Spatial Analysis of Water Quality Data Using Multivariate Spatial Outlier Detection Algorithms

Sweta Shukla [1] and S. Lalitha [2]

[1] *Department of Statistics*
University of Allahabad, Prayagraj-211002
Uttar Pradesh, India.
sswetashukla.1992@gmail.com

[2] *Department of Statistics*
University of Allahabad, Prayagraj-211002
Uttar Pradesh, India.
lalithadrs@gmail.com

Abstract

Spatial data consists of spatial and non-spatial attributes carrying information on spatial aspects of the object and about the behavioral aspects of the object respectively. An object is termed as a spatial outlier if its non-spatial attributes are different from those in its spatial neighborhood. In this paper, two different spatial outlier detection algorithms based on a geographically weighted approach is applied to water quality data (containing information about the river networks and multiple attributes on different pollution levels) and detected those water monitoring stations which require attention for the betterment of water quality in that respective area.

Subject Classification:[2020] 62-08, 62P12, 62M30, 62R07
Keywords: Geographically weighted methods, Spatial outlier, Mahalanobis distance, Water quality.

1 Introduction

Outlier detection is an important task in doing any kind of statistical analysis and also to understand the behavioral perplexity of the data. According to [6], "an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Study of outliers and their detection procedures find applications in varied areas like health care systems, credit card fraud detection, cybersecurity, earth and meteorological sciences and many more. The area of outlier detection has drawn a lot of attention resulting in different outlier detection methods available in the literature which can be found in [17] and [21]. Spatial data consists of two parts: (i) the spatial part, more formally called the spatial attributes carrying information about spatial nature of data, like location, shape, or size of the object under consideration, and (ii) the non-spatial part or the non-spatial attributes holding information about the behavioral nature of the object, like rates of employment, literacy, etc. for the states, counties or for the location of the object under consideration. With the advancement in GIS technologies and the database management systems there has been a substantial increase in the amount of Spatial data collected and for analyzing such data, detection of hidden anomalies becomes vital but unlike any traditional data outlier detection procedures available in literature, Spatial data requires special treatment as the traditional methods ignore the spatial nature of the data and thus the conventional methods available for outlier detection fails. According to [19], "A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood". Clearly, the concept of spatial neighborhood is highly critical to outlier detection procedures and can be found in [13]. The spatial outlier detection methods which use spatial autocorrelation as a base to detect spatial outliers include the local Moran's I index by [10], which is a local spatial autocorrelation statistic. There are some graphic based methods for identifying spatial outliers through visualization like Scatterplots by [9], Moran scatterplot by [11] which is a tool to visualize and identify the degree of spatial instability for a single attribute data. Detection of graph-based spatial outliers is given by [18]. The other challenging aspect of this data type is the number of attributes *i.e.* single or multiple, though many methods are available in the literature (discussed above) to deal with the single attributes case either based on graphical or statistical tests but complexity increases for the latter.

The research work combining both the spatial and multivariate aspects of the data are rare and the most recent works in this field include algorithms proposed by [4] for spatial outlier detection in multivariate data and [5] which provides robust methods both for single attribute and multi-attribute case. [16] provides a method for detection with the help of geographically weighted methods. In this paper, the method provided by [16] is applied on a multivariate spatial data.

2 Methodology

2.1 Geographically weighted methods (GW methods)

This approach was initially introduced by [1] and[2] and is quite popular in spatial statistics. It is quite similar but an advanced method than the locally weighted methods available in the literature. The GW methods are based on the moving window technique and have parallels with kernel-based methods.

In this approach, calibration of a model is done locally around a target location say, *(a, b)* which corresponds to the geographical co-ordinates (longitude, latitude) of an object o. Using this target location the model is localised by attaching weights to the objects nearby object o by some kernel function which decays according to the geographic distance between the object o and its neighborhood points and thus giving more weight to the objects which are close to the ones which are distant from the object o. The GW based methods have been extended in many forms like, GW summary statistics by [3], GW distribution analysis by [8], Fuzzy Geographically Weighted Clustering by [7], Geographically weighted Principal Component Analysis by [15], Geographically Weighted generalised linear models by [20], Geographically Weighted discriminant analysis by [14], Geographically Weighted Machine Learning by [12] and many more.

