International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Grid and Density Based Adaptive Clustering Algorithm for Spatio-Temporal Data Mining

Swati Meshram[1], Kishor P. Wagh[2]

Submitted: 26/01/2024 **Revised**: 04/03/2024 **Accepted**: 12/03/2024

Abstract: The Indian subcontinent experiences seismic activities which are visualized in India's Seismic map. These seismic spatio-temporal characteristics need to analyze to understand the evolution. Clustering is a machine learning technique to highlight the patterns of grouping similar objects in the spatio-temporal dimensional. Our research work in this paper proposes a novel algorithm to analyse the spatio-temporal data for patterns through clustering. This is a hybrid method based on grid and density clustering. We have devised a method to find the required total number of core points for density clustering. The efficiency of our algorithm is higher due to appropriate selection of core points with respect to the density in the region. In addition, proposed algorithm requires minimal user defined parameters and minimizes Euclidean distance computation to the neighboring core points in the current region and not with all of the core points. The algorithm has been experimentally tested for correctness of results and performance. It is observed from the results, the Earthquake spatio-temporal data has clustering tendency and the events indicate higher correlation with respect to frequency and time. The quality of clustering is effective and efficient with the silhouette index 0.93.

Keywords*: Clustering, core points, seismology, spatio-temporal data, pattern mining*

1. Introduction:

The progress made in machine learning has proven to be effective in handling the massive spatio-temporal data. Man-made intelligence models are crucial in handling the multidimensional aspects of spatio-temporal data with faster response. Spatial represents geographical coordinates and temporal relates to time information of events or data recording along with non-spatiotemporal information. These data is constantly created from various sources like GPS, satellites, sensors, robots etc. The ease in availability of massive but unpredictable spatio-temporal data presents troubles in deriving useful patterns of information. Machine learning models offer a solution for the process of

extraction and validating patterns or rules over the data.

Clustering is a classical machine learning technique performed by grouping similar spatio-temporal objects for mining patterns, trends, rules hidden in data[1]. Clustering in spatio-temporal environment mainly rely on particular characteristics of data which could represent a point on geographical coordinate system, a line representing trajectory or roads, a polygon representing a bounded region etc. So the spatial information records a fixed location or for moving objects, the change in location over time.

[1] *Government College of Engineering, Amravati, India.*
ORCID ID : 0000-0002-7569-7995
[2] *Government College of Engineering, Amravati, India.*
ORCID ID : 0000-0002-8189-3378
**Corresponding Author Email:*
swati.meshram@computersc.sndt.ac.in

The temporal dimension captures the time when the event of recording the information occurred or evolution in time. Then

the objective of clustering is to extract a population of data by a joint distribution or distance metric computation where clustering objects or instances have minimal distance with the centroids of the cluster. The overall clustering process is as given in Fig.1. as: acquiring the dataset and perform preprocessing or cleaning of data. Apply the clustering algorithm to generate clusters. Compute the distance metric for obtaining the similarity ratio and for determining neighbor instances. The clusters are validated and results are interpreted with drawing meaningful information at this stage. The widely known distance metrics are Euclidean, Manhattan, Minkowski etc. Clustering has applications in recommendation of product based on customer reviews, crime clusters, with NLP to retrieve similar documents, diseases analysis, event detection.

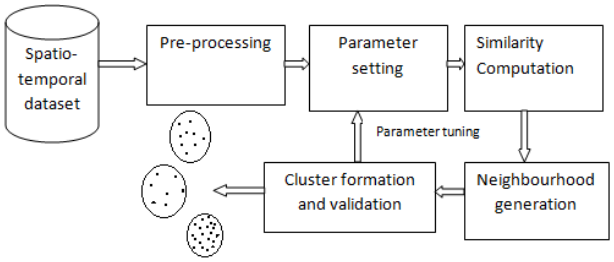

Fig.1. General Process of Spatio-Temporal Clustering.

