

Dynamic spatio-temporal pattern discovery: a novel grid and density-based clustering algorithm

Swati Meshram¹, Kishor P. Wagh²

¹Department of Computer Science and Engineering, Government College of Engineering, Amravati, India

²Department of Information Technology, Government College of Engineering, Amravati, India

Article Info

Article history:

Received Mar 22, 2024

Revised Jul 9, 2024

Accepted Jul 26, 2024

Keywords:

Centroids

Density

Distance

Earthquake dataset

Neighborhood

ABSTRACT

Clustering is a robust machine-learning technique for exploration of patterns based on similarity of elements over multidimensional data. Spatio-temporal clustering aims to identify target objects to mine spatial and temporal dimensions for patterns, regularity, and trends. It has been applied in human-centric applications, such as recommendation systems, urban development and planning, clustering of criminal activities, traffic planning, and epidemiology to identify the extent of disease spread. Although the existing research work in the field of clustering relies widely on partition and density-based methods, no major work has been carried out to handle the spatiotemporal dimension and understand the dynamics of temporal variation and connectivity between clusters. To address this, our paper proposes an algorithm to mine clustering patterns in spatiotemporal dataset using an adaptive, dynamic hybrid technique based on grid and density clustering. We adopt spatio-temporal partitioning of the virtual grid for distribution of data and reducing distance computation and increasing efficiency. Grouping the higher density regions along with neighborhood cluster density attraction rate to merge the clusters. This method has been experimentally evaluated over the Indian earthquake dataset and found to be effective with clustering silhouette index up to 0.93.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Swati Meshram

Department of Computer Science and Eng, Government College of Engineering

Amravati, Maharashtra, India

Email: swati.meshram@computersc.sndt.ac.in

1. INTRODUCTION

Advancement in computer technology, remote sensing, and location-based services has resulted in the generation of massive spatiotemporal data. Spatiotemporal data analysis is an emerging research area driven by development and application of intelligent computational techniques. Analyzing spatiotemporal data is beneficial for various human-centered applications like recommendation systems, identifying disease outbreak patterns, urban development clustering, infrastructure planning, and detection of criminal activities. Clustering is a valuable analysis tool for exploring and understanding rich information contained in the spatio-temporal datasets.

The aim of analyzing spatiotemporal data clustering is to detect and examine noteworthy patterns in the data that change both in space and time and help in understanding the dynamics or processes driving the patterns and trends. Anomalous patterns are detected to indicate rare but significant events with deviation from expected behavior. The analysis may also be used to design models for predicting future occurrences of similar events. It provides insights in evolution of cluster and its changes over time which is a valuable information in understanding trends. This may lead to proactive decision-making and improved resource allocation. Understanding the impact of human activities and natural processes and their interconnections. This is highly

relevant in the field of emergency response, public health, monitoring environmental variables, and urban resource allocation. Spatiotemporal clustering analysis allows researchers to uncover complex relationships and make informed decisions in dynamic and interconnected systems [1]. Spatio-temporal data contain the geographical location and temporal or time of occurrence of the events along with other non-spatiotemporal features describing the events. Spatio-temporal data clustering analysis is a machine learning technique to search patterns in a dataset by grouping data instances based on similarity measures. The intracusters instances exhibit high similarity whereas the same instances are incoherent with instances of other clusters to form distinct clusters. In other words, clustering, is an unsupervised classification technique that separates an unlabeled data set into a finite number of groups whose members are data instances that are more homogeneous to its group than to other groups. These groups are termed as clusters. Thus, clustering as a process is illustrated by the following example:

Given an input data set $X = \{x_1, x_2, \dots, x_n\}$, where each x_i has set of j features or dimensions. We attempt to derive 'k' clusters given as $C = \{c_1, c_2, \dots, c_k\}$ satisfying the following conditions. For all $i, j \in \{1, \dots, k\}$, each $|c_i| > 0$, and $C_i \cap C_j = \emptyset$, and $X = \bigcup_{i=1}^k C_i$.

Clustering has been applied in recommendation systems by reviewing customer feedbacks for product popularity. Clustering as a tool is also useful for observing the abnormal behavior of outliers that do not exhibit the same relationship as that of other clusters. This analysis helps detect rare but important patterns in urban planning [2], big climate data analytics [3]. One of the applications of clustering we tend to explore is detection of earthquake clusters of same severity, regions of clustering displaying foreshocks and aftershocks of main earthquake events. Earthquakes are natural events that cause tremors from Earth's core to the surface. These sudden, vibrations may destroy useful natural and man-made resources. Identification of such areas, which may have a trend or reach of earthquake impacts using machine learning pattern mining techniques is important. Hence, we focus our study on deriving clustering patterns through our proposed research work on Indian earthquake spatiotemporal data. We highlight our contribution in this research article as follows: i) a method for selection of centroids; ii) a method to convert tentative clusters to fixed clusters based on density; iii) outlier score and clustering quality; iv) detecting spatio-temporal referenced variables with respect to evolution over time; and v) the proposed algorithm is implemented and experimentally validated.

Our research paper adheres to the following structure: section 2 explores related literature. Section 3 outlines the methodology. Section 4 presents the results and subsequent discussion. Finally, section 5 provides the conclusion of the research.

