Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/050307

Multivariate Analysis of Crime Data using Spatial Outlier Detection Algorithm

Alok Kumar Singh[*]

Department of Statistics, University of Allahabad, Allahabad, India-211 002

Received: 31 May 2016, Revised: 24 Jul. 2016, Accepted: 28 Jul. 2016
Published online: 1 Nov. 2016

Abstract: A spatial outlier is a spatially referenced object whose non spatial attribute value is significantly different from the corresponding values in its spatial neighbourhood. In other words, a spatial outlier is a local instability, or an extreme observation which deviates significantly in its spatial neighbourhood, but may not be in the entire data set. In this paper, we have applied the well-known mean algorithm for detecting spatial outliers in the multiple attributes state wise crime data and predicted which states need more attention from the government so as to reduce crimes there. We have also done regression analysis between the populations of age group 15-19 years and separately for population of age 15 years with the crimes data of all states.

Keywords: Outlier,Spatial Outlier, regression analysis, Mahalanobis distance, Crime

1 Introduction

Hawkins [2] had defined an outlier as an observation that is significantly different from the remaining observation in a dataset so as to arouse suspicion that it was generated by different mechanism. Barnett and Lewis [13] had defined an outlier as an observation that deviates significantly in the sample in which they occur.

There are generally three types of outliers: points, contextual (or conditional) and collective [14]. Point outliers are the data points that are significantly different with rest of the data points. Contextual (or conditional) outliers are the data points which are different from the remaining data points in a specific context but not otherwise [15]. These types of outlier are defined with the help of two attributes: contextual and behavioural. The former one is used to define the context in which the outliers are assessed and later is the attribute used to decide outlierness of a data points. A cluster outlier is a group of observation that are clustered together having low variance among them but it is significantly different from the remaining dataset.

Spatial datasets are spatially referenced objects with some non-spatial attributes. In other words spatial datasets consist of two attributes: spatial and non-spatial. Spatial attributes include location, shape and other geometric or topological properties and non-spatial attributes include length, height, ownership, building age, name, etc. Spatial neighbourhood of a spatially referenced point is the subset of spatial datasets based on the spatial dimension using spatial relationship e.g., distance or adjacency. Spatial outlier is a spatially referenced object whose non spatial attribute is significantly different from the corresponding values in its spatial neighbourhood. In other words, we can say that a spatial outlier is a local instability or an extreme observation which is different in its neighbourhood but may not be in the entire dataset.

Outlier detection is very important in getting some interesting and useful patterns for further analysis. In literature, various tests for outlier detection, also known as discordancy tests, have been proposed by several authors [13,14]. The main idea behind these types of tests is to fit the datasets to a known distribution and to develop a test based on the distribution properties. These types of tests are based on the assumption that the data distribution is known, the observations are independently and identically distributed, distribution parameters are known and the number of

[*] Corresponding author e-mail: alok.rjnis@gmail.com

expected outlier must be known in advance. These tests are unsuitable when the above mentioned assumptions are not met. This condition usually arises in the case when the observations are from unknown distribution and also of higher dimension. Several non-parametric techniques have been proposed, including distribution based, density based, cluster based and depth based [6,5,8,4].

Detection of spatial outlier is useful in many applications like geographic information systems (GIS), spatial databases, transportation, ecology, public safety, public health climatology etc. but the above mentioned techniques are not suitable for detecting spatial outlier as they would generally identify global extreme observations as spatial outliers. In 2003, Shekhar et al. [11,12] introduced a technique for detecting spatial outliers in graph datasets. Their method was based on the distribution property of the difference between non-spatial attribute value at a point and average of non-spatial attribute values in its spatial neighbourhood. Several spatial outlier detection techniques are available in literature [10,1,7]. They can be classified in two categories: graphical methods and quantitative methods. Graphical methods, for example, variogram cloud and pocket plots, is based on the visualization of spatial dataset which highlights spatial outliers. Quantitative methods, for example scatter plots and Moran scatter plots, provide a test which distinguishes spatial outlier from the remaining data points.

The above mentioned approaches are only suitable in case of a single attribute only. They could not be successfully applied for the higher dimensional data due to sparsity of data in higher dimension. In literature, the two well-known approaches for dealing with multiple attributes, namely mean and median approach [3] used Mahalanobis distance which is more suitable for dealing with higher dimensional data as compared to Euclidean distance.