For any GW method, the geographical weight matrix *i.e.*

$$W = ((w_{ij})); i, j = 1, \ldots, n. \tag{2.1}$$

plays a very important role and thus, its construction becomes important too. Here, bi-square kernel is used to generate weights defined below.

$$w_{ij} = \begin{cases} (1-(\frac{dist_{ij}}{bw})^2)^2, & \text{if } dist_{ij} \le bw \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

where, $dist_{ij}$ is the distance between two spatial locations i and j, which could be any distance like Manhattan distance or great circle distance but here Euclidean distance is taken and ***bw*** is the bandwidth of the kernel function, which may be fixed bandwidth where the distance is fixed and the number of observations keeps varying or an adaptive bandwidth where the number of observations is kept fixed but the distance varies.

Here, adaptive bandwidth is used for the kernel functions in which the fixed observations are reported in the form of a percentage of the dataset. In a multivariate structure, the most commonly used distance measure for outlier detection purpose is the well known Mahalanobis distance (MD). The MD between an observation ***X*** to the centre ***C*** and shape ***V*** of data is given by:

$$MD(X,C) = (X-C)^T V^{-1}(X-C) \tag{2.3}$$

where ^{T}denotes the transpose of the matrix and $^{-1}$ denotes the inverse of the matrix. Conventionally, ***C*** and ***V*** are replaced by the mean vector and the variance-covariance matrix, but for the detection of outliers in a multivariate data robust estimation is needed and thus instead of conventional estimates Minimum Covariance Determinant (MCD) estimates are used.

According to [16], the calculation of Geographically Weighted Mahalanobis Distance (GWMD) and Geographically Weighted Principal Component Analysis (GWPCA) for an observation x_i having dependence on its spatial location ***(a, b)*** involves calculation local mean vector ***C(a, b)*** and local variance-covariance matrix ***V(a, b)*** defined below.

$$V(a,b) = X^T W(a,b) X \tag{2.4}$$

where, ***W(a,b)*** are the geographical weights as calculated in equation 2.2. GWPCA follows a similar procedure as the original PCA where the variance-covariance matrix is decomposed into the matrix of eigen values and eigen vectors. Similarly, for calculating the local principal components for GWPCA, local eigen values, and local eigen vectors are calculated for each location ***i = 1, . . . , n,*** as:

$$V(a_i,b_i) = E(a_i,b_i)\Gamma(a_i,b_i)E(a_i,b_i)^T \tag{2.5}$$

where, $V(a_i, b_i)$is the local variance-covariance matrix, $E(a_i, b_i)$ is the matrix of local eigen vectors and $\Gamma(a_i, b_i)$ is the diagonal matrix of local eigen vectors.

Now, for each spatial location $i = 1, \ldots, n$ local MDs *i.e.* $MD_k(a_i, b_i)$ with elements $k = 1, .., M$ and local principal components are found for an adaptive bandwidth of size M. For GWPCA, component scores (CS) method *i.e.* GWPCA (CS) is used where first few and last few components are retained. With the assumption that the $n \times p$ data X (n is the number of spatial locations and p is the number of non-spatial attributes) follows multivariate normal distribution,

$$MD(X,C) = (X-C)^T V^{-1}(X-C) \sim \chi^2_p(\alpha) \tag{2.6}$$

where, χ^2_p is the chi-square distribution with p degrees of freedom at α level of significance. Thus, the cutoff value (cv) is taken as $\chi^2_{p,0.975}$ which is the 97.5% quantile of χ^2 distribution with p degrees of freedom. For GWPCA (CS), the cv is taken to be ±2.5 for the large and small CS values respectively. Any spatial unit $\boldsymbol{i}$ is flagged as a spatial outlier if the local MD and local CS values are greater than $\boldsymbol{cv}$, when $\boldsymbol{j{=}i}$ in each local calibration.