2. RELATED WORK

The clustering distance-based method computes a distance metric to measure the spatial distance and cluster similar or neighbouring points. The distance metric used are Euclidean distance, dynamic time warping [4], longest common subsequence (LCSS) [5], edit distance on real sequence (EDR) [6], Hausdroff [7], and Fréchet [8] distance. Density based clustering performs the grouping of density satisfying regions into clusters [9]. Feature-based clustering, first extracts the features and then computes their similarity [10], [11]. Time series data analysis using kernel density was the work undertaken to develop the algorithm spatio-temporal density-based spatial clustering of applications with noise (ST-DBSCAN) [12]. A spike neural network architecture is developed to cluster spatiotemporal brain data [13]. Guo *et al.* [14] analysed foodborne diseases on people of Zhejiang has been studied using spatiotemporal clustering which includes methods such as statistical and spatial analysis along with spatiotemporal scanning. Here the temporal resolution found is large. Loiola *et al.* [15] explored a hybrid burned area algorithm based on moderate resolution imaging spectroradiometer (MODIS) thermal anomalies and NIR reflectance with spatial resolution of 250 m on MODIS data. Hotspots clusters were developed to discover fire active areas. All these studies reflect the algorithms are developed to tackle specific problems and thus they have limited applicability. Gong *et al.* [16] put forth a model that learns the dynamics of mobility in taxi trajectory data and uses it to predict mobility in specific route areas. Here the area of study is confined to a particular area. Another article on trajectory clustering is studied in [17], it identifies clusters based on Hausdroff distance in K-nearest neighbour method where the accuracy of the method heavily relies on appropriate value of 'K'. The research in [18], [19] trajectory analysis was used to extract road traffic, determine flow statistic, and detect congestion. According to Georgoulas *et al.* [20], a hybrid approach of clustering over seismic spatio-temporal data was adopted. It is based on density and hierarchical agglomerative clustering which extracts objects with unknown class labels. Here connectivity is based on single linkage to form the clusters and no emphasis is placed on the temporal parameter of the dataset. According to Nazia *et al.* [21], space time clusters were discovered using geographically weighted regression model. The model also utilizes local and global Moran's I to interpret the cluster distribution pattern. The model was verified using COVID dataset. Another work on COVID dataset is carried out in [22]. The authors adopted a partition dataset using medoids and improved the result gap statistics. While some of these studies have discussed about outliers, but they have not explicitly addressed to reduce the

outlier's ratio. A comparison of different types of spatiotemporal clustering is presented in Table 1. It highlights recent works in spatio-temporal clustering analysis with its applications and limitations of methods. However, these conventional approaches have limitations. It is a less studied topic and the impact of temporal resolution influencing the spatial events has not been thoroughly studied.

Table 1. Recent work on spatio-temporal clustering technique

Reference	Method category	Method/model name	Model validation	Application	Limitations in the article/method
[21]	Spatial regression and space-time scan statistics	Geographical weighted regression	AIC, R^2 , Log likelihood	COVID-19 cluster analysis	Outliers can have disproportionate impact on model prediction.
[22]	spatial auto correlation	K-medoid, Spectral density matrix	Gap Statistics	COVID-19 cluster analysis	Computational complexity is higher.
[23]	agglomerative hierarchical clustering	-	accuracy - 92-96%, recall and precision	Drought Analysis	Uses NLP bag of words to capture the location which is imprecise.
[24]	agglomerative hierarchical clustering	Ward clustering	system minimum variance	Distribution of social enterprises across provinces in China	Single factor detection analysis, multiple features not included.
[25]	spatio-temporal Clustering	Kulldorff's space-time scan statistic, discrete Poisson model	log likelihood ratio test	Hotspot detection of COVID-19 cases in Johor, Malaysia.	Factors triggering cluster formation were unclear.
[26]	Deep learning	MuSTC, spatial correlation	MAE - 0.2304, RMSE- 0.3527	Sea Surface Temperature prediction	Additionally, Regional information required.
[27]	Statistical	The Gertis Ord Gi* for hotspot analysis	p-value<0.05, z-score >1.96 with confidence level 90,95,99%	disease control-Bovine anaplasmosis across Zimbabwe	Experiments conducted on data of two years. Shorter period.
[28]	Hierarchical clustering	Average Linkage criterion, partitioning around medoids, Smith and Schlather model, Hopkins statistic	Gap Statistics and Silhouette methods, p-value>0.05, Hopkins statistic>0.5,	Drought analysis of Lowveld in the Limpopo Province in South Africa.	Study doesnot include full range of dependence of parameters. Relied on partial extremal dependence of parameters.
[29]	Space-time scan statistics	Binomial regression model-Poisson model.	p-value<0.05, monte-carlo likelihood	Clustered attacks of leopards on humans in Himachal Pradesh, India 2004-2015	Cannot detect irregular shape clusters
[30]	Density based and regression model	DBSCAN, multiscale geographically weighted regression (MGWR)	Akaike information criterion (AIC)	spatial agglomeration of the catering industry	spatial relationship between the service industry and residential is explored and not explored other industry.
[31]	Hierarchical clustering	Self-organized maps (SOM)	Adjusted randomized index (ARI), Silhouette score, Calinski-Harabasz score	-	The SOM architecture requires modification for different dataset. Experiments carried on standard dataset from UCI repository.
[32]	Space-time scan statistics	Kulldorff's space-time scan statistic, discrete Poisson model	Loglikelihood ratio (LLR), Relative Risk (RR), $p<0.05$, LLR= 886,097.7, RR= 5.55, $P<0.05$	spatiotemporal clusters of malaria incidences	Study is confined to a specific region dataset. Cannot detect irregular shape clusters
[33]	Spatio-temporal scan statistics	discrete Poisson model, window scan	LLR, RR, Monte-Carlo statistical significance.	cluster of human brucellosis	Study is confined to a specific region dataset. Cannot detect irregular shape clusters

3. METHODOLOGY

This section discusses the proposed methodology using hybrid grid and density-based clustering approach on spatio-temporal data as shown in the Figure 1. The proposed method is implemented using Python programming in Colab environment which offers Google free cloud space for storage of data along with CPU processing capability. We have employed the concept of grid structure, density of the grid cells and neighborhood of instances, centroids and grids to form clusters. Further use density attraction rate of neighboring clusters to merge the clusters and derive final clusters.

