Geospatial Consistency in Clustering: Assessing Latitude and Longitude Stability

Praveen Kumar V.S*
Mahatma Gandhi University, Priyadarsini Hills P.O., Kottayam – 686 560.
Dr. Sajimon Abraham
Mahatma Gandhi University, Priyadarsini Hills P.O., Kottayam – 686 560.
Mr. Sijo Thomas
Mahatma Gandhi University, Priyadarsini Hills P.O., Kottayam – 686 560.
Dr. Nishad A
Department of Higher Secondary Education, Kerala.
Dr. Benymol Jose
Marian College, Kuttikkanam , Kerala.

Abstract

Understanding the movement of objects through spatio-temporal data is important for timely interventions in complex areas related to human mobility and the trajectories of moving objects. The spatio-temporal data forms the basis for the development of applications in mobility management that have an influence on every aspect of human life and object tracking. With latitude, longitude, and time data, continuous mobility tracking is possible and provides very valuable insights for applications that depend on unique mobility characteristics. Mobility data supports many research studies and predictive applications. This includes travel behavior analysis, geomatic applications, and transportation system evaluations. It is very important for analysis in human mobility data since it will be critical to epidemic modeling and traffic prediction in which there is a requirement of quantitative models that would reflect the statistical patterns of individual trajectories. Such models add up to urban planning, traffic forecasting, location-based services, and modeling pandemic spread. Semantically annotated regions are integrated for enrichment of meaningful attributes over trajectories data; this results in an attribute-enriched trajectory. This study also includes the SemTraClus algorithm [6] to cluster and prioritize semantic regions within spatio-temporal trajectories, The performance is analyzed by comparing DBSCAN clusterings with K-means and BIRCH methods, and the evaluation made will be also based on weightage participation of users and Silhouette scores. GeoLife Trajectory Dataset from Microsoft Research Asia [18] is used in the purpose.

Key Words: Moving object trajectory; Point of Interest; Spatio-temporal data; Clustering Comparison.

*Mahatma Gandhi University, Priyadarsini Hills P.O., Kottayam – 686560. Email: praveenplavila@yahoo.co.in

1 Introduction

Compared to activity recognition, predicting activities is a more challenging task because it involves inferring future activities based on existing features in the current phase [1]. Activity prediction relies solely on historical trajectory data features, which may or may not incorporate contextual information. Statistical or machine learning techniques are applied to generate predictions for future activities before the current phase. In essence, while an individual is in motion, the application acquires their location information as raw trajectories—a sequence of spatio-temporal points collected over time [2]. With the increasing prevalence of context-sensing applications that rely on location data, the generation and storage of mobility data have become common practices. Consequently, there is a growing demand for efficient analysis and knowledge extraction from this data across various application domains [3].

In light of the proliferation of the Internet of Things (IoT) and the deluge of Big Data generated on the Internet, such as weather channels and social network interactions (e.g., Flickr, Facebook, Twitter, Foursquare), it is now possible to collect vast volumes of movement data pertaining to people, animals, and objects such as cars, buses, drones, etc. [4]. The prediction of an object's activity based on trajectory data necessitates proper clustering and consideration of other attributes associated with that object. In this study, we primarily focus on clustering applications with trajectory data. Nishad A and Sajimon Abraham propose an algorithm named SemTraClus [6], which extracts revisited points, stay points, and user participation weights in different geographical areas. For the implementation of the SemTraClus algorithm [6], they exclusively employ the DBSCAN clustering method.

In this paper, we implement and evaluate the clustering method (DBSCAN) used in the SemTraClus algorithm, and we also implement and evaluate other clustering methods, namely BRICH and K-means, using the same dataset and algorithm.

Our evaluation demonstrates that the BRICH clustering method yields more accurate results in clustering based on the Silhouette score [17]. This increased accuracy leads to more meaningful results in trajectory data processing, achieved by incorporating additional attributes.

2 Related Works

Moving Object Data processing is emerging as a noteworthy area of research. Various studies on Moving Object data cover diverse aspects of Big Data, including representation, indexing, retrieval, and analysis of trajectory data. In this context, we explore some notable works in the field of Points of Interest extraction. In a study published in 2016 [7], human mobility patterns are discerned from space-time points recorded on social networking sites. The outcome of this research is a semantically enriched dataset that opens up new possibilities for modeling human movement behavior. The authors have also published a paper [5] proposing a Business Intelligence tool named "Predict-Move." This tool assesses the potential for further customer movement from a Point of Interest (POI) to other businesses within large commercial establishments, enhancing customer services, potentially boosting business volume and productivity. In a work published in 2008 [8], trajectories are characterized as sequences of stops and movements. Stops represent crucial points in the movement track, tailored to specific contexts, such as tourist destinations in the realm of tourism, storage facilities in freight management, or traffic hotspots in transportation management. This method marks one of the earliest documented instances of semantic trajectory processing. In another model presented in [9], the authors introduce an innovative approach to identifying interesting places within trajectories, with a primary focus on directional variations. The proposed approach has been tested with real trajectory data from oceanic fishing vessels, with the goal of automatically detecting the locations where vessels engage in fishing activities. Marco A. Beber et al. [10] propose a novel method for recognizing multiple activities occurring at a single location and identifying all individuals involved in group activities. This is achieved by analyzing people's trajectories and extracting insights from social media data. Abraham S and Lal [11] put forth a method for identifying the similarity of moving objects along a restricted path, using a combination of structural and sequential similarity in movement trajectories. They also introduce an encoding technique for managing road network information. In the work titled "Developing a Spatial-Temporal Contextual and Semantic Trajectory Clustering Framework," published in 2017 [12], the authors introduce a two-dimensional trajectory representation method that encompasses attributes beyond spatio-temporal aspects. This method extracts and categorizes the contextual and semantic dimensions of traveling object data to provide meaningful analysis. Contextual information pertains to the surrounding factors associated with the moving object, while semantic information characterizes the motivation for the object's movement.