3 Results and Discussion

The algorithms GWCOM and GWPCA(CS) are applied to a ***Water Quality Data 2016*** which is sourced from *ENVIS Centre on Control of Pollution Water, Air and Noise* hosted by *Central Pollution Control Board* and sponsored by Ministry of Environment, Forest and Climate Change, Government of India. Here, only a subset of the original data is taken which spans over 101 water quality monitoring stations in nine states namely, Himachal Pradesh, Punjab, Uttarakhand, Bihar, Jharkhand, West Bengal, Haryana, Delhi, and Uttar Pradesh. The dataset contains information about the water quality of 4 rivers: Ganga, Beas, Satluj, and Yamuna on 16 different variables : Temperature (min, max) in °C, Dissolved Oxygen (min, max) in mg/l, pH (min, max), Conductivity (min, max) in $\mu mhos/cm$, B.O.D. (min, max) in mg/l, Nitrate-N + Nitrite-N (min, max) in mg/l, Faecal Coliform (min, max) in $MPN/100ml$ and Total Coliform (min, max) in $MPN/100ml$.

Thus, the dataset comprises of $\boldsymbol{n{=}101}$ spatial locations (spatial units) and $\boldsymbol{p{=}16}$ variables (non-spatial attributes). The algorithms are applied to this data considering a varied range of bandwidths i.e. 55%, 65%, 75%, 85% and 95%. Bandwidths lower than 55% is not considered because singularity problems for MCD occurred. The results obtained are summarized below.

Tab. 1: Spatial outlier detection by GWMD method for different bandwidths

Bandwidth (bw)	**Total spatial outliers detected**
55%	17
65%	20
75%	21
85%	41
95%	56

Tab. 2: Top ten Spatial outliers detected by GWMD method for different bandwidths

S.no.	**bw=55%**	**bw=65%**	**bw=75%**	**bw=85%**	**bw=95%**
1.	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI(Himachal Pradesh)
2.	BEAS D/S KULLU (Himachal Pradesh)	BEAS D/S KULLU (Himachal Pradesh)	GANGA AT BAHARAMPORE (West Bengal)	BEAS D/S KULLU (Himachal Pradesh)	BEAS D/S KULLU (Himachal Pradesh)
3.	EAST KALI BEIN FALLING INTO RIVER BEAS (Punjab)	BEAS D/S AUT (Himachal Pradesh)	BEAS D/S AUT (Himachal Pradesh)	BEAS D/S AUT (Himachal Pradesh)	BEAS D/S AUT (Himachal Pradesh)
4.	SATLUJ D/S BHAKHRA (Himachal Pradesh)	GANGA AT BUXAR (Bihar)	BEAS AT D/S PANDOH DAM (Himachal Pradesh)	BEAS D/S MANDI (Himachal Pradesh)	BEAS EXIT OF TUNNEL DEHAL POWER HOUSE (Himachal Pradesh)
5.	SATLUJ AT BOAT BDG. DHARMKOT-NAKODAR ROAD, JALANDHAR (Punjab)	SATLUJ AT BOAT BDG. DHARMKOT-NAKODAR ROAD, JALANDHAR (Punjab)	GANGA AT PUNPUN, PATNA (Bihar)	BEAS AT D/S PANDOH DAM (Himachal Pradesh)	BEAS AT D/S PANDOH DAM (Himachal Pradesh)
6.	SATLUJ D/S EAST BEIN BASIN (Punjab)	SATLUJ D/S EAST BEIN BASIN (Punjab)	SATLUJ AT BOAT BDG. DHARMKOT-NAKODAR ROAD, JALANDHAR (Punjab)	EAST KALI BEIN FALLING INTO RIVER BEAS (Punjab)	EAST KALI BEIN FALLING INTO RIVER BEAS (Punjab)
7.	GANGA D/S HARIDWAR (Uttarakhand)	GANGA D/S HARIDWAR (Uttarakhand)	SATLUJ D/S EAST BEIN BASIN (Punjab)	SATLUJ D/S BHAKHRA (Himachal Pradesh)	SATLUJ D/S BHAKHRA (Himachal Pradesh)
8.	GANGA AT GARHMUKTESHWAR (Uttar Pradesh)	GANGA AT NARORA (BULANDSAHAR) (Uttar Pradesh)	GANGA D/S HARIDWAR (Uttarakhand)	SATLUJ AT NEPTHA ZAKHAI (Himachal Pradesh)	SATLUJ AT NEPTHA ZAKHAI (Himachal Pradesh)
9.	GANGA AT NARORA (BULANDSAHAR) (Uttar Pradesh)	GANGA AT BITHOOR (KANPUR) (Uttar Pradesh)	GANGA AT NARORA (BULANDSAHAR) (Uttar Pradesh)	SATLUJ AT BOAT BDG. DHARMKOT-NAKODAR ROAD, JALANDHAR (Punjab)	SATLUJ AT BOAT BDG. DHARMKOT-NAKODAR ROAD, JALANDHAR (Punjab)
10.	GANGA D/S KANPUR (JAJMAU PUMPING STATION) (Uttar Pradesh)	GANGA AT BAHARAMPORE (West Bengal)	OKHLA BRIDGE (INLET OF AGRA CANAL) (Delhi)	GANGA AT PUNPUN, PATNA (Bihar)	GANGA AT PUNPUN, PATNA (Bihar)