The dataset is obtained from the <https://seismo.gov.in> [34] which is the Government of India portal for Seismic events containing earthquake spatiotemporal data for the Indian subcontinent. The data is available in CSV file. With 6506 samples have been employed in our experiment from the year August 2019 to January 2024 as mentioned in Table 2. The attributes of the dataset are spatial longitude, latitude, timestamp, and depth of the event along with comments. Table 3, shows the types of earthquake severity levels based on magnitude

and depth. We imported the earthquake catalogue and deleted the comments describing the textual location of the earthquake. Table 4, describes the parameters of the proposed algorithm with its initialization. The visualisation of different levels of earthquakes recorded are shown in Figures 2(a) and 2(b).

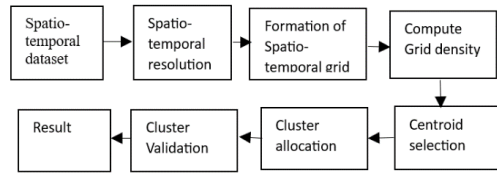


Figure 1. The proposed spatio-temporal clustering framework

Table 2. Spatio-temporal dataset and instances

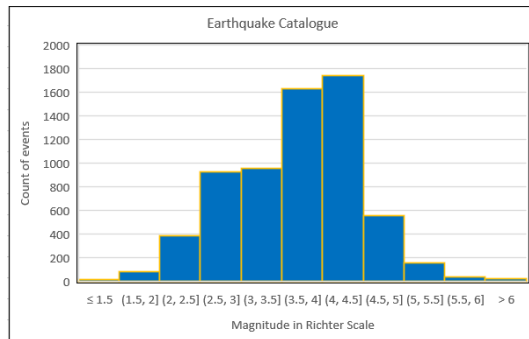
Dataset	Total instances
Indian Earthquake Catalogue August 2019 to January 2024	6506

Table 3. Descriptive statistics of earthquake dataset

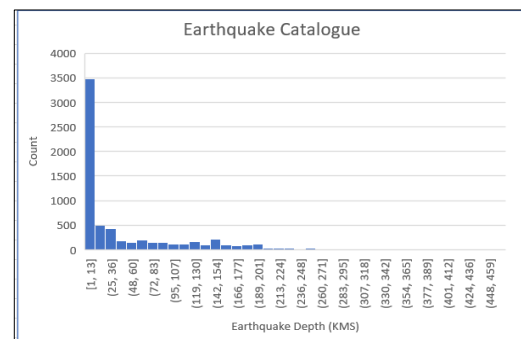
Attribute	Min	Max	Mean	Median
Magnitude	1.2	7	4.4	3.9
Depth	0.8	471	48	10

Table 4. Parameters of the algorithm with initialization

Parameter	Description	Initialization and range
λ	Total equally spaced cells	10,20
γ	Minimum probable centroids in a grid cell	2
r	Increment rate	1 to λ
\min_{Density}	Minimum density of a cluster	10
\min_{TH}	Minimum Threshold	0.1



(a)



(b)

Figures 2. Summary of different w.r.t count of earthquake dataset based on (a) magnitude and (b) depth levels

Spatiotemporal data is a sequence of data points in increasing order of time and is expressed as (1).

$$ST = \{st_1, st_2, \dots, st_n\} \quad (1)$$

where ST is a collection of spatio-temporal events dataset and 'n' is the total number of spatiotemporal events present in the dataset. st_i represents the i^{th} datapoint that records the longitude, latitude as location coordinates along with occurrence time of the event.

Grid: G is a multidimensional logical grid that geographically and temporally divides the spatiotemporal space. The division is based on the longitude, latitude, and time. In addition, the dataset also records non-spatial information related to events.

Distance Measure: This measure describes the closeness of two data points based on their spatial-temporal distance and similarity, producing lower values for low similarity and higher values for high

similarity. As a spatio-temporal distance measure, we adopt the Haversine distance formula for spatial distance along with the temporal distance measured in days. We assume that the time advancement between any two events was at least 1. The Haversine distance formula for two spatial instances is expressed as (2):

$$Hdist(O_i, O_j) = 2R \cdot \arcsin \sqrt{\sin^2(B - A) + \cos(A) \cdot \cos(B) + \sin^2(D - C)} \quad (2)$$

where R is radius of the Earth with value as 6371 kms. A, B, C, D represents:

$$A = O_i.lat, B = O_j.lat, C = O_i.lon, D = O_j.lon \quad (3)$$

$$Tdist(O_i, O_j) = \begin{cases} to_days(O_i.time - O_j.time), & \text{if } (O_i.time \neq O_j.time) \\ 1, & \text{if } (O_i.time = O_j.time) \end{cases} \quad (4)$$

Step 1: Determine minimum and maximum longitude and latitude coordinates of the dataset.

$$\begin{aligned} G(\minlon, \minlat) &= \forall_i \min(st_i.longitude, st_i.latitude) \\ G(\maxlon, \maxlat) &= \forall_i \max(st_i.longitude, st_i.latitude) \end{aligned} \quad (5)$$

$$\begin{aligned} t_{min} &= \forall_i \min(st_i.time) \\ t_{max} &= \forall_i \max(st_i.time) \end{aligned} \quad (6)$$

Step 2: Calculate the step size.

$$\begin{aligned} \Delta lat &= \frac{(\maxlat - \minlat)}{\lambda} \\ \Delta Lon &= \frac{(\maxlon - \minlon)}{\lambda} \\ \Delta t &= \frac{(t_{max} - t_{min})}{|ST|} \end{aligned} \quad (7)$$

where λ is initialised to 20.