Spatiotemporal data types can be categories into events, time series, moving objects, trajectories, and geo-referenced variables [9]. Spatio-temporal event data is recorded for a location with corresponding timestamp. The spatial and the

temporal information is static and no movement of data is recorded. In case of epidemics, affected region may change the size. Earthquakes, cyclone are good example of data representing events. In the time series data, the object evolution with respect to time is captured and recorded. A long history of the object evolution is stored. Then analysis for correlation in different geographical time series is performed. In case of moving objects, spatial location of the object changes with change in time. Moving vehicles monitoring can be a good example for moving objects data collection. Trajectories involve recording the history of moving objects in sequence. Clustering analysis for similar movement behavior may be carried out. Geo-referenced variable records the evolution of the phenomena at a fixed location.

Clustering has applications in recommendation of product based on customer reviews, crime clusters, with NLP to retrieve similar documents, diseases analysis, event detection. A famous application of clustering recorded sucessfully in 1854 was for disease spread analysis, where Dr John Snow established correlation to form clusters of cholera cases found around a public water pump, and was the source of the spread of cholera.

One of the applications of clustering is detection of earthquake clusters of same severity, regions of clustering displaying foreshocks and aftershocks of main earthquake events. Earthquakes are the natural events that cause tremors from the Earth's core to the surface. These sudden vibrations may cause destruction of useful natural and man-made resources. Leading to the identification of such areas, which may have a trend or reach of Earthquake impacts, is of importance. Hence, we focus our study to derive clustering patterns through our proposed rese arch work on Indian Earthquake spatio-temporal data.

Aim:

The primary aim that this study addresses is developing a clustering algorithm for spatio-temporal data for pattern mining, which offers better performance and quality in event-based setting.

Our contribution in the paper covers the follow objective:

1. Develop a hybrid clustering algorithm to derive patterns in spatio-temporal data.
2. To state a method to select appropriate core points' from the dataset.
3. To minimize the user defined parameters to be used for the algorithm.
4. To experimentally test the algorithm on spatio-temporal Indian Earthquake dataset.

The remaining paper follows the structure as: Section II presents the study of the state of art research works. Section III presents the proposed method. Section IV carries out the result and discussion and Section V puts forth the conclusion of the research work.

2. Related Work:

The approaches to clustering are highlighted in table 1. K-means[2], a partitioning based approach extracts k partitions iteratively to form clusters using Euclidean type of distance metric. The method being simple and interpretable is widely applied. The results are influenced by outliers. Density based approach identifies maximal density connected regions to form clusters based on neighborhood where parameters like maximum distance and minimum number of datapoints to form clusters are used. It results into arbitrary shaped clusters. DBSCAN[3] algorithm works on density-based strategy. Grid based methods adopt formation of grid of non overlapping bins and are determines dense grid bins to form clusters. CLIQUE[4] is a well-known grid-based algorithm. MAFIA[5] is an algorithm of this approach. Hierarchical uses top down or bottom-up approach along with linkages is incorporated for clustering. Chameleon[6-7] algorithm is a type of hierarchical clustering method. Trajectory clustering has an application in animal movement. TRACLUS[8] is a trajectory clustering approach. In this trajectory are partition into line segments and grouped to form clusters. Hybrid approach uses two or three approaches to utilize their benefits.

Table 1: Traditional approaches to spatio-temporal clustering.

Clustering Approach	Strategy	Characteristics
Partitioning	Distance, centroids.	Outliers influence the results.
Density	Neighborhood, core points,density reachable points.	Minimizes noise.
Grid	Non-overlapping bins, dense spaces, buffer area.	Easy to parallelize, incremental
Hierarchical	Agglomerative, divisive, linkages.	Sensitive to outliers
Trajectory	Group lines of trajectories. Angular distance.	Efficient due to indexing.
Hybrid	A mix of above approaches.	Robust, scalable, flexible.