Tab. 3: Spatial outlier detection by GWPCA (CS) method for different bandwidths

Bandwidth (bw)	**Total spatial outliers detected**
55%	54
65%	64
75%	70
85%	81
95%	92

Tab. 4: Top ten Spatial outliers detected by GWPCA (CS) method for different bandwidths

S.no.	**bw=55%**	**bw=65%**	**bw=75%**	**bw=85%**	**bw=95%**
1.	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)	BEAS U/S MANALI (Himachal Pradesh)
2.	BEAS D/S MANDI (Himachal Pradesh)	BEAS D/S MANDI (Himachal Pradesh)	SATLUJ D/S KIRATPUR SAHIB (Punjab)	SATLUJ D/S KIRATPUR SAHIB (Punjab)	BEAS D/S MANDI (Himachal Pradesh)
3.	SATLUJ D/S KIRATPUR SAHIB (Punjab)	SATLUJ D/S KIRATPUR SAHIB (Punjab)	SATLUJ U/S BUDHA NALLAH (UPPER) (Punjab)	SATLUJ U/S BUDHA NALLAH (UPPER) (Punjab)	SATLUJ D/S KIRATPUR SAHIB (Punjab)
4.	SATLUJ D/S NFL (Punjab)	SATLUJ U/S BUDHA NALLAH (UPPER) (Punjab)	GANGA AT KADAGHAT, ALLAHABAD (Uttar Pradesh)	GANGA AT KADAGHAT, ALLAHABAD (Uttar Pradesh)	SATLUJ U/S BUDHA NALLAH (UPPER) (Punjab)
5.	SATLUJ U/S BUDHA NALLAH (UPPER) (Punjab)	GANGA AT KALA KANKAR, RAEBARELI (Uttar Pradesh)	GANGA U/S VARANASI (ASSIGHAT) (Uttar Pradesh)	GANGA U/S VARANASI (ASSIGHAT) (Uttar Pradesh)	GANGA AT KADAGHAT, ALLAHABAD (Uttar Pradesh)
6.	GANGA AT BITHOOR (KANPUR) (Uttar Pradesh)	GANGA AT ALLAHABAD (RASOOLABAD) (Uttar Pradesh)	GANGA AT BUXAR (Bihar)	GANGA AT BUXAR (Bihar)	GANGA U/S VARANASI (ASSIGHAT) (Uttar Pradesh)
7.	GANGA AT DALMAU (RAI BAREILLY) (Uttar Pradesh)	GANGA AT KADAGHAT, ALLAHABAD (Uttar Pradesh)	GANGA AT BUXAR, RAMREKHAGHAT (Bihar)	GANGA AT BUXAR, RAMREKHAGHAT (Bihar)	GANGA AT BUXAR (Bihar)
8.	GANGA AT KALA KANKAR, RAEBARELI (Uttar Pradesh)	GANGA AT ALLAHABAD D/S (SANGAM), (Uttar Pradesh)	GANGA U/S KHURJI, PATNA (Bihar)	GANGA AT CONFLUENCE OF SONE DORIGANJ, CHAPRA (Bihar)	GANGA AT BUXAR, RAMREKHAGHAT (Bihar)
9.	GANGA D/S VARANASI (MALVIYA BRIDGE) (Uttar Pradesh)	GANGA U/S VARANASI (ASSIGHAT) (Uttar Pradesh)	GANGA AT CONFLUENCE OF SONE DORIGANJ, CHAPRA (Bihar)	GANGA U/S MOKAMA (Bihar)	GANGA AT CONFLUENCE OF SONE DORIGANJ, CHAPRA (Bihar)
10.	GANGA AT TRIGHAT (GHAZIPUR) (Uttar Pradesh)	GANGA D/S VARANASI (MALVIYA BRIDGE) (Uttar Pradesh)	GANGA DARBHANGA GHAT AT PATNA (Bihar)	GANGA AT MUNGER (Bihar)	GANGA U/S MOKAMA (Bihar)