Step 3: Form spatio-temporal grid with grid cells.

$$\begin{aligned} Lat_1 &= \minlat + r \times \Delta Lat, r \leq \lambda \\ Lat_2 &= Lat_1 + \Delta Lat, Lat_2 \leq \maxlat \\ Lon_1 &= \minlon + r \times \Delta Lon, r \leq \lambda \\ Lon_2 &= Lon_1 + \Delta Lon, Lon_2 \leq \maxlon \end{aligned} \quad (8)$$

$$Gspatial(Lat_1, Lat_2, Lon_1, Lon_2); Gtemporal(t_j, t_j + \Delta t) \quad (9)$$

Step 4: Allocate the data-points to the grid cells.

$$\begin{aligned} G_{ijk} &= \{O_m | (G_{ijk}.minlat \leq O_m.lat \leq G_{ijk}.maxlat) \wedge (G_{ijk}.minLon \leq O_m.lon \leq G_{ijk}.maxlon) \\ &\quad \wedge O_m.time \in [k-1, k]\} \end{aligned} \quad (10)$$

where O_m represents the spatio-temporal instances, with its event time belonging to the interval k-1 to k.

Step 5: Compute the density of each grid cell as

$$Density(G_{ijk}) = |O_m|, \forall O_m \in G_{ij} \wedge O_m.time \in [k-1, k] \quad (11)$$

where G_{ijk} represents the grid cell index number.

Step 6: Determine 'p', the maximum number of probable centroids.

$$p = \lceil \log_2 (Density(G_{ijk})) + \gamma \rceil \quad (12)$$

where γ takes value as 2.

Step 7: Distribution of probable centroids (PCentre).

$$\begin{aligned} &\text{If } Density(G_{ijk}) > \min_{Density} \\ &\quad /* \text{ Select probable centroids P */} \\ PCentre(G_{ijk}) &= \{Random(O_m) | O_m \in G_{ijk}\} \end{aligned} \quad (13)$$

$$|PCentre(G_{ijk})| \leq p \quad (14)$$

where p is the maximum number of probable centroids.

Step 8: Assign each grid cell data instances to the closest probable centroids with minimum distance to form probable clusters.

$$PCluster_q(O_m) = \operatorname{argmin}_{q=1..p} \{ \operatorname{dist}(O_m, PCentre_q) \} \quad (15)$$

where dist is the spatio-temporal distance given as

$$\operatorname{dist}(O_m, PCentre_q) = H\operatorname{dist}(O_m, PCentre_q) * T\operatorname{dist}(O_m, PCentre_q) \quad (16)$$

$H\operatorname{dist}$ is the Haversine spatial distance between two locations. $T\operatorname{dist}$ is the temporal distance between the two events converted into days.

Step 9: Compute the average radius of the probable clusters.

$$\operatorname{Radius}(PCluster_q) = \frac{\sum_{O_m \in PCluster_q} \operatorname{dist}(O_m, PCentre_q)}{|PCluster_q|} \quad (17)$$

Step 10: Calculate the density attraction rate of probable centroids.

$$\operatorname{DensityAttractionRate}(PCentre_q) = \frac{|PCluster_q|}{|G_{ijk}|} \quad (18)$$

Step 11: Sort the density attraction rate of each grid cell.

The density attraction rate is used to merge the clusters with minimum points joins to strong clusters in the neighbourhood.

Step 12: For each grid cell $G_{ijk} \in G$ do

For each $PCentre_q \in G_{ijk}$ do

If $(\operatorname{DensityAttractionRate}(PCentre_q) \leq \min_{TH})$ then

$$PCentre \leftarrow PCentre - PCentre_q \quad (19)$$

If $(\operatorname{DensityAttractionRate}(PCentre_q) \geq 0.3)$ then

/* Include the cluster in final list of clusters */

$$Cluster \leftarrow PCluster_q \quad (20)$$

If $(\operatorname{DensityAttractionRate}(PCentre_q) < 0.3)$ then

Go to step 13

Step 13: Construct centroid to centroid distance matrix, $p \times p$ for p centroid configuration.

$$\operatorname{Exteriordist}(PCentre_q, PCentre_r) = \operatorname{dist}(PCentre_q, PCentre_r) - \operatorname{Radius}(PCentre_q) - \operatorname{Radius}(PCentre_r) \quad (21)$$

Step 14: Find the neighbour clusters using exterior distance.

$$\operatorname{Neighbour}(PCentre_q) = \min \operatorname{Exteriordist}(PCentre_q - PCentre_r) \quad (22)$$

Step 15: Find the neighbour cluster density attraction rate to merge the cluster.

If $(\operatorname{DensityAttractionRate}(PCentre_q) \leq \operatorname{DensityAttractionRate}(PCentre_r))$ then

$$PCluster_r \leftarrow PCluster_r \cup PCluster_q$$

$$PCluster \leftarrow PCluster - PCluster_q$$

$$PCentre \leftarrow PCentre - PCentre_q \quad (23)$$

Continue step 12

If $(\operatorname{DensityAttractionRate}(PCentre_q) > \operatorname{DensityAttractionRate}(PCentre_r))$ then

$$PCluster_q \leftarrow PCluster_q \cup PCluster_r$$

$$PCluster \leftarrow PCluster - PCluster_r$$

$$PCentre \leftarrow PCentre - PCentre_r$$

$$\text{Continue step 12} \quad (24)$$

Step 15: Find outlier ratio.

$$\operatorname{Outlier_ratio} = \frac{|U_{vq}Cluster_q|}{|ST|} \quad (25)$$

Step 16: Calculate silhouette index for quality.

$$Silhouette_index = \frac{\mu - M}{\max(\mu, M)} \quad (26)$$

where μ is the mean distance from the centroid to all other data instances within the cluster. M is the mean distance to all other clusters data instances.

Step 17: Stop.

