International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

A Novel and Efficient Spatio-Temporal Clustering Algorithm to Discover Patterns in Indian Earthquake Data

[1]Swati Meshram, [2]Dr.Kishor P.Wagh
Department of Computer Science and Engineering,
Government College of Engineering.
Amravati, Maharashtra,India,
[1]swati.meshram@computersc.sndt.ac.in
[2]kishorpwagh2000@gmail.com

Abstract— The seismic activity in India, including the Himalayas, the North-East region, and the Andaman-Nicobar Islands, is prominently displayed on the seismic map. It is important to analyze the specific characteristics of seismic events on the Indian sub-continent. This research paper introduces a new algorithm for data analysis using a partitioning technique that mines clustering patterns. The algorithm generates clusters of spatial and spatio-temporal data by distributing the data into spatial bins or buckets, identifying neighboring bins, and minimizing distance calculations. Furthermore, the selection of centroids is based on the density of data in the spatial region, utilizing a random selection method. It requires minimal parameter settings, enhancing its efficiency and practicality for analyzing large-scale spatio-temporal datasets. We validate the algorithm's efficacy through experiments conducted on the Indian earthquake spatio-temporal dataset, demonstrating its ability to detect spatio-temporal patterns and identify earthquake-prone regions.The conducted experiments reveal a correlation between the frequency of earthquake events and the number of clusters formed, indicating that regions with a higher occurrence of earthquakes exhibit a greater clustering tendency, signifying their susceptibility to seismic activity. The results imply promising clustering quality, with Silhouette index in the range of 0.88 to 0.93.

Keywords- Pattern recognition; pattern mining; centroid; neighbourhood; seismology; spatial bins.

I. INTRODUCTION

The progress made in machine learning has introduced fresh possibilities and challenges in the realm of spatio-temporal computing. Machine learning models are essential in addressing spatio-temporal issues, where "spatial" refers to geographical coordinates and "temporal" pertains to the timing of events or data recording. Spatio-temporal data consists of recorded events with both spatio-temporal attributes and non-spatio-temporal features. Massive spatio-temporal data is constantly generated from diverse sources like satellites, GPS, lidar, drones, and sensors. The abundance and complexity of spatio-temporal data pose challenges in extracting meaningful insights and patterns. Machine learning techniques offer a solution to process and analyze spatio-temporal data, enabling pattern identification, prediction, and forecasting.

Classification techniques, such as clustering, are well-suited for pattern mining in spatio-temporal data analysis. Clustering, an unsupervised classification method, groups data points that exhibit high similarity [1], [2]. The clustering process involves reading the input dataset and defining the distance metric. A clustering algorithm is then applied to process the data and generate clusters, which are subsequently validated to obtain optimal results. Cluster validation entails assessing the quality of clusters based on internal information about the objects being clustered and external comparisons with other results or by varying different parameters.

Distance or similarity measures are utilized to determine the similarity in formation of clusters [3]. Common distance metrics include Euclidean, Manhattan, Minkowski, and Cosine [4]. Clustering has a wide range of applications in pattern recognition, such as market research and recommendation systems, where it involves grouping similar customer reviews to evaluate product popularity. It is also employed in tasks like spam email classification, dynamic trend detection, image processing, and identifying areas with similar land use. Cluster analysis is valuable for identifying outliers, which are objects that do not conform to any group. Outlier detection finds practical application in areas such as detecting online credit card and insurance fraud, as well as understanding spending patterns among extremely high or low-income groups.

Spatio-temporal data clustering focuses on revealing meaningful correlations among spatial and temporal features [5]. This type of clustering analysis finds utility in diverse applications, includingcrime hotspot detection,climate modeling, event modeling and agricultural monitoring.

Earthquakes occur when energy is suddenly released from the Earth's crust, resulting in ground vibrations [6]. The Indian sub-continent exhibits a wide range of seismic characteristics. Analyzing earthquake data is vital for comprehending and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

assessing seismic hazards.A few research focusing on earthquake data analysis utilizing machine learning techniques has been carried out, there is still a lack of a straightforward clustering-based approach in this domain. Therefore, this study aims to fill this gap and provide decision-makers with valuable insights into identifying highly seismic vulnerable regions in the Indian subcontinent.

II. LITERATURE REVIEW

In spatio-temporal clustering problem, the frequently employed and recent algorithms are ST-DBSCAN (Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise) [7]. This algorithm extends the traditional DBSCAN algorithm to handle spatio-temporal data. It clusters data points based on their spatio-temporal density.It is highly sensitive to the determination of appropriate parameters. ST-DBSCAN requires the setting of two parameters: the spatial density threshold (Eps) and the temporal density threshold (Eps_t). Choosing optimal values for these parameters can be challenging, as they directly impact the clustering results. Setting them too low may result in over clustering, while setting them too high may lead to under clustering or merging of distinct clusters. STDBSCAN struggles to handle datasets with complex density distributions.