The state-of-art research work applied in spatio-temporal clustering has been in the applications of analyzing the impact of disease spread like COVID-19 etc in [11][12].The

disease mapping and modeling is required to identify high-risk clusters which requires distinguishing high-risk areas from low-risk areas and smoothing the relative risk as given in [13]. In [11], Moran's local and global I is computed to find the autocorrelation to identify hot spot regions through clustering. Here variation in disease spread rate was highly related to the determinants as income of families and foreign visitors in the country. In [12], study is based on Kulldroff's time space scanning statistics along with Poisson probabilistic model, logistic regression for evaluation of results. In [13], the case study of Dengue disease is carried out, where agglomerative hierarchical clustering is applied for mapping of clusters. Further with spatiotemporally varying coefficient using Bayesian model, optimal clusters are selected. It finds the optimal cluster configuration. It was found risk of Dengue varies with space and time and is related to weather variables.

In [14], hierarchical trajectory clustering framework has been developed using semantic information such as speed, direction, and time. Clustering is based on single linkage between clusters. This method uses reference spots and long sequences of trajectories.

Spatio-temporal clustering can be carried out with geographical weighted regression as studied in [15-16].Some use varying coefficient model as given in [17]. The varying coefficient in generalized linear models along with hierarchical Bayesian method is used latent random variables analysis.

Kernel density estimation is widely used non-parametric method for discovering high density geographical events. The most efficient kernel function is Gaussian kernel function [18-19].

In [20], study of flow of vehicles for traffic congestion condition is carried out by computing travel time index (TTI) for zones while k-means with Naïve Bayes classifier is applied for the clustering process and analysis.

3. **Proposed Methodology**:

The objective of spatio-temporal dataset clustering on Indian Earthquake dataset is to uncover correlation or associations properties of inside the data. Application of clustering computations is to recognize regions that show practically identical seismic traits. This would be useful in understanding distribution of events in the Indian subcontinent.

Dataset and Study Area

The dataset of earthquake was selected for experiment as the data contains this gives a good collection of spatio-temporal events. The analysis is required for evolution of clusters. The earthquake dataset was obtained from the portal of National Centre of Seismology, ministry of Earth Sciences, Government of India for the events from 2019 to Jan 2024 with 6506 events. The events represent information as the date and time, location as longitude and latitude, with non-spatial features as magnitude and depth of earthquakes. The data captured is for the region covering 0^0 to 40^0 N in latitude and 60^0 to 100^0 in longitude.

Pre-Processing of Raw Data

The events data obtained from the website contained some attributes that have not been used for the clustering process like the comments attribute. This attribute was dropped. The Temporal information was in timestamp format, where time was represented in the form of date, hours, minutes, seconds which had to be converted to seconds and then further used in next stages of the processing.

Grid Formation and Allocation of Data Instances to Grid.

While forming the virtual grid, focus is to ascertain the range of longitude and latitude coordinates, temporal range. In order to have every grid bin to be equally spaced as per the data coordinates, the grid coordinates have to be determined. For this process, division of coordinates range is performed. It is necessary to locate the neighbouring bins in all direction of the current grid which is completed her. After setting the grid further stage requires allocation of data instances to the grid bins as per its coordinates. This process is simple as it only checks the instance coordinates and it falls in the range of respective grid bin coordinates.

Density of Grid Bins

Further stage requires to find the density population of every grid bin. The density decides to set the core point in the grid and number of core points for the grid bin. If the density is found to near to zero, no core points are selected. Then there is increase, in the number of core points as per the density. The assignment of core points is controlled by the density of the bin. In order to have selection of the core points a random process selection is applied. In order to not have clusters too close there is distance maintained between core points.