4. RESULTS AND DISCUSSION

Figures 3 and 4 present the experimental result of proposed hybrid clustering algorithm. The results display the magnitude and depth of every cluster with respect to time in the Indian subcontinent. This reflects the density of the formed clusters in space and time dimensions. Densely populated cluster 25 is in Fayzabad, Afghanistan, Pakistan, and Jammu Kashmir regions of India with 1219 events and a mean magnitude of 4.15 richter scale. The size of cluster 25 has been the highest as reflected in Figure 4. It is observed that the recurrence duration given by the mean time between the events is 99 hours. The next highly populated cluster 7, is in the eastern India region that include Mizoram, Arunachal Pradesh, and Manipur. The mean magnitude is 3.48 richter scale. Then the next highest being the Himachal Pradesh, Uttarakhand, and certain region of Jammu Kashmir as cluster 19 with mean magnitude of 3.08 richter scale. Next is the Andaman Nicobar Islands cluster 1 with a mean magnitude of 4.45 richter scale. The data analysis also reflects that there are fewer earthquakes with higher magnitudes and many number of it with moderate magnitudes i.e. majorly the events fall in the range from 3.5 to 5. It is observed that the stronger earthquakes are followed by lesser magnitude earthquake in the near surrounding regions. The aftershocks are even felt after several days from the stronger earthquakes. The outlier ratio was 0.005% towards the 38 events not closer to any of the clusters in space-time dimension which is constrained by the distance threshold. Outliers are events that belong to regions such as Oman and Maldives, that have not been in the range of spatio-temporal grid and could not be assigned to any clusters, due to the spatial distance being greater than the distance threshold value.

The analysis has shown spatial and temporal interactions, and changing the resolution, provides an effective algorithm for earthquake modeling. If the distance and time thresholds are maximum, will result in merging many clusters in one or could result in overlap of clusters. The study confirms that magnitude and frequency are correlated in spatiotemporal dimension and tends to generate clusters. Across and within the cluster distance and variation differentiate regions with high-risk earthquake zones. Low-risk clusters appear in the region. The Andaman, Nicobar, Jammu and Kashmir regions comprise high-risk clusters. High-risk events with magnitudes above six richter scales were observed in cluster 24. Greater depth events are identified in cluster 25 that make it high-risk clusters. These are highly complex regions due to many fault lines periodically releasing the tectonic stresses in the form of earthquakes.

The model produces better result for earthquake classification model. The results of the hybrid spatio-temporal clustering are best due to the accuracy of the results is higher. The advantage that is observed is the allocation of data instances to grids reduces the burden of comparison to farthest centroids not in their neighborhood. Eliminating the unnecessary computations. Given that centroid selection is a random process, may result into increase in the mean intracluster distance. For even distribution of centroids to form clusters with strong connectivity, the distribution of probable centroids, computation of density attraction and merging of clusters locally and globally have led to better clustering quality. We used the discard policy to eliminate the clusters and merge to those neighbor clusters with higher density attraction rate has shown the proposed method is adaptive.

The proposed method in this study tended to have an inordinately higher proportion of data clusters based spatial and temporal density, which is in contrast to the result as shown in STK-means methods as shown in the Figures 5 and 6. STK-means clusters are time slice of data instances irrespective to spatial distance whereas Figure 6 shows non overlapping distinct clusters. Our study suggests that higher number of clusters formed due to the spatial dimensions of the dataset and density attraction ratio, results into reducing the outliers and increase in the clustering quality. Which is evident from the result of ST_DBSCAN and proposed method shown in the Figures 7 and 8. The proposed methods benefits from forming clusters based on centroid selection for the grid cell depending on its density. The results also demonstrates that the minthreshold parameter minTH is crucial in determining the outlier threshold. If the minthreshold is close to zero, will result in assignment of every data instance to some cluster. But even a slight increase in minthreshold, start to increase the outlier proportion. Further the other parameter that also is of importance is the connectivity linkage distance between clusters. As this distance between clusters increases, it results into more tightly coupled clusters with higher density towards the core. On the contrary when the connectivity linkage between the clusters is made small, the clusters are more arbitrary in shape, and are loosely coupled to the core centroids as the distance between centroid and the data instances on the border's increases. The silhouette index is about 0.93 reflecting good

clustering result. It is observed that on this dataset the proposed algorithm has shown better clustering quality. With Davis Bouldin DB index 2.337 as shown in the Table 5. The number of clusters as are increased, the spatio-temporal distances between the clusters are reduced, forming strong clusters.