ST-KMeans (Spatio-Temporal K-Means): This algorithm applies the K-Means clustering approach to spatio-temporal data [8]. It iteratively assigns data points to clusters based on their distance to cluster centroids, aiming to minimize the intra-cluster variance. Spatio-Temporal K-Means clustering is sensitive to the initialselection of cluster centroids. The K-Means algorithm requires the initial assignment of centroids, which can significantly impact the resulting clusters. In spatio-temporal data, where the distribution and density of data points can vary across both spatial and temporal dimensions, the initial placement of centroids may not accurately represent the underlying patterns and structures in the data. Consequently, this can lead to suboptimal clustering results.The earthquake data from the Bengkulu Province, Indonesia dataset has also been evaluated by K-means clustering in[9].

The paper [10] proposes trajectory clustering that aims to identify target behavioral patterns by utilizing the k-nearest spatial-temporal Hausdorff distance (STHD). The clustering technique leverages the STHD measure to determine the similarity between trajectories in both spatial and temporal dimensions, enabling the identification of common patterns in the data.Selecting an appropriate value for k is crucial, as it can significantly impact the clustering results.

A new method incorporates concepts such as the NMAST (Neighbourhood Move Ability and Stay Time) density function and NT (Noise Tolerance) metrics to assess the spatio-temporal density of data for clustering purposes [11]. The NMAST density function measures the motion characteristic of trajectories. NMAST requires setting parameters related to the neighborhood size and the duration for considering trajectory points as "stay" points. The accuracy and effectiveness of the clustering results can be influenced by the specific values chosen for these parameters.

Hidden Markov Models (HMM) are employed to cluster multivariate time series data by utilizing prior knowledge of the initial classes [12].The effectiveness heavily depends on the availability and accuracy of this initial class information. The need for initial class information can be a limitation in scenarios where such information is not readily available or difficult to obtain. An instance of such multivariate time series data is climate data collected from sensors for climate informatics. The identification of spatial and temporal patterns of association is performed to establish clusters in [13].

ST-Grid [14] utilizes grid cells to partition the spatio-temporal dimension. The grid cell size and configuration can significantly impact the accuracy and granularity of the clustering results. In [15], a density cube-based spatio-temporal clustering approach is introduced, which incorporates a distance threshold and a density compensation calculation method. The performance of the algorithm is highly sensitive to the distance threshold parameter. IMSTAGRID focuses on data partitioning and interval expansion. A study is conducted on earthquake time-series analysis, specifically de-clustering sequences and regular background events that follow a Poisson process in the time domain. This research relies on COV(T) (coefficient of variation of inter-event times) and inter-event time statistics using a sliding temporal window method [16].

To address the aforementioned challenges, this study introduces innovative approaches to tackle the issues at hand. The primary contributions of this research can be outlined as follows:

1. Applying clustering, an unsupervised classification technique, to an Indian earthquake dataset for the purpose of identifying regions that are at risk.
2. Our clustering approach involves dividing the dataset into spatial bins based on the boundaries defined by the spatial coordinates.
3. We employa conception of random selection of centroidsbased upon spatial density of the bin and in the proximity to denser areas.
4. The approach necessitates the specification of a sole parameter, minPts, and entails low computational complexity.

The rest of the paper is organized as follows: Section III introduces the proposed method, Section IV details the conducted experiments with subsequent discussions, and finally, Section V presents the conclusions.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

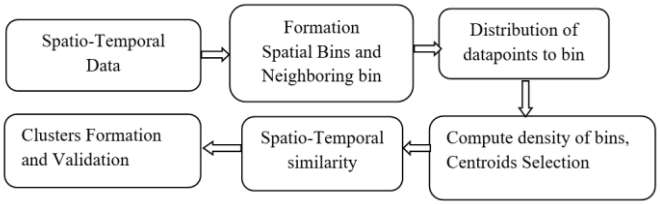

Figure 1: Process flow of the proposed clustering algorithm.

III. PROPOSED SPATIO-TEMPORAL CLUSTERING ALGORITHM

The objective of performing spatio-temporal data clustering on Indian earthquake data is to uncover meaningful patterns, relationships, and clusters within the data. By applying clustering algorithms specifically designed for spatio-temporal data, the objective is to identify regions or areas that exhibit similar seismic characteristics or are prone to earthquakes. This clustering analysis can provide valuable insights for understanding the distribution of earthquakes in the Indian subcontinent.