Distance metric

The approach used here is using the Haversine distance formula to determine the spatial distance between any two events location. Haversine distance is the Earth surface distance between any two location based on latitude and longitude coordinates. The temporal distance is computed based taking time difference between two events. This requires conversion of time into seconds and finding the lag. Since time difference will outweigh the spatial distance, we use a smoothing factor weight with time attribute to balance the spatial distance and temporal distance effect.

Clustering

Assign the data instances to the closest grid bin or neighboring bins core points based on minimal distance

metric including the neighboring bin. In order to not influence the results by bin boundaries, the algorithm obtains the closet core point which may belong to its own bin or neighboring bins. At this stage the clusters are formed. Determine the mean within the cluster distance. Determine the mean across the cluster distance.

Validation of results

Every cluster is interpreted. We have checked are the clusters overlapped or wrongly assigned. Compute the clustering quality measure silhouette index. Set the clusters can be made final clusters or go for optimization of parameters.

Proposed algorithm for spatio-temporal clustering:

Step 1: Read the data and find the total number of events as objects n.

Step 2: Find the grids extreme geographical coordinates from data, to construct a virtual grid(i.e. latitude, longitude).

Step 3: Form the geographical coordinates grid containing bins.

Step 4. For every grid bin, find neighbors grid bins that surround the grid in all directions.

Step 5: Assign the data points to grid bins.

Step 6: Find the density of every grid bin.

Step 7: Classify the grid bin as per the density in the category as below minimum threshold, above minimum threshold, above average.

Step 8:Find the number of core points for every bin.

$$|corepoints| = \log_2|gridbindatapoints| + \vartheta \quad (1)$$

where ϑ takes value in the range [1,3].

Step 9:Allocate core points to the grid bin using random process.

Step 10: Maintain minimum distance 'δ' between core points of the same grid bin.

Step 11: Allocate the data points to the nearest grid core points using spatio-temporal distance inspired by Haversine formula.

$$\text{Haversine}(I_i,I_j) = 2 * \text{Radius of Earth} * \arcsin\sqrt{\text{Sin}^2(\text{I}_\text{j}.\text{Lat} - \text{I}_\text{i}.\text{Lat}) + \text{Cos}(\text{I}_\text{i}.\text{Lat}).\text{Cos}(\text{I}_\text{j}.\text{Lat}) + \text{Sin}^2(\text{I}_\text{j}.\text{Lon} - \text{I}_\text{i}.\text{Lon})} \quad (2)$$

$$\text{Time}(I_i,I_j) = \text{I}_\text{i}.\text{time} - \text{I}_\text{j}.\text{time} + 1 \quad (3)$$

Distance = Haversine distance +time distance (4)

Step 12: Form the clusters.

Step 13:Compute the silhouette index for cluster quality.

Step 14: Stop

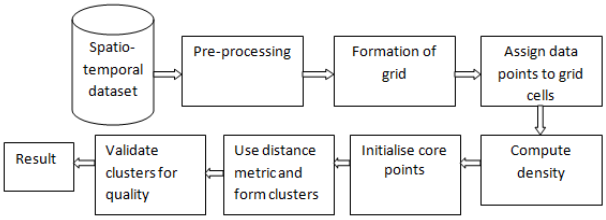

Fig 2: Proposed Spatio-temporal Clustering framework.

The algorithmic framework has been diagrammatically represented in Fig.2.

4. Result and Discussion*:*

The proposed algorithm was implemented in python programming language on Colab platform of Google which provides cloud storage. The algorithms works in stages. First, formation of grid and assignment of datapoints to grid. Second, density computation of grid bins and allocation of grid core points. Third, computing distance between datapoints and assignment to cluster. Last, computing cluster quality score.

Fig.3. shows that selection of core points within the grid by maintaining the distance implements to not have overlapping clusters and Fig.4. shows result of core point selection. Figures 5 and 6. Shows the clustering results with different colors as different clusters. Figure 7, shows the computation silhouette index score reflecting quality of clustering as 0.93 which is a good clustering quality. Figures 8 and 9 illustrate the effect of altering the minimum distance between the core points/centroids given by δ influences the number of core points/centroids selected and the resulting mean distances both between and within clusters. As δ increases, fewer centroids are chosen, leading to an increase in mean distances between clusters and within clusters. These observations indicate that the resulting clusters are distinct, non-overlapping, and possess arbitrary shapes.