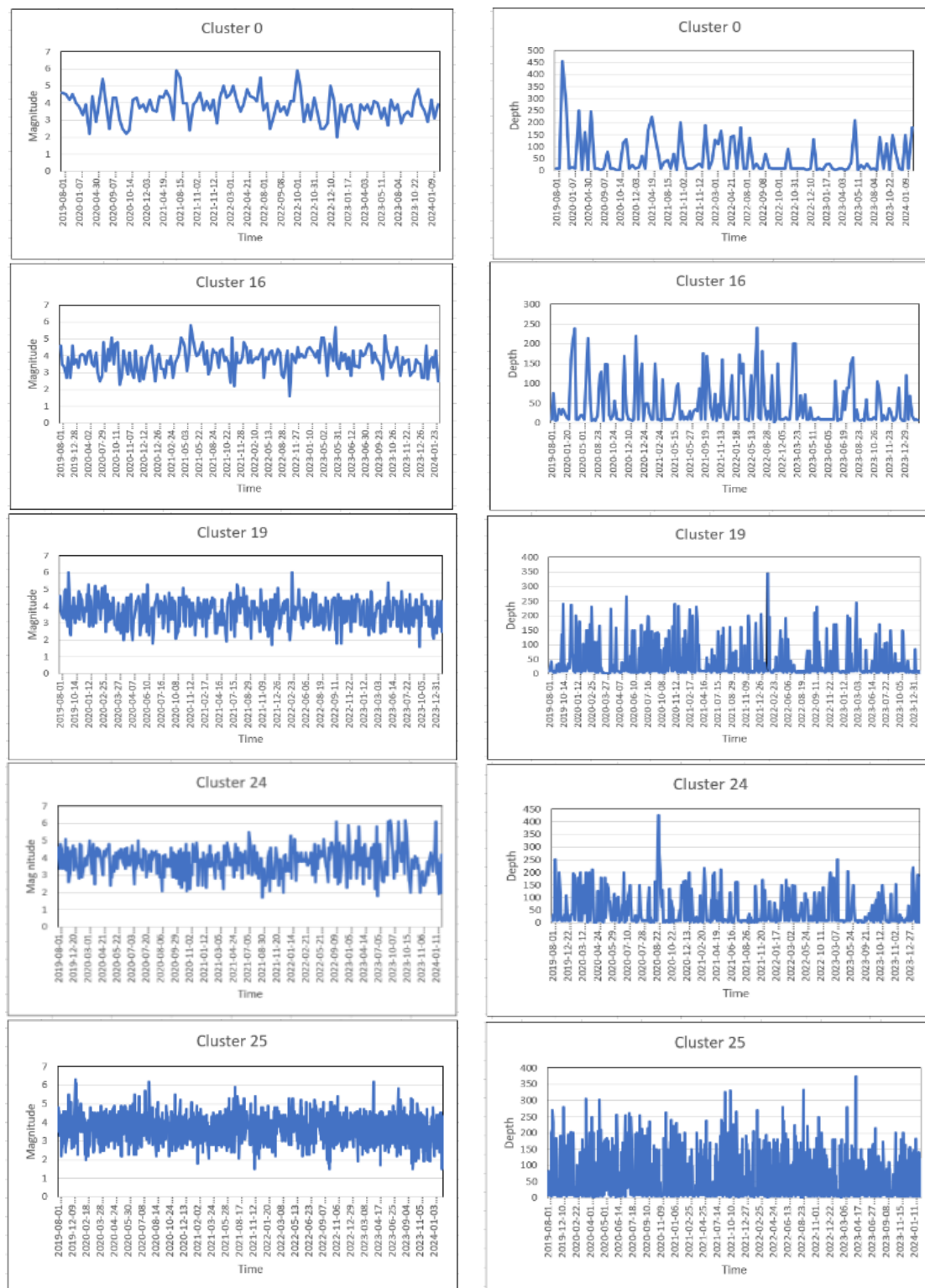


Figure 3. Result of proposed clustering algorithm on the earthquake dataset showing distinct density of the clusters with magnitude and depth with respect to time plot

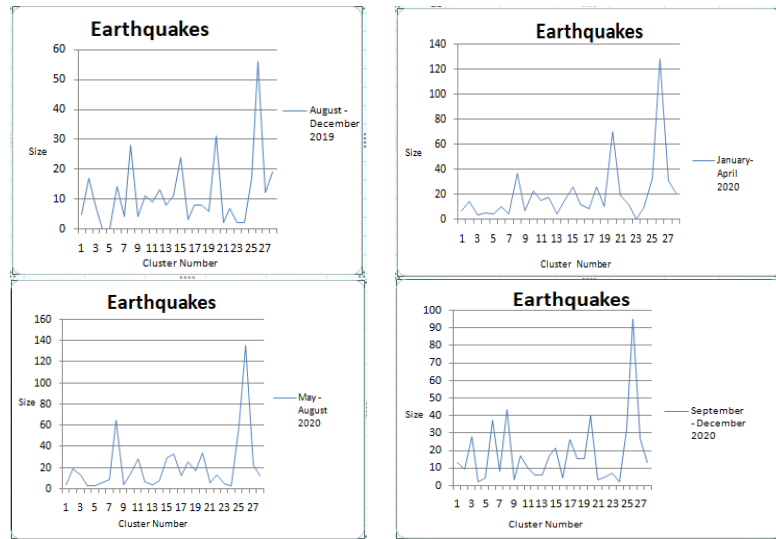


Figure 4. Result of proposed clustering algorithm on the earthquake dataset showing trend of events with clustering size

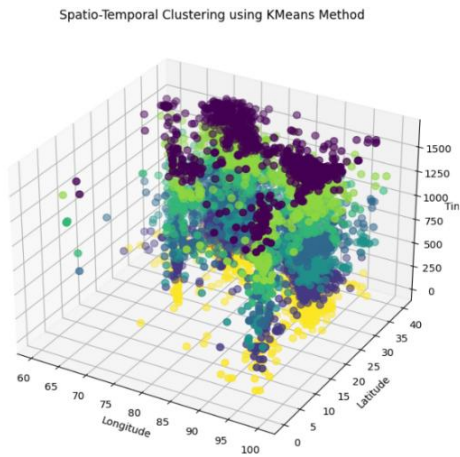


Figure 5. Result of STK-means on Indian subcontinent earthquake dataset producing seven clusters

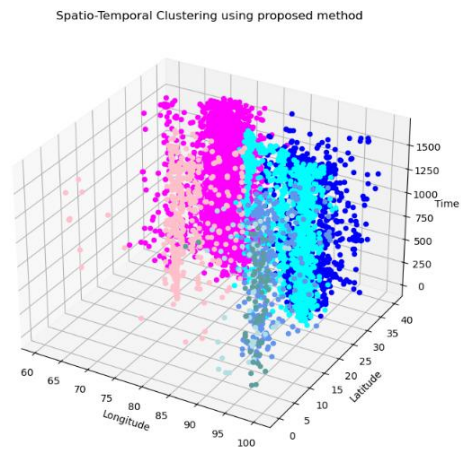


Figure 6. Result of proposed clustering algorithm on Indian earthquake dataset producing seven clusters