In view of the above, the present study deal with the mean approach for detecting spatial outlier in a multiple attribute data and applied regression analysis technique between the population of age group 15-19 years and separately for population of age 15 years in a given state with the crimes in that state.

2 The Methodology

Let X=$\{x_1,x_2,\ldots,x_n\}$ be a set of spatially referenced objects and for each x_i,there are multiple attributes $y_{i1},y_{i2},\ldots,y_{iq}$. So we have a function $f : X \to R^q$, where R^q denotes the q dimensional Euclidean space such that each x_i in X is associated with an element $f(x_i)$=$(f_1(x_i), f_2(x_i),\ldots, f_q(x_i))$=$(y_{i1},y_{i2},\ldots,y_{iq})$ of R^q. This function is called an attribute function associated with X. Denote the set $\{y_1,y_2,\ldots,y_n\}$ by Y. Next, for each x_i , there is a set of k nearest neighbours in the spatial neighbourhood of x_i, denoted by $NN_k(x_i)$.If k is a variable for each x_i, then it is usually denoted by k_i,for $i = 1,2,\ldots,n$ and in this case the spatial neighbourhood of x_i is denoted by $NN_{k_i}(x_i)$. g is a neighbourhood function from X to R^m , whose i^{th} component gives the summary statistics of attribute value y_i in the spatial neighbourhood of x_i i.e. $NN_k(x_i)$(or $NN_{k_i}(x_i)$). A comparison function h is the function of f and g, whose domain is X and range is R^s, where $s \le m$. If we take $h = f - g$, then $s = m$ and if $h = f_1/g_1$, then $s = 1$. Clearly , dimension of the range of h depends on the choice of h. In this paper, we have taken $h = f - g$, so $s = m$. In the mean algorithm, $h(x)$ is assumed to be normally distributed as $N_q(\mu,\sum)$ i.e., m dimensional vector $h(x)$ follows the multivariate normal distribution with mean vector μ and variance-covariance matrix $\sum$, where μ and $\sum$ are given by: $\mu = \frac{1}{n}\sum\limits_{i=1}^{n} h(x_i)$ and $\sum = \frac{1}{n}\sum\limits_{i=1}^{n}[h(x_i)-\mu][h(x_i)-\mu]^T$.

Then $d^2(x) = [h(x_i)-\mu]^T\sum^{-1}[h(x_i)-\mu]$ follows χ^2_q, where χ^2_q is chi-square distribution with q degrees of freedom.

3 Mean Algorithm

1.Given a set X=$\{x_1,x_2,\ldots,x_n\}$ of spatially referenced objects and a predefined threshold value θ, depending on the confidence interval.

2.Standardize the attribute function f_j by replacing $f_j(x_i)$ by $\frac{f_j(x_i)-\mu_{(f_j)}}{\sigma_{(f_j)}}$,for each $j = 1,2,\ldots,m$.

3.For each x_i, compute the neighbourhood function $g = (g_1,g_2,\ldots,g_m)$, where $g_j(x_i)$ is obtained by taking average value of $f_j(x_i)$ in the neighbourhood of x_i i.e. $g_j(x_i) = \frac{1}{k_i}\sum_{x_i\in NN_{k_i}(x_i)} f_j(x_i)$.

4.Calculate the comparison function h by $h(x_i) = f(x_i) - g(x_i)$,for $i = 1,2,\ldots,n$.

5.Compute $d^2(x_i) = [h(x_i)-\mu]^T\sum^{-1}[h(x_i)-\mu]$ for $i = 1,2,\ldots,n$.
If $d^2(x_i) > \theta$, then x_i is a spatial outlier with respect to Y.