DATASET

The Earthquake data catalogue includes the earthquake events captured by the National Centre of Seismology, Ministry of Earth Sciences, Government of India for the period 2019 to 2022. Events are captured with the attributes as timestamp, geographical coordinates as latitude and longitude, depth, and magnitude of events. The Earthquake events occurred in the region of 0^0 to 40^0 N and 60^0 to 100^0N coordinates are present in it.Earthquakes catalogue is a good example of spatio-temporal data as it includes geographical location and occurrence time of events.

The spatiotemporal data collectively is expressed as a vector

$$D = [d_1, d_2, d_3, .., d_r] \quad (1)$$

where D represents the spatiotemporal earthquake dataset,$d_1$is a temporal attribute. The attributes d_2,d_3 are spatial attributes, along with non-spatiotemporal attributes d_r that pertains to earthquake events. The dataset includes n total number of earthquake events and 'i' and 'j' represents the subscript used to indicate the 'i^{th}' and 'j^{th}' earthquake event or bin.

Firstly, the study seeks to ascertain the total number of records available in the earthquake catalogue.

Step 1: Read the total number of records present in the earthquake catalogue as 'n'.

Step 2: Determine minimum (or least) and maximum (or top) latitudes and longitudes of events.

These extreme latitude and longitude values are utilized to establish the geographical coordinate range.

Step 3: Compute total number of spatial coordinates bins and the coordinates of each bin.

Subsequently, this range is divided into an equal number of bins (B). Furthermore, the minimum and maximum latitude and longitude values for each bin are computed.

A logical representation of a bin is expressed as

$$[B_i.LeastLat=Lt1] \wedge [B_i.TopLat=Lt2] \wedge [B_i.LeastLong=L1] \wedge [B_i.TopLong=L2] \quad (2)$$

and

$$Lt1 < Lt2, L1 < L2 \quad (3)$$

where Lt1, Lt2, LeastLat, TopLat are latitudes and L1, L2, LeastLong, TopLong are Longitudes.

Step 4: Discover the neighbourhood bins.

Determine adjacent bins in a list, for each bin B_i by identifying neighboring bins, denoted as Neighbours(B_i.). Neighbour bins are cells that surround current cell in all directions like North, South, West, East, North-West, North-East, South-West and South-East.

Step 5: Perform allocation of earthquake events from the data catalogue to the spatial coordinates'bins.

Assign each datapoint D_j, to the bin B_i as

$$Bin(D_j) = \{i \mid (B_i.LeastLat < D_j.Lat \leq B_i.TopLat) \wedge (B_i.LeastLong < D_j.Long \leq B_i.TopLong)\} \quad (4)$$

This assignment results in distribution of events from the catalogueto the bins based on the geography of occurrence and putting together the events that have occurred in the proximity.

Step 6: Determine density of each bin by counting datapoints.

$$n_{Bi} = \forall_j count(D_j) \text{ such that } 1 \leq j \leq l \text{ and } Bin(D_j)=i \quad (5)$$

Density of every bin infers the vulnerability percentage or degree to which the area is prone to seismic activity.

Step 7: Compute an average or meanof data points for the entire study area.

$$mean = \frac{1}{b}\sum_{i=1}^{b} n_{Bi} \quad (6)$$

where b is the total number of buckets.

Step 8: Classify bins as discard bins, bins less than average, bins greater than average datapoints.

$$Bi \in \begin{cases} greater\_than\_avg, & if\ n_{Bi} \geq mean \\ Less\_than\_avg, & if\ \min\_pts < n_{Bi} < mean \\ discard\_bins, & if\ 0 \leq n_{Bi} \leq \min\_pts \end{cases} \quad (7)$$

Where greater_than_avg signifies list of bins having the count of datapoints greater than mean. The less_than_avg is a list of binshaving datapoints count in the range of minimum points and mean. Discard bin has datapoints count in the range from zero to minPts. This classifies the bins in the groups as per their density.

Step 9: Perform sorting of the bin list greater_than_avg in descending order.

Step 10: Find the highest density count of bins which indicates the greatest number of earthquake events.

$$Bin_{max} = max(n_{Bi}) \quad (8)$$

Step 11: Compute δ

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

$\delta =$ surfaceDistance$(B_i.LeastLat, B_i.LeastLong, B_i.TopLat, B_i.TopLong)/2$ (9)

Step 12:

Calculate the centroids count for each bin as per the datapoints density.

$$count\_centroid(B_i) = \begin{cases} \lceil log_{10}(n_{Bi}) + 2 \rceil, if\ n_{Bi} \geq Bin_{max} \\ \lceil log_{10}(n_{Bi}) * 3/4 \rceil,\ if\ 3 * mean\ < n_{Bi} < Bin_{max} \\ 1,\ if\ mean\ \leq n_{Bi} \leq 2 * mean \\ 0,\ if\ 0 \leq n_{Bi} < mean \end{cases} \quad (10)$$

Centroid is a point of importance which is used to determine proximity and formation of clusters.