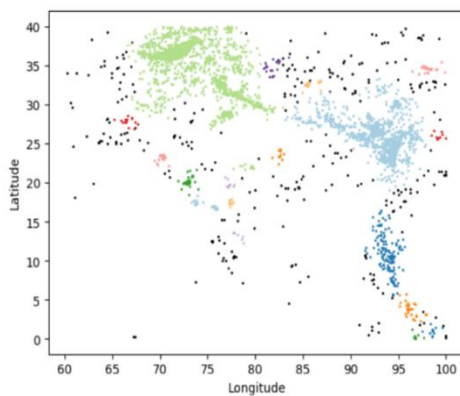


Figure 7. Result of STDBSCAN on Indian earthquake dataset with outliers shown in black color. Epsilon1=1.15, Epsilon2=500 km, min_samples=10, forming 19 clusters

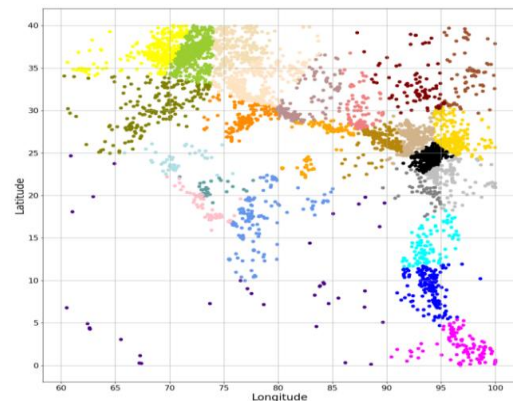


Figure 8. Proposed algorithm on Indian earthquake dataset, min_distance_threshold=500 km, min_density=10, forming 24 clusters and outliers shown in violet color

Table 5. Clustering quality index of proposed method

Evaluation Parameter	Result
Silhouette Index	0.93
Davis Bouldin Index	2.337

5. CONCLUSION

This paper proposes a novel and adaptive method of clustering. The method has been experimentally evaluated on real and standard earthquake dataset of Indian subcontinent. The clustering technique uses grid and density-based partitioning of data instances. Restricting the analysis to the effects of space and time, provides us with the information that events that are high intensity events are followed by weak events in the same clustering region reasoning to aftershocks. Our proposed method has found distinct, non-overlapping arbitrary shaped clusters on spatial and temporal data with reducing outlier ratio and distance metric computation by taking advantage of grid structure. The silhouette index is about 0.93 shows good clustering result. The proposed method for Spatio-temporal clustering is experimented on earthquake dataset but it can be applied on other Spatio-temporal dataset to study the dynamics of data. Further research direction we would take up is to minimize the parameter required for the method.




REFERENCES

- [1] S. Meshram and K. P. Wagh, "Mining intelligent spatial clustering patterns: A comparative analysis of different approaches," in *Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom 2021*, 2021, pp. 325–330, doi: 10.1109/INDIACom51348.2021.00056.
- [2] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 1–55, Oct. 2014, doi: 10.1145/2629592.
- [3] F. Hu *et al.*, "ClimateSpark: An in-memory distributed computing framework for big climate data analytics," *Computers and Geosciences*, vol. 115, pp. 154–166, 2018, doi: 10.1016/j.cageo.2018.03.011.
- [4] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 1994, vol. 398, pp. 359–370.
- [5] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proceedings 18th International Conference on Data Engineering*, pp. 673–684, doi: 10.1109/ICDE.2002.994784.
- [6] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, Jun. 2005, pp. 491–502, doi: 10.1145/1066157.1066213.
- [7] B. Guan, L. Liu, and J. Chen, "Using relative distance and hausdorff distance to mine trajectory clusters," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 1, 2013, doi: 10.11591/telkomnika.v11i1.1877.
- [8] M. M. Fréchet, "Sur quelques points du calcul fonctionnel," *Rendiconti del Circolo Matematico di Palermo*, vol. 22, no. 1, pp. 1–72, 1906, doi: 10.1007/BF03018603.
- [9] Z. Cheng, L. Jiang, D. Liu, and Z. Zheng, "Density based spatio-temporal trajectory clustering algorithm," *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2018, pp. 3358–3361, 2018, doi: 10.1109/IGARSS.2018.8517434.
- [10] J. Du and L. Aultman-Hall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 3, pp. 220–232, 2007, doi: 10.1016/j.tra.2006.05.001.
- [11] N. Pelekis, I. Kopanakis, I. Ntoutsis, G. Marketos, and Y. Theodoridis, "Mining trajectory databases via a suite of distance operators," in *2007 IEEE 23rd International Conference on Data Engineering Workshop*, Apr. 2007, pp. 575–584, doi: 10.1109/ICDEW.2007.4401043.
- [12] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data and Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007, doi: 10.1016/j.datak.2006.01.013.
- [13] M. G. Doborjeh and N. Kasabov, "Dynamic 3D clustering of spatio-temporal brain data in the NeuCube spiking neural network architecture on a case study of fMRI data," in *Neural Information Processing*, vol. 9492, 2015, pp. 191–198, doi: 10.1007/978-3-319-26561-2_23.
- [14] J.-X. Guo, T. Liu, X.-J. Qi, J. Chen, and S.-Y. Ai, "Application of spatio-temporal scanning in the analysis of spatio-temporal clusters of foodborne diseases in Zhejiang Province," *Chinese Preventive Medicine*, vol. 21, no. 11, pp. 1171–1177, 2020, doi: 10.16506/j.1009-6639.2020.11.003.
- [15] J. L. -Loiola, G. Otón, R. Ramo, and E. Chuvieco, "A spatio-temporal active-fire clustering approach for global burned area mapping at 250 m from MODIS data," *Remote Sensing of Environment*, vol. 236, Jan. 2020, doi: 10.1016/j.rse.2019.111493.
- [16] S. Gong, J. Cartledge, R. Bai, Y. Yue, Q. Li, and G. Qiu, "Extracting activity patterns from taxi trajectory data: a two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation," *International Journal of Geographical Information Science*, vol. 34, no. 6, pp. 1210–1234, 2020, doi: 10.1080/13658816.2019.1641715.
- [17] Y. Yang, J. Cai, H. Yang, J. Zhang, and X. Zhao, "TAD: A trajectory clustering algorithm based on spatial-temporal density analysis," *Expert Systems with Applications*, vol. 139, Jan. 2020, doi: 10.1016/j.eswa.2019.112846.
- [18] A. I. J. Tostes, F. D. L. P. Duarte-Figueiredo, R. Assunção, J. Salles, and A. A. F. Loureiro, "From data to knowledge," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, Aug. 2013, pp. 1–8, doi: 10.1145/2505821.2505831.
- [19] A. Muñoz-Villamizar, E. L. Solano-Charris, M. AzadDisfany, and L. Reyes-Rubiano, "Study of urban-traffic congestion based on Google Maps API: the case of Boston," *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 211–216, 2021, doi: 10.1016/j.ifacol.2021.08.079.
- [20] G. Georgoulas, A. Konstantaras, E. Katsifarakis, C. D. Stylios, E. Maravelakis, and G. J. Vachtsevanos, "'Seismic-mass' density-based algorithm for spatio-temporal clustering," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4183–4189, 2013, doi: 10.1016/j.eswa.2013.01.028.
- [21] N. Nazia, J. Law, and Z. A. Butt, "Spatiotemporal clusters and the socioeconomic determinants of COVID-19 in Toronto