The proposed algorithm is an improvement over existing methodology based on distance computation. Firstly, it avoids the expensive computation of distances between individual data points and all core points by utilizing distance calculations with nearby core points to identify clustering patterns. Secondly, it narrows down the search scope by employing neighborhood grid bins and exploring proximity in all directions. Thirdly, the grid bin-based approach does not restrict cluster formation solely to the shape of the grid bin, as the nearest core points may be found

in neighboring grid cells. Additionally, the algorithm does not repeatedly alter cluster assignments iteratively but instead focuses on achieving accurate assignments initially. Moreover, it operates independently of prior knowledge or background information regarding clusters, core points, or neighborhoods.

The algorithm's effectiveness is highlighted by experimental findings, as it partitions the dataset into spatial-temporal space bins and focuses on a limited number of grid bins. This approach reduces unnecessary searches and calculations of spatio-temporal distances calculation, resulting in a compact search space and minimal memory usage.

```
count=0
for j in range(len(X)):
  if(Bucket[j]==i and count<centroid_count_bucket[i]):
    if(count<1):
      centroid.append(X[j])
      centroid_bin.append(i)
      previous_centre_lat=X[j][1]
      previous_centre_long=X[j][2]
      count=count+1
    else:
      Dont_append_flag=1
      for l in range(len(centroid)):
        if(distance_d1(centroid[l][1],centroid[l][2],X[j][1],X[j][2])<300):
          Dont_append_flag=0
          break
      if(Dont_append_flag):
```

Fig. 3: Maintain minimum distance between core points.

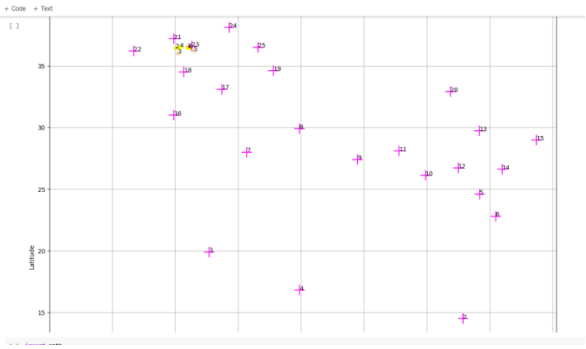

Fig. 4: Core points selection.

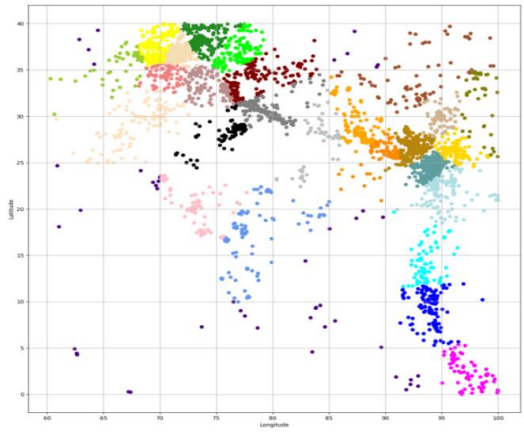

Fig. 5: Execution of proposed spatio-temporal clustering with formation of 25 clusters, distance between the core points is 300km.