Step 13: Select random centroids for bins based on the respective count_centroid (B_i) and should be atleast δ distance from other centroids. Append centroids to list C_k.

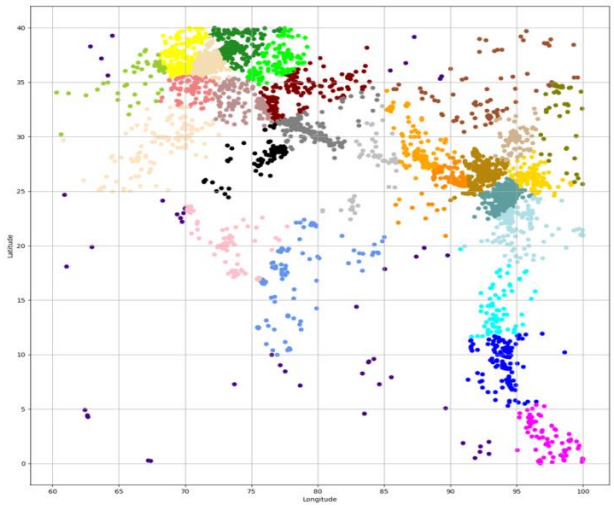

Figure 2: Execution of proposed spatio-temporal clustering with formation of 25 clusters, distance between the centroids is 300km.

If (D[j]$\in B_i$^surfaceDistance(D[j],{C_k})>δ) then

Append(D[j],{C_k}) (11)

where the surface distance is based on the Haversine formula given as

$$P = 2 * radius * arcsin\sqrt{sin^2 + \cos Lat1 . cosLat2 + sin^2(Long2 - Long1)} \quad (12)$$

and the radius of the Earth is 6371 kms.

Step 14: For each datapoint 'i', find the bin it belongs to and the closet centroid based on the current bin and neighboringbins.

Assign datapoints to its closet centroid based on the Euclidean distance formula and form clusters.

if(Bin[i]=centroidBin[j]) or (NeighbourBinCentroid(Bin[i],neighboursofcentroid[j]) is true then 14.1. ComputeEuclideanDistance(D[i], Centroid[j])

14.2. AppendD[i] to the closet centroid cluster

where the Euclidean Distance between two events is computed as

$$Euclidean\ Distance = \sqrt{\Delta S^2 + \alpha \Delta t^2 + \Delta NS^2} \quad (13)$$

with ΔS as spatial attributes difference,ΔNS as non-spatial attributes difference, Δt is the temporal difference.The α parameter is weight assigned to time attribute. The timestamp is converted into days to compute the temporal difference.

Step 15: Validate the clusters using silhouette index.

Step 16: If no more datapoints to assign, then terminate.

Figure 1, outlines the process flow of the proposed spatio-temporal clustering algorithm.

IV. RESULT AND DISCUSSION

Assignment of data points and centroids to bins assists in classification and brings spatio temporal datapoints closer to their geographical bins.Bin allocation is based on geographical coordinates as earthquake events are rare events and there might be absence of seismic activity for several intervals together.

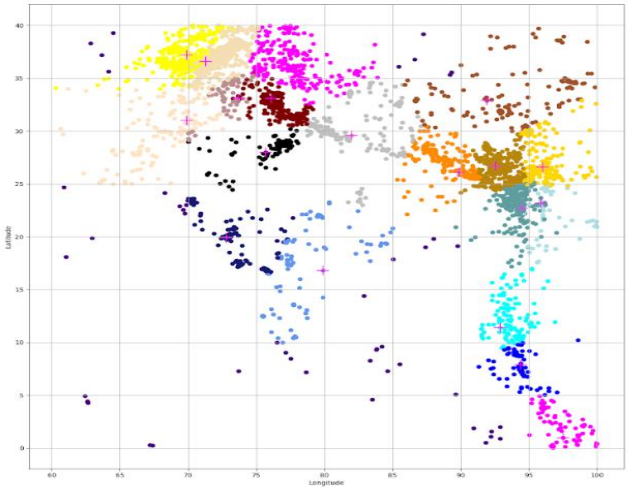

Figure 3: Execution of proposed spatio-temporal clustering with formation of 19 clusters, distance between the centroids is 500km

intervals. Resulting in a few numbers of events to be clustered for several intervals. Hence, distribution of bins is based on the spatial dimensions.