Fig. 1: Water monitoring locations detected as spatial outliers by GWMD method for different bandwidths.

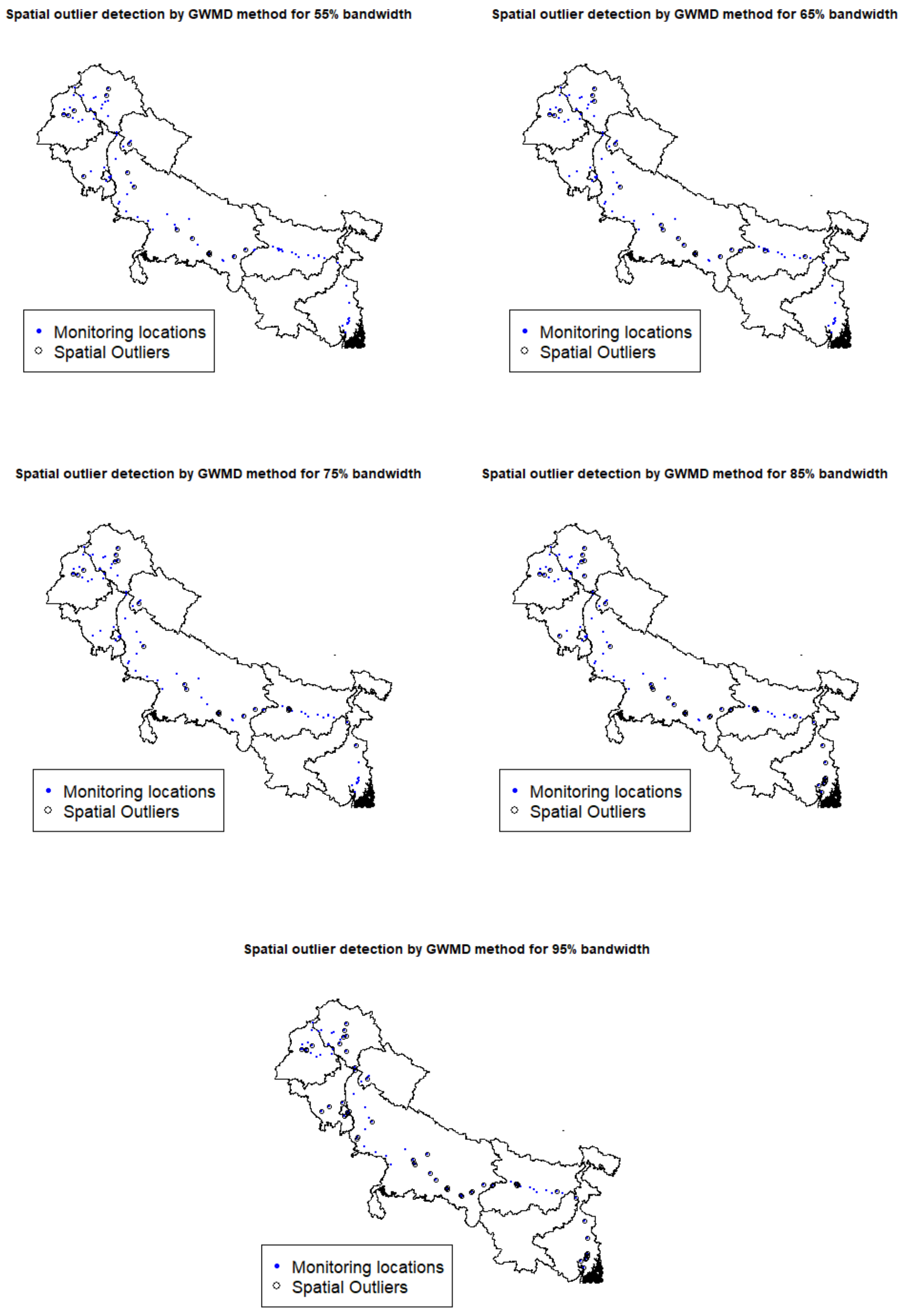
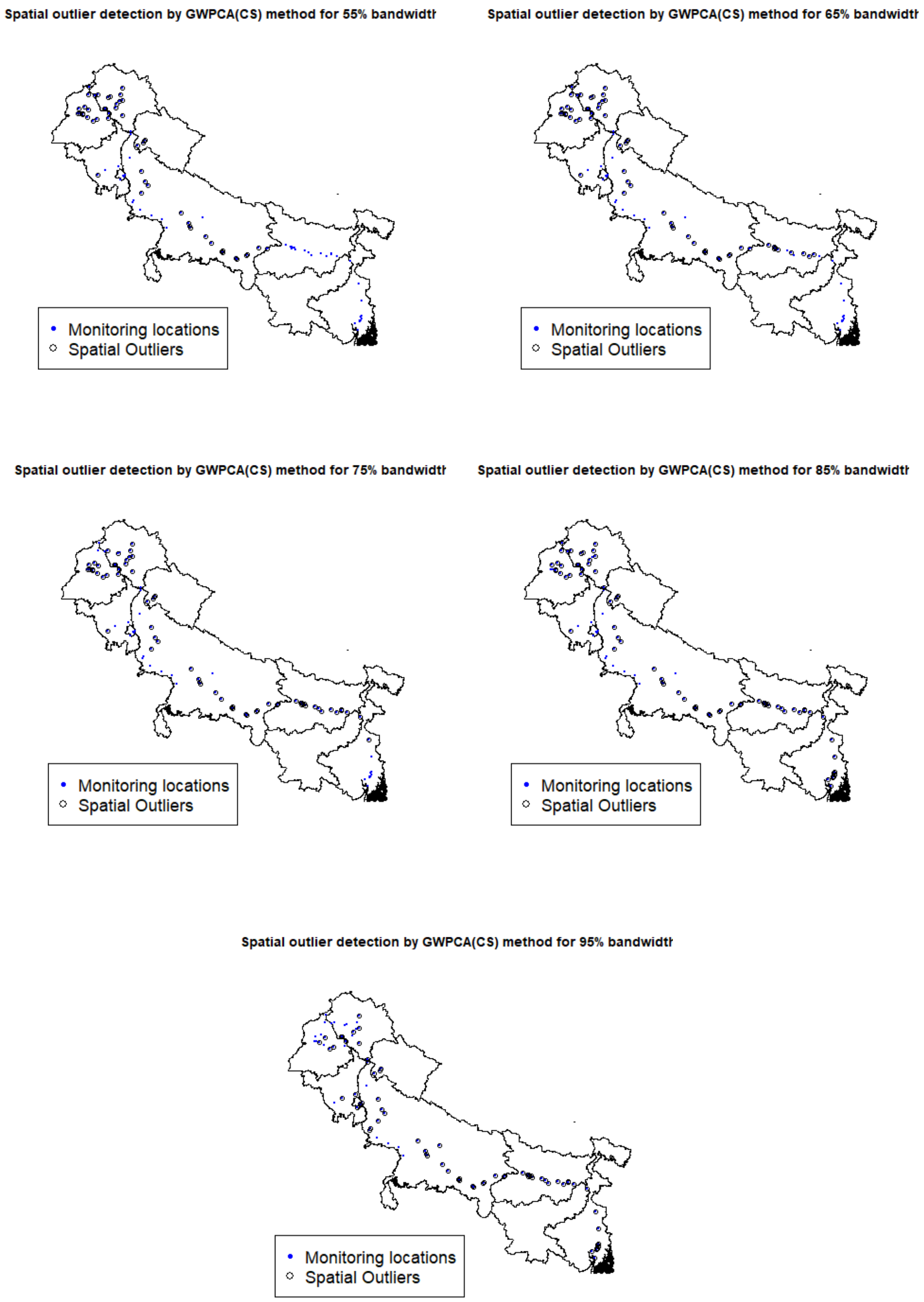