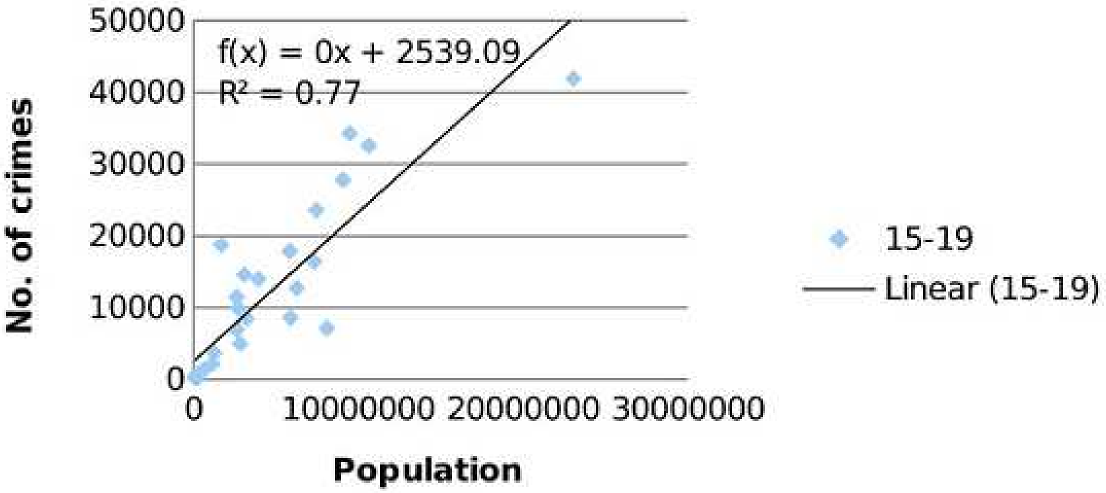

Fig. 1: Graph showing number of crimes as a function of population of age group 15-19 years

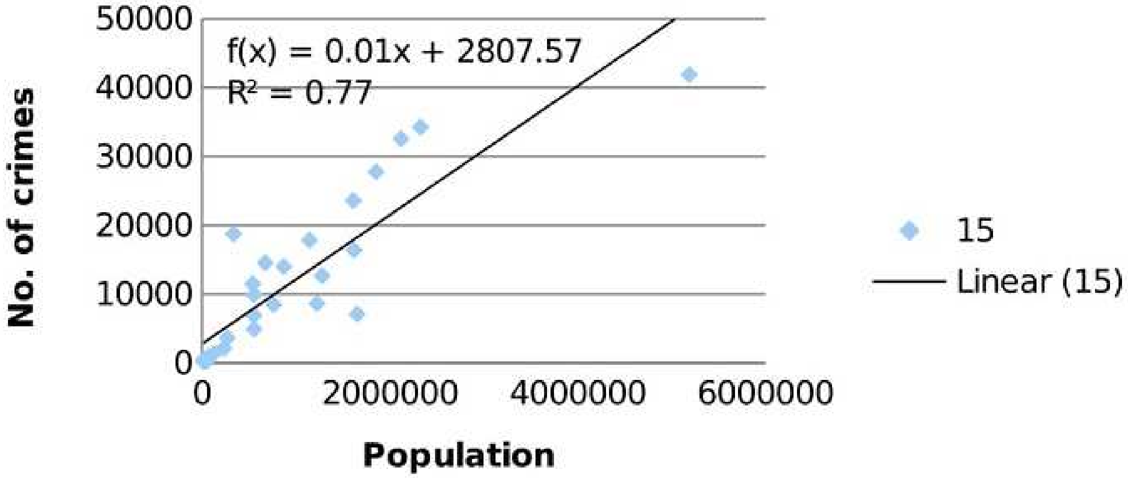

Fig. 2: Graph showing number of crimes as a function of population of age 15 years

4 Experiment and Discussion

We have applied the well-known mean algorithm to the crime data of India, 2014 which has been compiled by Ministry of Home Affairs [9]. This data is a type of spatial data where spatially referenced points were given to be set of states and union territories. We have considered figures of 29 states and one union territory, Delhi. So in our case we have 30 spatial points, whose spatial attributes are chosen to be set of latitudes and longitudes passing through that location and non-spatial attributes are given to be figures of different types of crimes, which are already given state wise. In our experiment, we are considering 13 non spatial attributes (means 13 types of crimes). For each spatial referenced point x_i, the spatial neighbourhood is the set consisting of all spatial points which share common boundaries with x_i. So for each x_i, number of members in the spatial neighbourhood were taken to be variable, which we denote here by k_i, $i = 1, 2, \ldots, n$. The dataset has been standardized and we have applied the mean algorithm and obtained the results.