- neighbourhoods, Canada,” *Spatial and Spatio-temporal Epidemiology*, vol. 43, Nov. 2022, doi: 10.1016/j.sste.2022.100534.
- [22] S. Deb and S. Karmakar, “A novel spatio-temporal clustering algorithm with applications on COVID-19 data from the United States,” *Computational Statistics and Data Analysis*, vol. 188, 2023, doi: 10.1016/j.csda.2023.107810.
- [23] J. Sodge, C. Kuhlicke, and M. M. de Brito, “Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4178096.
- [24] X. Liu and D. Lv, “Spatial and temporal characteristics, spatial clustering and governance strategies for regional development of social enterprises in China,” *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26246.
- [25] F. Y. Foo, N. A. Rahman, F. Z. S. Abdullah, and N. S. A. Naeem, “Spatio-temporal clustering analysis of COVID-19 cases in Johor,” *Infectious Disease Modelling*, vol. 9, no. 2, pp. 387–396, 2024, doi: 10.1016/j.idm.2024.01.009.
- [26] H. Peng, W. Li, C. Jin, H. Yang, and J. Guan, “MuSTC: A multi-stage spatio-temporal clustering method for uncovering the regionality of global SST,” *Atmosphere*, vol. 14, no. 9, 2023, doi: 10.3390/atmos14091358.
- [27] N. Madyavanhu, M. D. Shekede, S. Kusangaya, D. M. Pfukenyi, S. Chikerema, and I. Gwitira, “Bovine anaplasmosis in Zimbabwe: spatio-temporal distribution and environmental drivers,” *Veterinary Quarterly*, vol. 44, no. 1, pp. 1–16, 2024, doi: 10.1080/01652176.2024.2306210.
- [28] M. M. Nemukula, C. Sigauke, H. Chikoore, and A. Bere, “Modelling drought risk using bivariate spatial extremes: application to the Limpopo Lowveld Region of South Africa,” *Climate*, vol. 11, no. 2, 2023, doi: 10.3390/cli11020046.
- [29] S. Shivakumar *et al.*, “Examining leopard attacks: spatio-temporal clustering of human injuries and deaths in Western Himalayas, India,” *Frontiers in Conservation Science*, vol. 4, 2023, doi: 10.3389/fcsc.2023.1157067.
- [30] R. Tang, G. Hou, and R. Du, “Isolated or colocated? exploring the spatio-temporal evolution pattern and influencing factors of the attractiveness of residential areas to restaurants in the central Urban Area,” *ISPRS International Journal of Geo-Information*, vol. 12, no. 5, May 2023, doi: 10.3390/ijgi12050202.
- [31] K. Tripathi, “The novel hierarchical clustering approach using self-organizing map with optimum dimension selection,” *Health Care Science*, vol. 3, no. 2, pp. 88–100, 2024, doi: 10.1002/hcs2.90.
- [32] T. Z. Nigussie, T. T. Zewotir, and E. K. Muluneh, “Detection of temporal, spatial and spatiotemporal clustering of malaria incidence in northwest Ethiopia, 2012–2020,” *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-07713-3.
- [33] X. Yu *et al.*, “Epidemiological characteristics and spatio-temporal analysis of brucellosis in Shandong Province, 2015–2021,” *BMC Infectious Diseases*, vol. 23, no. 1, 2023, doi: 10.1186/s12879-023-08503-6.
- [34] National Center for Seismology, “Seismological data: earthquake catalogue, Aug. 2019 to Jan. 2024,” *National Center for Seismology, Ministry of Earth Sciences, Government of India*. [Online]. Available: <https://riseq.seismo.gov.in/riseq/earthquake/archive>

BIOGRAPHIES OF AUTHORS



Ms. Swati Meshram    is a Research Scholar in Computer Science and Engineering, has Master of Engineering degree in computer engineering and is working as an Assistant Professor, P.G. Department of Computer Science, SNDT Women’s University, Maharashtra, India. Her areas of interest are data mining, machine learning, and databases. She can be contacted at email: swati.meshram@computersc.sndt.ac.in.



Dr. Kishor P. Wagh    is a Research Supervisor and working as an Assistant Professor at Department of Information Technology, Government College of Engineering, Amravati. He has more than 20 years of experience in teaching and research in the field of computer science and engineering and information technology. His area of interest is in data mining, machine learning, databases, and object oriented methodology. He can be contacted at email: kishorpwagh2000@gmail.com.