Spatial outlier detection by GWMD method for 55% bandwidth
• Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWMD method for 65% bandwidth
• Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWMD method for 75% bandwidth
• Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWMD method for 85% bandwidth
• Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWMD method for 95% bandwidth
• Monitoring locations
○ Spatial Outliers

Fig. 2: Water monitoring locations detected as spatial outliers by GWPCA_CS method for different bandwidths.


Spatial outlier detection by GWPCA(CS) method for 55% bandwidth
● Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWPCA(CS) method for 65% bandwidth
● Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWPCA(CS) method for 75% bandwidth
● Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWPCA(CS) method for 85% bandwidth
● Monitoring locations
○ Spatial Outliers
Spatial outlier detection by GWPCA(CS) method for 95% bandwidth
● Monitoring locations
○ Spatial Outliers

It can be seen clearly from Table 1 and Table 3, that the total number of spatial outliers detected vary a lot for the algorithms. The GWMD method declares a reasonable number of spatial units as spatial outliers, while in the GWPCA (CS) method more than 50% of spatial units are being declared as spatial outliers which shows high swamping errors (non-outliers that were classified as outliers). For the GWMD method as seen in Table 2, it performs well for the bandwidth 75% because it detects spatial outliers from almost all the states and their respective local MDs are also very high. Figure 1 and Figure 2, plots all the water monitoring stations in the map and respective spatial outliers detected by both GWMD and GWPCA (CS) methods. In Figure 2, it can be seen that the spatial outliers detected get clustered in an area, whereas for the GWMD method in Figure 1, it can be seen that the spatial outliers detected are well spaced over the entire area.

4 Conclusion

Two geographically weighted methods GWMD and GWPCA (CS) are applied to a water quality dataset and spatial outliers are detected for different bandwidths. For both the algorithms top 10 spatial outliers are listed in Table 2 and Table 4. For all the bandwidths both the algorithms detect the water monitoring location BEAS U/S MANALI located in Himachal Pradesh as the top-most outlier. Manali is one of the top tourists' destinations in India and is affected by the heavy water pollution in turn. The government should look upon such areas and take preventive measures for same by educating the people about the water pollution causes, reasons and how to help in maintaining the water quality of the stations up to the mark. Water Pollution is a big concern at present time causing health problems, environmental and ecological problems. Since the water quality data is dependent on the geographical area and thus on geo-locations the detection of spatial outliers helps in detecting a local instability in an area and hence, the application of geographically weighted methods to such data is quite appropriate for the detection of such water monitoring locations where the attention and involvement of the government is strongly required.