As can be seen from Table 1 that Delhi is detected as the top most spatial outlier because it has the largest mahalanobis diatsance 27.02. This largest mahalanobis distance mainly comes from the contribution of the corresponding attribute crime (standardised value of rate of violent crime 4.44) as compared to its neighbouring states, Haryana(0.57), Uttar Pradesh(-0.63) and Rajasthan(-0.39). The second detected outlier, Bihar has also high rate of violent crimes (standardised

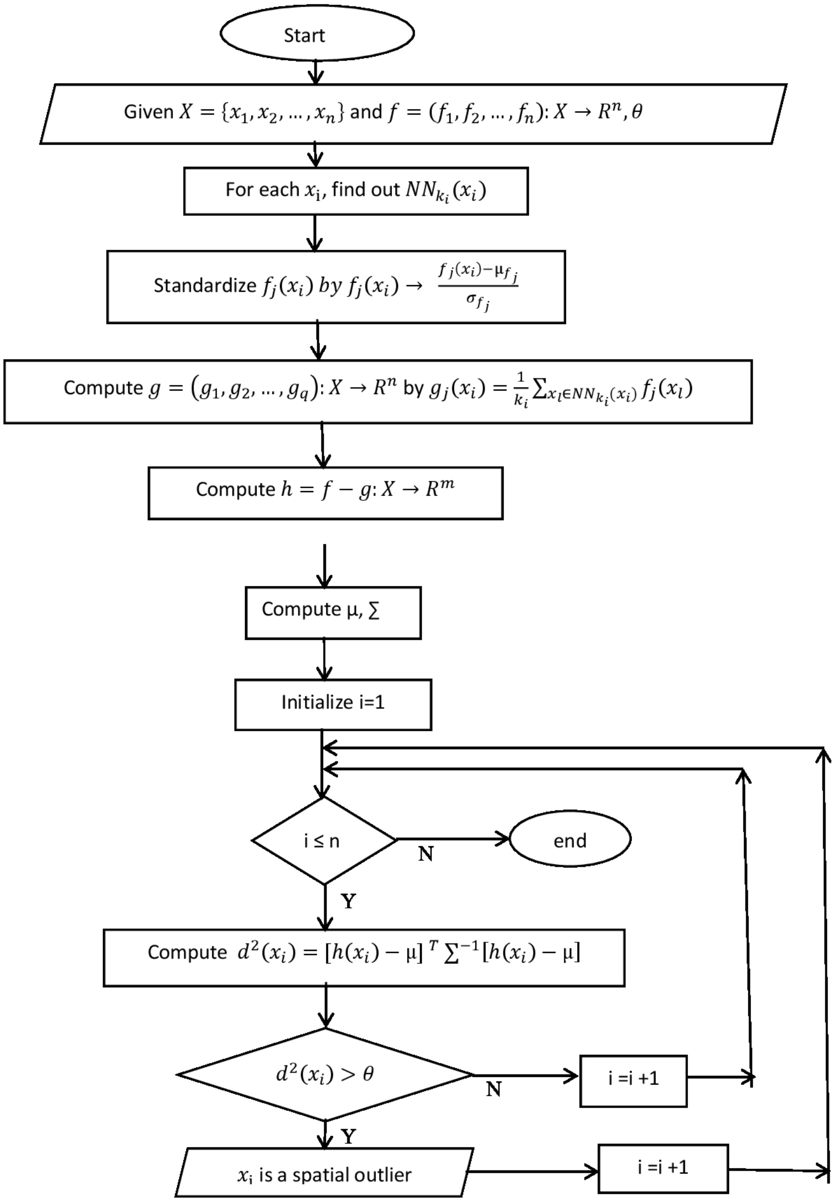
Start
Given $X = \{x_1, x_2, \dots, x_n\}$ and $f = (f_1, f_2, \dots, f_n): X \to R^n, \theta$
For each x_i, find out $NN_{k_i}(x_i)$
Standardize $f_j(x_i)$ by $f_j(x_i) \to \frac{f_j(x_i) - \mu_{f_j}}{\sigma_{f_j}}$
Compute $g = (g_1, g_2, \dots, g_q): X \to R^n$ by $g_j(x_i) = \frac{1}{k_i} \sum_{x_l \in NN_{k_i}(x_i)} f_j(x_l)$
Compute $h = f - g: X \to R^m$
Compute μ, ∑
Initialize i=1
i ≤ n
N
end
Y
Compute $d^2(x_i) = [h(x_i) - \mu]^T \sum^{-1} [h(x_i) - \mu]$
$d^2(x_i) > \theta$
N
i =i +1
Y
x_i is a spatial outlier
i =i +1

Fig. 3: Flow chart of Mean Algorithm

Table 1: Table showing top 5 outlier

Rank	States	Mahalanobis Distance
1	Delhi UT	27.0249
2	Bihar	25.8729
3	Kerala	25.3219
4	Odisha	23.1297
5	Uttar Pradesh	22.8712

value 0.34) as compared to its spatial neighbourhood which consists of Jharkhand (-0.20), Uttar Pradesh (-0.63) and West Bengal(0.10). This same argument is true for other detected outliers also. Next, we have done the regression analysis between population of age group 15-19 years and separately for population of age 15 years with the crime values in all states. We have obtained the linear relationships (please see Figure 1 and Figure 2) in both the cases with the same regression coefficient R=0.77.