The neighbouring bins determination eliminates the need to perform comparison and Euclidean distance computation with datapoints or centroids that are situated far and are not in vicinity. The only requirement imposed is to perform the distance computation and comparison with the nearby centroids and datapoints. In other words, the location of the datapoints instances which are farther away, will not be considered for Euclidean distance computation, as datapoints

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

instances beyond a certain distance are not contributing to the same clusters. Thus, it minimizes the comparison and Euclideandistancecomputation.

The optimal selection of centroids is of prime consideration for working of the algorithm. Appropriately choosing earthquakes with larger magnitudes as centroids is the method chosen here. Further, constraint on 'centroids' imposes the requirement of selecting more centroids from the region with higher density of earthquakes. Whereas choosing either no centroids from abin if the area is less prone to earthquakes. In addition, allowing to choose few centroids if the bin density is in range above minPts to less than average earthquake belonging to bins. Bins with less than minPts are considered as discard bins where no processing is required. Datapoints get attached to the closest centroids in the neighbouring bin which is less than the maximum distance between two centroids. Else such points are regarded as outliers. The rare patterns in this case are outliers.Resultingin minimizing the error rate of clustering assignment ofdatapoints.

To incorporate the constraint of keeping clusters distant apart, once a centroid is selected from a bin or a region, all the other centroids must be atleast δ km apart. This condition assists in maintaining the mean distance between clusters.Forming easily separable and non-overlapping clusters.This is done by employing the Haversine distance formula with the Earth's radius as 6371 kms and spatio-temporal Euclidean distance.

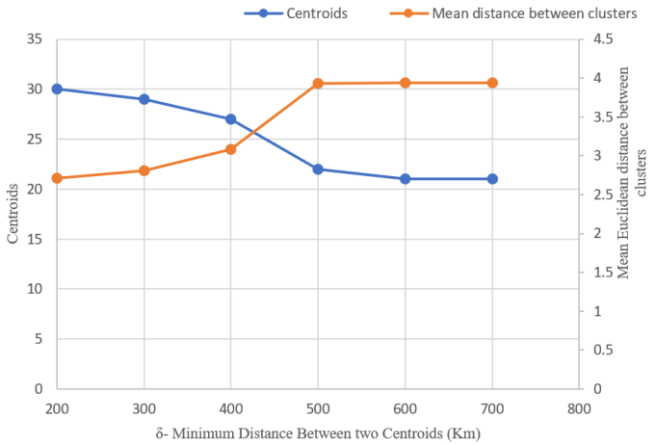

Figure 4: Effect of change in maximum distance between two centroids on number of centroids selection and mean distance between the clusters.Size of the catalogue is 5444 records.

The timestamps in the dataset are extracted to retrieve time in seconds, which introduces a large variation in Euclidean distance calculation. In order to the balance this large variation in values of timestamp, they are converted from seconds to days and used in further computation of the algorithm.

We have adopted weighted lambda to further normalize the data and bring the smoothing effect in clustering results.Shape of formed clusters is arbitrary. Each cluster is assigned a different color. The outcome, illustrated in Figure 2, demonstrates that when the minimum distance between centroids is less than 300km, more clusters are created, specifically 25 clusters.Conversely, when the minimum centroid-centroid distance is increased, the number of clusters formed decreases, as shown in Figure 3, with only 19 clusters. Generally, the density of the clusters is higher than the average density of data points in the bins. The ratio of outliers is very small, representing points where only a few earthquakes have occurred. Increasing the minimum number of data points required for a cluster (minpts) would result in more outliersbeing identified. As the number of cluster centroidsincreases,computational complexity also increases. If the cluster centroids are not appropriately selected, it may lead to the formation of cluster centroids that are very close to each other, resulting in non-separable and overlapping clusters being generated.

Measuring the clustering quality with Silhouette index given as

$$Silhoutteindex_i = \frac{y(i) - z(i)}{\max\big(z(i), y(i)\big)} \qquad (14)$$

where

z is within cluster, mean distance from datapoint 'i' to all other datapoints.

y is the mean distance to all datapoints of any other cluster.

The experiment results in the Silhouette index in the range of 0.88 to 0.93, which reflects denser clusters with good clustering quality on the Indian Earthquake dataset.