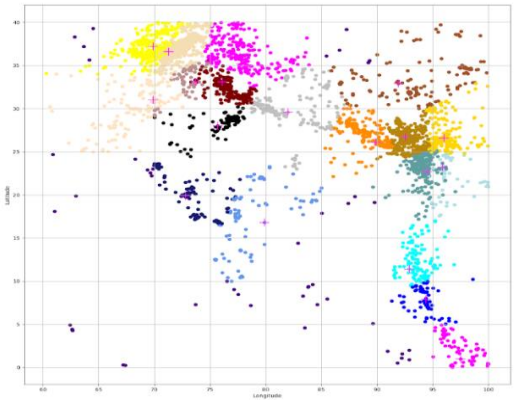

Fig. 6: Execution of proposed spatio-temporal clustering with formation of 19 clusters, distance between the core points is 500km

```
[3] Silhoutte_index=0
    value1=0
    value1=max(Cluster1_dist_avg3,avg_dist_bet_clusters)
    Silhoutte_index=(avg_dist_bet_clusters-Cluster1_dist_avg3)/value1
    print(Silhoutte_index)

    0.933783428047153
```

Fig.7: Silhoutte index for clustering quality.

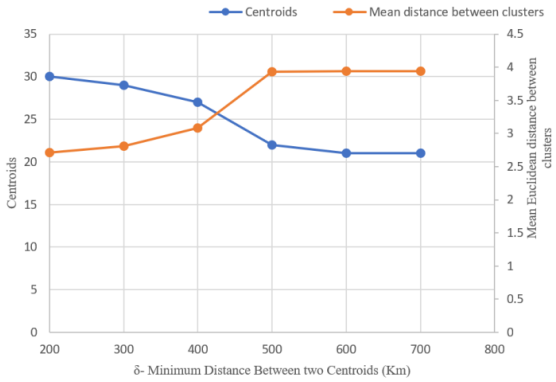

Fig. 8: Effect of change in maximum distance between two centroids on number of centroids selection and mean distance between the clusters.

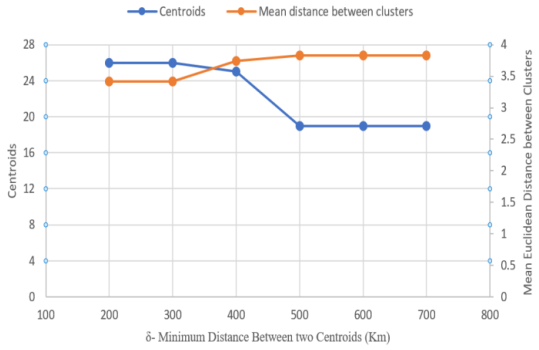

Fig. 9: Effect of change in maximum distance between two core points/centroids on number of centroids selection and mean distance between the clusters.

5. Conclusion:

In this research article we have proposed an algorithm for clustering patterns on spatio-temporal data. The algorithm has been applied on the Earthquake data of Indian subcontinent and resulted in 27 clusters. The experimental evaluation ensures with the clustering results effective performance of the algorithm. We conclude to use the proposed clustering algorithm as it offers reduction in distance computation and good clustering quality with silhouette index as 0.93.In future recent advances in deep learning such as graph convolution networks can be explored for clustering approach.

Acknowledgements: Nil

Author contributions :

First author- conceptualization, programming, paper writing.

Second author – review and supervision.

Conflicts of interest

The authors declare no conflicts of interest.

References:

[1] S. Meshram and K. P. Wagh, "Mining Intelligent Spatial Clustering Patterns: A Comparative Analysis of Different Approaches," in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2021, pp. 325–330

[2] Pelleg, Dan; Moore, Andrew (1999). "Accelerating exact k -means algorithms with geometric reasoning". Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States: ACM Press. pp. 277–281. doi:10.1145/312129.312248. ISBN 9781581131437. S2CID 13907420

[3] Ester, Martin; Kriegel, Hans-Peter; *Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.).* A density-based algorithm for discovering clusters in large spatial databases with noise (PDF). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).* AAAI Press. *pp. 226–231*. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.

[4] Palla, Gergely; Derényi, Imre; Vicsek, Tamás (2006). "The Critical Point of k-Clique Percolation in the Erdős–Rényi Graph". Journal of Statistical Physics. 128 (1–2): 219–227.

