peak, spreading outward, then reducing the height, and then the spread to the original size. For each of the increments, the assignment operator and resolution were also varied.

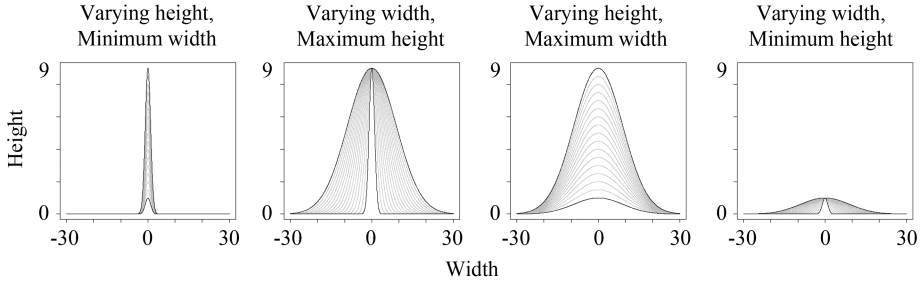

Figure 1. The Gaussian variable space in the experiments. The height and width were varied independently. First the height and width were each increased, then decreased. Spatial resolution and assignment operators were also varied at every increment.

Reference data were created for each parameter set by labeling the Gaussian outlier from its center to three standard deviations out. Also, the raw detection outlier scores were thresholded at the 90th percentile to label outlier results. An accuracy assessment was conducted by calculating for each permutation the 'true positive rate' and 'false positive rate' metrics, which are determined with a classification confusion matrix and Equations (1) and (2).

$$\textit{True positive rate} = \textit{resulting positives} / \textit{reference positives} \quad (1)$$

$$\textit{False positive rate} = \textit{resulting positives} / \textit{reference negatives} \quad (2)$$

For comparison of the effects of each variable between the algorithms, the true positive rate and false positive rate of all permutations were aggregated as both mean and median values and plotted. Generally, it is not desirable to aggregate results for analysis of extreme-valued data, but the changes between mean and median values provide additional information about the distribution of the results at each incremental step of the variable.

3. Results and Discussion

As shown in Figure 2, fine resolution (i.e. many cells across the area of analysis) had a major impact of reducing the true positive rate. Coarser resolution performed best, but with the area subdivided by at least 27 cells, the true positive rate was less than 50 percent. Also at coarse resolution, the bulk of the results distribution was high as evidenced by the median being above than the mean. A multi-scale outlier detection process is suggested to address the issue of a an outlier not being detected due to neighboring cells having similar values with fine spatial sampling. Regarding false positives, although the 'Z-test' did poorly with coarse resolution, all methods generally performed well across each of the variables.

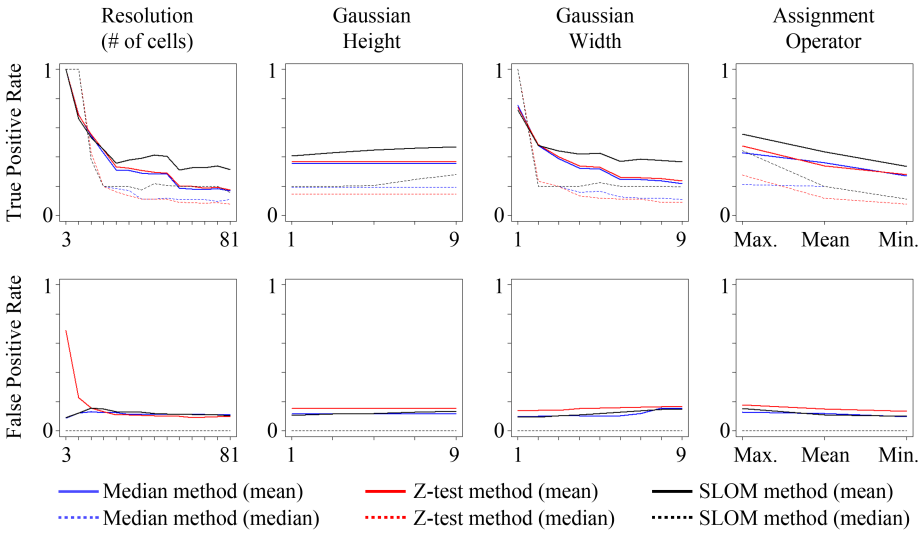

Figure 2. The mean and median true positive rates and the false positive rates of the detection results compared to the reference data, along the top and bottom rows, respectively. The four isolated variables, from left-to-right, are spatial resolution, Gaussian height, Gaussian width, and the assignment operators 'maximum', 'mean', and 'minimum'.