Figures 4 and 5 show the effect of change in the δ distance between centroidson the number of centroids selected and the effect on mean distance between the clusters. As δ increases, the number of centroids selected decreases and the mean distance between the clusters and within the clusters also

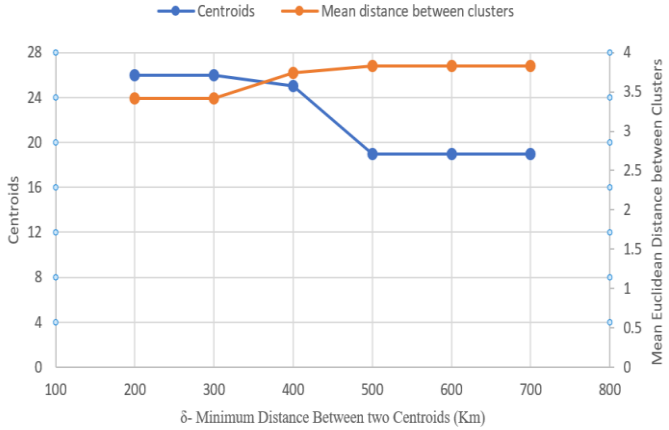

Figure 5: Effect of change in maximum distance between two centroids on number of centroids selection and mean distance between the clusters.Size of the catalogue is 4630 records

increases. The results indicate the clusters are non-overlapping and distinct with arbitrary shape.

The proposed algorithm is an improvement over existing algorithms based on distance computation. First it avoids

expensive distance calculation of a datapoint with all the centroids by using distance computation with the nearby centroids to find clustering patterns. Secondly, the neighbourhood bins limits the scope of searching as well as explores the proximity in all directions.Thirdly, bin-based approach does not limit the formation of clusters to merely take the shape of bin. Nor the bin size has impact on the clusters, as the closet centroid may be situated in the neighbouring bins. Fourth, the process does not iteratively change the assignment of clusters. It focuses on performing most accurate assignment in the process. Rather, it is more focused on centroid selection. Fifth, the process is not sensitive to the initialization of many parameters. It requires initialization of only one parameter, which is minPts. Furthermore, no prior knowledge or background information of clusters, centroids or neighbourhood is utilized. So it is independent of prior information usage.

The effectiveness of the algorithm is evident and supported by the experiment because it partitions the dataset into spatial temporal space bins and the focus is on few bins. Thus, avoiding repeated searching and computation of Euclidean distance. In turn, assuring a small search space and memory requirement.

The results demonstrate assignment of datapoints to its closest clusters. For the clustering pattern, where the quality is concerned results in no overlap of datapoint instances being assigned to two clusters and no overcounting of object instances being done.

Characteristics of earthquake events clusters in Indian subcontinent:

In Table 1, the study area has identified the highly active zones. Among these zones, clusters 0, 1, and 2 are located in the Andaman and Nicobar Islands, with cluster sizes of 76, 123, and 74 events, respectively. Cluster 3 encompasses a longitude range of 70.03° to 77.69° and a latitude range of 16.77° to 23.63°. The selected centroid for this cluster is located 93km northwest of Mumbai, Maharashtra, India, towards the Gujarat state.Within this cluster, there have been a total of 216 earthquake events. The mean time difference between these events is calculated to be 5.29 days. Notably, 81 events occurred with a time difference of 0 days. The largest magnitude earthquake recorded in this cluster is 5.3 ML, while the average magnitude is 2.94 ML. The mean depth of these earthquakes is 8km.

Cluster 5 encompasses a longitude range of 91.79° to 96.77° and a latitude range of 21.7° to 26.7°. The chosen centroid for this cluster is located 35km southeast of Imphal, Manipur, India. Within this cluster, there have been a total of 443 earthquake events. The mean time difference between these events is calculated to be 2.34 days. Notably, the majority of events, specifically 155, occurred with a time difference of 0 days, and an additional 87 events occurred with a time difference of 1 day. In total, 346 events occurred within 0, 1, 2, or 3 days of time difference. The largest magnitude earthquake recorded in this cluster is 5.3 ML, while the average magnitude is 3.52 ML. The mean depth of these earthquakes is 42.28 km. Specifically, 23 events were recorded in the 7th month of 2020, and 22 events were recorded in the 11th month of 2021 within this cluster.

Cluster 8 encompasses a longitude range of 76.6° to 84.52° and a latitude range of 27.07° to 33.85°. The selected centroid for this cluster is located 46km northwest of Pithoragarh, Uttarakhand, India. Within this cluster, there have been a total of 224 earthquake events. The mean time difference between these events is calculated to be 5.00 days. In particular, 57 events occurred with a time difference of 0 days. The majority of events, totaling 138, occurred within 0, 1, 2, or 3 days of time difference. The largest magnitude earthquake recorded in this cluster is 6.3 ML, while the average magnitude is 3.14 ML. The mean depth of these earthquakes is 11.5 km.