[5] S. Goil, H. Nagesh, and A. Choudhary, "Mafia: efficient and scalable subspace clustering for very large data sets," Technical Report CPDC TR-9906-010, Northwestern University, 1999

[6] G. Karypis, E. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," Computer, vol. 32, no. 8 pp. 68– 75, 1999

[7] T. Barton, T. Bruna, and P. Kordik, "Chameleon 2: An Improved Graph-Based Clustering Algorithm," ACM Transactions on Knowledge Discovery from Data, vol. 13, no. 1, 2019

[8] Jae-Gil Lee, Jiawei Han, Kyu-Young Whang," Trajectory Clustering: A Partition-and-Group Framework∗", SIGMOD'07, June 11–14, 2007, Beijing, China.

[9] Kisilevich, S., Mansmann, F., Nanni, M. and Rinzivillo, S., 2010. Spatio-temporal clustering (pp. 855-874). Springer US.

[10] Shi Z, Pun-Cheng LSC. Spatiotemporal Data Clustering: A Survey of Methods. *ISPRS International Journal of Geo-Information*. 2019; 8(3):112. https://doi.org/10.3390/ijgi8030112

[11] .M. Siljander, R. Uusitalo, P. Pellikka, S. Isosomppi, O. VapalahtiSpatiotemporal clustering patterns and sociodemographic determinants of COVID-19 (SARS-CoV-2) infections in Helsinki, Finland"Spat. Spatio-Tempor. Epidemiol., 41 (2022), Article 100493.

[12] S.-Q. Yang, Z.-G. Fang, C.-X. Lv, S.-Y. An, P. Guan, D.-S. Huang, W. Wu,"Spatiotemporal cluster analysis of COVID-19 and its relationship with environmental factors at the city level in mainland China", Environ. Sci. Pollut. Res., 29 (9) (2022), pp. 13386-13395.

[13] Jaya,I.G.N.M. and Folmer, H. (2021), Identifying Spatiotemporal Clusters by Means of Agglomerative Hierarchical Clustering and Bayesian Regression Analysis with Spatiotemporally Varying Coefficients: Methodology and Application to Dengue Disease in Bandung, Indonesia. Geogr Anal, 53: 767-817. https://doi.org/10.1111/gean.12264

[14] D. Zhang, K. Lee, I. Lee,"Hierarchical trajectory clustering for spatio-temporal periodic pattern mining",Expert Syst. Appl., 92 (2018), pp. 1-11

[15] Fotheringham, S., C. Brunsdon, and M. Charlton. (2002). Geographically Weighted Regression, The Analysis of Spatially Varying Relationships. New York, NY: Wiley.

[16] Ndiath, M., B. Cisse, J. L. Ndiaye, J. Gomis, O. Bathiery, A. Dia, and B. Faye. (2015). "Application of GeographicallyWeighted Regression Analysis to Assess Risk Factors for Malaria Hotspots in Keur Soce Health and Demographic Surveillance Site." Malaria Journal 14(463), 1–11.

[17] Gelfand, A., H. J. Kim, C. Sirmans, and S. Banerjee. (2003). "Spatial Modeling With Spatially Varying Coefficient Processes." Journal of the American Statistical Association 98(462), 387–96.

[18] Scott, D.W. Multivariate Density Estimation: Theory, Practice, and Visualization; John Wiley & Sons: Hoboken, NJ, USA, 2015
[19] Silverman, B.W. Density Estimation for Statistics and Data Analysis; CRC Press: Boca Raton, FL, USA, 1986; Volume 26.
[20] Naoufal Rouky, Abdellah Bousouf, Othmane Benmoussa, MouhseneFri," A spatio temporal analysis of traffic congestion patterns using clustering algorithms: A case study of Casablanca" Decision Analytics Journal24 January 2024