5 Conclusion

In this paper we have used the well-known mean algorithm for multiple attributes for detection of spatial outliers in the state wise crime data of India, 2014 with multiple attributes. We have detected top 5 outliers by this algorithm. Further, we have done regression analysis for the population of age group 15-19 years and separately for the population of age 15 years with the crimes data of all states. We have obtained linear relationships in both cases with significant values regression coefficient R. The detected spatial outliers give us the states and union territories which need more attention from the government so as to reduce crime there. Next, the regression analysis results show that a significant proportion of teenagers are involved in crimes. However, these cases are the result of various socio-economic conditions prevailing in our society. These are also accompanied by several factors that become obstacles in their development process. According to researchers, almost all juvenile criminals are illiterate or limited to primary education and rapid growth of information technology has led them to cyber-crimes. Also, opportunities are limited, so this leads to their involvement in criminal activities. Special measures should be taken by the government to create, implement and monitor a national action plan, to promote primary prevention response, integrate crime prevention into social and educational policies etc. Also, they should think to give proper guidance, corrective treatments, education, healthcare etc. for children so as to increase possibility for their better life. Also, the government could think of reducing juvenile age further, whenever necessary and also to pay more attention to enforce law and maintain criminal justice etc.

Acknowledgement

The author wish to thank the editor, and the anonymous reviewers for their constructive comments that have helped to improve the earlier version of the manuscript.

References

[1] A. Luc, Exploratory Spatial data analysis and geographic information system, (1994). In M.Painto, editor, *New tools for spatial analysis, 45-54*.

[2] D. Hawkins (1980), *Identification of outliers*, Chapman and Hall.

[3] D.Chen, Y. Kou, Detecting spatial outlier with multiple attributes, (2003). *Proceeding of 15th IEEE international (ICTAI?03)*.

[4] E. Acuna and C.A. Roduguez(2004), Meta analyses of outlier detection methods in classification , *In proceeding IPSI 2004, Venice.*

[5] E. Knorr and R. Ng (1997), A unified notion of outliers: properties and computation, *In proceeding of the international conference of knowledge discovery and data mining, 219-222*.

[6] F. Preparata and M. Shamos (1988), *Computational Geometry: An introduction*, Springer-Verlag.

[7] J. Haslett , R. Brandley , P. Craig ,A. Unuin and G. Wills,(1991). Dynamic Graphics for exploring spatial data with applications to locating global and local abnormities, *The American Statistician, 45, 234-242*.

[8] M. M. Breuning, H.P. Kriegel, R. T. Ng and J. Sandon (2000) , Identifying local outlier, *In proceeding of PKDD?99 Prague, Czech Republic, Lecture Notes in computer Science, pp. 262-270, springer-verlag.*

[9] Ministry of Home affairs,(2014) http://mha.nic.in.

[10] R. Haining, (1993) *Spatial data analysis in the social and Environmental Sciences*, Cambridge University Press.

[11] S. Shekhar, C. Lu and P. Zhang, (Aug 2001). Detecting Graph based spatial outliers: Algorithm and Applications (A summary of Results), *In Proceeding of the Seventh ACM-SIGKDD international conference on knowledge Discovery and Data mining*.

[12] S. Shekhar, C. Lu and P. Zhang, (2002). Detecting Graph based spatial outliers, *intelligent data analysis, An international Journal, 6(5), 451-468.*

[13] V. Barnett and T. Lewis (1994),*Outliers in statistical data*, John Wiley.

[14] V. Chandola, A. Banerjee and V. Kumar (2009), Anomaly detection: a survey, *ACM Computing Survey, 41(3)*.

[15] X. Song, J. Wang, W. Huang, L. Liu, G.Yan, R.Pu (2009), The delineation of agricultural management zones with high resolution remotely sensed data, *Precision agriculture, 10, 471-487.*

Alok Kumar Singh pursuing PhD in Statistics at Univeristy of Allahabad, India. His research interests are in the areas of outlier, spatial outlier, classical and bayesain inference, applied demography, applied mathematical modeling etc