Cluster 12 encompasses a longitude range of 90° to 94.42° and a latitude range of 24.2° to 28.9°. The chosen centroid for this cluster is located 29km west of Tezpur, Assam, India. Within this cluster, there have been a total of 315 earthquake events. The mean time difference between these events is calculated to be 3.52 days. In particular, 113 events occurred with a time difference of 0 days. The largest magnitude earthquake recorded in this cluster is 6.4 ML, while the average magnitude is 3.16 ML. The mean depth of these earthquakes is 18 km. Specifically, 58 events occurred between April 4th, 2021, and May 22nd, 2021, within this cluster.

Cluster 19 covers a longitude range of 75.23° to 84.13° and a latitude range of 31.33° to 38.2°. The selected centroid for this cluster is located 53km north-northeast of Leh, Ladakh, India. Within this cluster, there have been a total of 378 earthquake events. The mean time difference between these events is calculated to be 2.88 days. Notably, the majority of events, specifically 144, occurred with a time difference of 0 days. The largest magnitude earthquake recorded in this cluster is 6.1 ML, while the average magnitude is 3.99 ML. The mean depth of these earthquakes is 20km.

Cluster 21 exhibits a mean time difference of 2.32 days between earthquake events. The majority of events, specifically 166, occurred with a time difference of 0 days. The largest magnitude earthquake recorded in this cluster is 5.5 ML, with an average magnitude of 3.95 ML. The longitude range for this cluster spans from 68.06° to 72.03°, while the latitude range is from 35.53° to 40°. The centroid for this cluster is located at a longitude of 69.9° and a latitude of 37.2°, positioned 58km west of Fazyabad, Afghanistan. The

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

mean depth of the earthquakes in this cluster is 103.89km. A total of 431 earthquake events are recorded within this cluster.

Cluster 23 represents a cluster located near Fazyabad, Afghanistan, characterized by a maximum magnitude of 6.3 ML and a depth of 220km. The longitude range for this cluster extends from 69.79° to 73.58°, while the latitude range spans from 35.3° to 39.5°. The mean time difference between earthquake events in this cluster is calculated to be 1.697 days. A noticeable pattern observed within Cluster 23 is that the majority of time differences between events are less than 2 days. The centroid for this cluster is situated at a longitude of 71.3° and a latitude of 36.6°. A total of 579 earthquake events have been recorded within this cluster. Specifically, 233 events occurred with a time difference of 0 days, and 145 events occurred with a time difference of 1 day.

Cluster 24 exhibits a mean time difference of 3.77 days between earthquake events. The largest magnitude earthquake recorded in this cluster is 6 ML, with an average magnitude of 4.01 ML. The longitude range for this cluster spans from 71.84° to 77.25°, while the latitude range is from 35.26° to 40°. The centroid for this cluster is located at a longitude of 74.3° and a latitude of 38.1°. A total of 248 earthquake events are recorded within this cluster. The majority of events occurred with time differences of 0, 1, 2, 3, or 4 days. Specifically, 66 events occurred with a time difference of 0 days. Regions within the longitude range of 65° to 80° and latitude range of 30° to 40°, as well as the longitude range of 85° to 97° and latitude range of 30° to 40°, are prone to a higher occurrence of earthquakes. Consequently, more clusters are formed in these areas.

V. CONCLUSION

Machine learning is a versatile approach used to address various problems by creating models based on data analysis, enabling the identification of problems and the generation of feasible solutions. Clustering, a fundamental technique in spatial and spatio-temporal data analysis plays a significant role in knowledge discovery. Spatio-temporal datasets, such as seismic events like earthquakes, provide an excellent example of data that includes both location and time of occurrence.Our proposed method utilizes a few iterations to compute the clusters. It relies on neighbourhood bin searching and employs the Euclidean distance metric. The number of clusters formed is determined based on the density of earthquakes in the region. In the clustering process, isolated data points that are far away from all clusters are considered noise or outliers. These outliers, having a substantial spatial distance from any clusters, do not influence the remaining data points (inliers) in our method.The clustering results demonstrate the formation of distinct and non-overlapping clusters. To assess the quality of the clusters, we employ the Silhouette index, which measures the cohesion and separation of data points within clusters. The Silhouette index for our method falls within the range of 0.88 to 0.93, indicating the formation of good clusters.