References

[1] Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1996): Geographically weighted regression: a method for exploring spatial nonstationarity, Geog. Anal., 28(4):281-298.

[2] Brunsdon, C., Fotheringham, A.S. and Charlton, M. (1998): Spatial nonstationarity and autoregressive models, Environ.PlanA., 30(6):957-973

[3] Brunsdon, C., Fotheringham, A.S. and Charlton, M. (2002): Geographically weighted summary statistics a framework for localised exploratory data analysis, Comput. Environ. Urban Syst., 26(6), 501-524.

[4] Lu, C.T., Chen, D. and Kou, Y. (2004): Multivariate spatial outlier detection, Int. J. Artif. Intell. Tools, 13(04), 801-811.

[5] Lu, C.T., Chen, D., Kou, Y. and Chen, F. (2008): On detecting spatial outliers, Geoinformatica, 12(4), 455-475.

[6] Hawkins, D.M. (1980): Identification of outliers, vol 11, Springer.

[7] Mason, G., Jacobson, R. (2007): Fuzzy geographically weighted clustering, In: Proceedings of the 9th International Conference on Geocomputation, 1998, pp 1-7.

[8] Dykes, J. and Brunsdon, C. (2007): Geographically weighted visualization: interactive graphics for scale varying exploratory analysis, IEEE TransVis. Comput. Graph., 13(6), 1161-1168.

[9] Anselin, L. (1994): Exploratory spatial data analysis and geographic information systems, In: Painho M (ed) New tools for spatial analysis. Eurostat, Luxembourg, pp 45-54.

[10] Anselin, L. (1995): Local indicators of spatial association - LISA, Geog. Anal.,27(2), 93-115.

[11] Anselin, L. (1996): The moran scatterplot as an esda tool to assess local instability in spatial, Spatial Analytical, 4:111.

[12] Li, L. (2019): Geographically weighted machine learning and downscaling for high resolution spatiotemporal estimations of wind speed, Remote Sens., 11(11), 1378.

[13] Worboys, M. (1995): Next-Generation Systems, GIS A Computing Perspective, pp.331-332.

[14] Foley, P. and Demar, U. (2013): Using geovisual analytics to compare the performance of geographically weighted discriminant analysis versus its global counterpart, linear discriminant analysis, Int. J.Geogr. Inf. Sci., 27(4), 633-661.

[15] Harris, P., Brunsdon, C. and Charlton, M. (2011): Geographically weighted principal components analysis, Int. J.Geogr. Inf. Sci., 25(10), 1717-1736.

[16] Harris, P., Brunsdon, C., Charlton, M., Juggins, S. and Clarke, A. (2014): Multivariate spatial outlier detection using robust geographically weighted methods, Math. Geosci., 46(1), 1-31.

[17] Rousseeuw, P.J. and Leroy, A.M. (2005): Robust regression and outlier detection, vol 589. John Wiley & Sons.

[18] Shekhar, S., Lu, C.T. and Zhang, P. (2001): Detecting graph-based spatial outliers: algorithms and applications (a summary of results), In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp 371-376.

[19] Shekhar, S., Lu, C.T. and Zhang, P. (2003): A unified approach to detecting spatial outliers, GeoInformatica, 7(2), 139-166.

[20] Nakaya, T., Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2005): Geographically weighted poisson regression for disease association mapping, Stat.Med., 24(17), 2695-2717.

[21] Barnett, V. and Lewis, T. (1984): Outliers in statistical data, Osd..