Varying height had very little affect on the results, suggesting the methods detected minor or major outliers equally, however the SLOM methods slightly improved with higher anomalies. On the other hand, varying the Gaussian width

had strong effects for all the methods: smaller widths resulted in much higher true positive rates. Larger widths failed to perform as well, likely again due to similarity of nearby values. Though each of the assignment operators just managed to represent the outlier enough for detection, in this case the outlier was positive-valued and the 'maximum' cell assignment operator improved the outlier representation and detection.

An interesting outcome was the variety of patterns that emerged from the digital outlier simulation and detection methods, as illustrated in Figure 3. The first three patterns along the top, from left to right, are examples of an 'original' Gaussian, the reference data, and raw outlier scores. The other two patterns along the top are examples of thresholded results from the 'Median' method and the second and third rows show 'Z-test' and 'SLOM' results. The diversity of patterns show how a simple, smooth Gaussian function can result in complex outlier patterns by means of discrete-valued data and analysis. If one were to 'detect' such a pattern in their data set, it would be quite a reach to intuitively or computationally derive the original distribution. It is also noteworthy that the highest attribute value at the center rarely resulted in a detected outlier, but the transitional, sloping sides did.

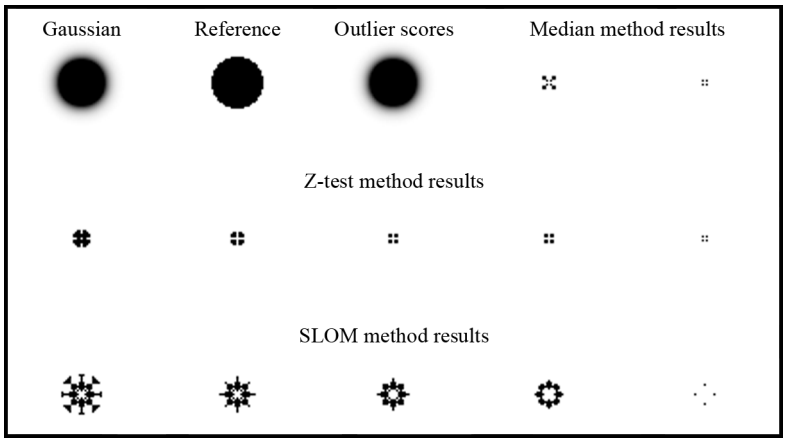

Figure 3: Simulations and detected results of the same Gaussian distribution.

References

Achtert, E., Kriegel, H.-P., Schubert, E., Zimek A. (2013), "Interactive Data Mining with 3D-Parallel-Coordinate-Trees." *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, New York City, NY, 1009-1012 p.

Aggarwal, C. (2013), *"Outlier Analysis"*, Springer Science + Business Media, New York, NY, 446 p.

Anselin, L. (1993), "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association". *GISDATA Specialist Meeting on GIS and Spatial Analysis*, The Netherlands, December 1-5, 1993.

Anselin, L. (1995), Local Indicators of Spatial Association – LISA. *Geographical Analysis*, Vol. 27(2): 93-115.

Barnett, V., Lewis, T. (1994), *"Outliers in Statistical Data"*, John Wiley and Sons Ltd., West Sussex, England, 575 p.

Chandola, V., Banerjee, A., Kumar, V. (2009), "Anomaly Detection: A Survey", A*CM Computing Surveys*, Vol. 41(3), Article 15: 1-58.

Chawla, S., Sun, P. (2006), "SLOM: a new measure for local spatial outliers". *Knowledge Information Systems*, Vol. 9(4): 412-429.

Cressie, N., Wilke, C. (2011), "*Statistics for Spatio-temporal Data"*, John Wiley & Sons, Inc., Hoboken, NJ, 588 p.

Getis, A., Ord, J.K. (1992), "The Analysis of Spatial Association by Use of Distance Statistics", *Geographical Analysis*, Vol. 24(3): 189-206.

Lu, C.-T., Chen, D., Kou, Y., (2003), "Algorithms for Spatial Outlier Detection", *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, Florida, USA, 597-600 p.

Moran, P.A.P. (1950), "Notes on Continuous Stochastic Phenomena". *Biometrika,* Vol. 37(1): 17–23.

R Development Core Team, (2013), R: A language and environment for statistical computing. Vienna, Austria: *R Foundation for Statistical Computing.*

Shekhar, S., Lu, C.-T., Zhang, P. (2003), "A Unified Approach to Detecting Spatial Outliers" *GeoInformatica*, Vol. 7(2): 139-166.

Tobler, W. (1970), "A Computer Movie Simulating Urban Growth in the Detroit Region". *Economic Geography*, Vol. 46(2): 234-240.