TABLE I. CLUSTERING RESULT WITH DATA POPULATION

CLUSTER NUMBER	LONGITUDE MEAN	LATITUDE MEAN	DEPTH MEAN	MAGNITUDE MEAN	COUNT OF DATA POINTS
0	97.495	2.715	138	5.3	76
1	94.94	8.62	95	4.65	123
2	93.57	15.68	112	4.55	74
3	73.86	20.2	15.9	3.35	216
4	80.2	16.21	21	2.95	125
5	94.28	24.2015	95.2	3.7	443
6	95.51	21.51	104	4.8	101
7	74.215	28.065	27.5	3.2	138
8	80.56	30.96	76.5	4.15	224
9	83.225	27.63	102.5	3.85	60
10	88.625	24.755	75.4	3.6	124
11	87.475	30.225	238	4.3	136
12	92.21	26.55	85.5	4.3	315
13	94.98	30.375	102.5	4.2	81
14	96.855	26.625	129	3.85	90
15	98.385	30.25	144	5.3	48
16	66.67	28.95	150	4.7	113
17	74	33.325	102	4.3	64
18	70.8	32.445	150.5	4.75	135
19	79.68	34.765	121	4.1	378
20	92.63	35.09	125	4.65	73
21	70.045	37.765	216	4.15	431
22	64.35	34.125	120	4.6	47
23	71.685	37.4	179.5	4.85	579
24	74.5325	37.63	230.5	4.4	248

DECLARATION

It is Ph.D research work contributed by Ms Swati Meshramunder the supervision of Dr Kishor Wagh.Both the authors have contributed in preparation of manuscript. This research has not received financial support from third party. As per my understanding authors have no conflict of interest with third person/party.The dataset is available on https://seismo.gov.in/

(MSID: 64ba64fa-f6e1-475e-8cee-392966d1e869) has been posted as a preprint on Research Square through the In Review preprint service offered by SpringerNature during journal submission with the doihttps://doi.org/10.21203/rs.3.rs-3068567/v1
Further, We declare this work has not been published in other journal.

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] K. B. Chimwayi and J. Anuradha, "Clustering West Nile Virus Spatio-temporal data using ST-DBSCAN," *Procedia Computer Science*, vol. 132, pp. 1218–1227, 2018, doi: 10.1016/j.procs.2018.05.037.

[3] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005, doi: 10.1109/TNN.2005.845141.

[4] Z. Shi and L. S. C. Pun-Cheng, "Spatiotemporal Data Clustering: A Survey of Methods," *ISPRS International Journal of Geo-Information*, vol. 8, no. 3, Art. no. 3, Mar. 2019, doi: 10.3390/ijgi8030112.

[5] M. Burch, I. Tauroseviciute, and G. M. Guridi, "Visual Analysis of Spatio-Temporal Earthquake Events," in Proceedings of the 15th International Symposium on Visual Information Communication and Interaction, Chur Switzerland: ACM, Aug. 2022, pp. 1–5. doi: 10.1145/3554944.3554959.

[6] "Development Of Probabilistic Seismic Hazard Map Of India,Technical Report Of The Working Committee Of Experts (WCE) Constituted By The National Disaster Management Authority Govt. Of India, New Delhi."

[7] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, Jan. 2007, doi: 10.1016/j.datak.2006.01.013.

[8] Li, Ling, and Dezhong Yao. "A new method of spatio-temporal topographic mapping by correlation coefficient of k-means cluster." Brain topography 19 (2007): 161-176.

[9] P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means cluster analysis in earthquake epicenter clustering," *International Journal of Advances in Intelligent Informatics*, vol. 3, no. 2, Art. no. 2, Jul. 2017, doi: 10.26555/ijain.v3i2.100.

[10] 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, p. 112846, Jan. 2020, doi: 10.1016/j.eswa.2019.112846

[11] 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, p. 112846, Jan. 2020, doi: 10.1016/j.eswa.2019.112846.

[12] L. M. D. Owsley, L. E. Atlas, and G. D. Bernard, "Automatic clustering of vector time-series for manufacturing machine monitoring," in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Apr. 1997, pp. 3393–3396 vol.4. doi: 10.1109/ICASSP.1997.595522.

[13] G. P. K. Wu and K. C. C. Chan, "Discovery of Spatio-Temporal Patterns in Multivariate Spatial Time Series," *ACM/IMS Trans. Data Sci.*, vol. 1, no. 2, p. 11:1-11:22, May 2020, doi: 10.1145/3374748.

[14] M. Wang, A. Wang, and A. Li, "Mining Spatial-temporal Clusters from Geo-databases," in Advanced Data Mining and Applications, X. Li, O. R. Zaïane, and Z. Li, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 263–270. doi: 10.1007/11811305_29.

[15] D. Fitrianah, H. Fahmi, A. N. Hidayanto, and A. M. Arymurthy, "Improved partitioning technique for density cube-based spatio-temporal clustering method," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, Part A, pp. 8234–8244, Nov. 2022, doi: 10.1016/j.jksuci.2022.08.006.

[16] R. K. Vijay and S. J. Nanda, "Earthquake pattern analysis using subsequence time series clustering," Pattern Anal Applic, vol. 26, no. 1, pp. 19–37, Feb. 2023, doi: 10.1007/s10044-022-01092-1