Effective clustering is essential for categorizing trajectory points according to their application context. Various clustering methods have been developed, implemented, and evaluated in various research studies and publications. The most frequently used clustering algorithm is DBSCAN.

In a study published in 2014 [13], the evaluation of different versions of DBSCAN and its variations is carried out, and their limitations are documented.

Another work titled "Differentially Private and Utility-Aware Publication of Trajectory Data," published in 2020 [14], explores the application scenarios of two clustering algorithms, K-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The study analyzes and presents the advantages and disadvantages of each algorithm using actual ship's Automatic Identification System (AIS) data, facilitating further information mining of trajectory data.

The clustering algorithm BRICH [15], first published in 1997, is implemented in a system named BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Extensive research is conducted to assess its performance in terms of memory requirements, processing time, clustering quality, stability, and scalability. The study also includes comparisons with other available methods, concluding that BIRCH stands as the most suitable clustering method for handling large datasets.

3 Methodology

3.1 Overview

The study employs three mobility clustering methods and compares their efficiency. The baseline method utilized is the recently published SemTraClus algorithm [6]. This algorithm computes users' intersection points, stay points, revisited points, and weightage participation based on their trajectories. The chosen user trajectories are sourced from the Geo-life Microsoft dataset [16], and they serve as the foundation for this research. Within the dataset, the clustering algorithms DBSCAN, K-Means, and BRICH are applied, generating clusters for each respective algorithm. Additionally, the weightage participation (WP) of users at different locations is extracted and compared using evaluation criteria. The efficiency and validity of the clustering methods are assessed through the Silhouette score [17].

3.2 Data Description

This GPS trajectory dataset was collected within the Geolife project at Microsoft Research Asia [18]. It comprises data from 182 users over a span of more than five years, ranging from April 2007 to August 2012. Each GPS trajectory in this dataset is represented as a sequence of time-stamped points, each of which includes information regarding latitude, longitude, and altitude. The dataset encompasses a total of 17,621 trajectories, covering a distance of 1,292,951 kilometers and a cumulative duration of 50,176 hours. These trajectories were recorded

using various GPS loggers and GPS phones, resulting in a wide range of sampling rates. Notably, 91.5 percent of the trajectories feature dense representation, with data points recorded every 1 to 5 seconds or at intervals of 5 to 10 meters.

This dataset captures a diverse spectrum of users' outdoor movements, encompassing not only everyday routines like commuting to work and going home but also leisure and sports activities such as shopping, sightseeing, dining, hiking, and cycling. Researchers can employ this trajectory dataset in numerous domains, including mobility pattern mining, user activity recognition, location-based social networks, location privacy, and location recommendation. While this dataset is extensively distributed across more than 30 cities in China and even some cities in the USA and Europe, the majority of the data originates from Beijing, China.

3.3 Application of DBSCAN in SemTraClus

Clustering is a popular machine learning technique used to group similar data points together based on their similarities or differences. Clustering algorithms aim to partition data points into different clusters to discover hidden patterns and structures in the data.

DBSCAN [19] is a fundamental density-based clustering algorithm. Its advantage lies in its ability to discover clusters with arbitrary shapes and sizes. The algorithm typically treats clusters as dense regions of objects in the data space that are separated by regions of low-density objects. The algorithm has two input parameters: radius ε and Min Pts. To understand the process of the algorithm, some concepts and definitions must be introduced.

Definition 1: The neighborhood within a radius ε of a given object is called the ε-neighborhood of the object.

Definition 2: If the ε-neighborhood of an object contains at least a minimum number σ of objects, then the object is called a σ-core object.

Definition 3: Given a set of data objects D, we say that an object p is directly density-reachable from object q if p is within the ε-neighborhood of q, and q is a σ-core object.

Definition 4: An object p is density-reachable from object q with respect to ε and σ in a given set of data objects, D, if there is a chain of objects p1, p2, p3, ..., pn, where p1 = q and pn = p, and each pi is directly density-reachable from pi-1 with respect to ε and σ.

Definition 5: An object p is density-connected to object q with respect to ε and σ in a given set of data objects, D, if there is an object o€D such that both p and q are density-reachable from o with respect to ε and σ.

3.3.1 Steps of DBScan in SemTraClus

- Step 1: Preprocess the data.
- Step 2: Transform the data points using the DBSCAN algorithm with clustering criteria, specifying a minimum of 4 clusters and a minimum of 14 points within each cluster.
- Step 3: Partition the data into 4 clusters. Any data points that do not belong to any cluster are treated as noise and subsequently removed.
- Step 4: Visualize the data points allocated to different clusters.

3.4 Application of k-means in SemTraClus

K-means clustering stands out as one of the most widely utilized and straightforward clustering algorithms due to its efficiency.

K-means clustering, a partition-based algorithm, segments a dataset into k non-overlapping clusters. The primary objective is to minimize the sum of squared distances between each data point and its nearest cluster center, often referred to as the within-cluster sum of squares (WCSS).

The algorithm operates as follows:

1. Initialization: Randomly select k initial centroids from the dataset.
2. Assignment: Allocate each data point to the nearest centroid, thus forming k clusters.
3. Update: Reassess the centroid of each cluster as the mean of all data points assigned to it.
4. Repeat steps 2 and 3 until either the centroids no longer change significantly or a maximum number of iterations is reached.

K-means clustering boasts several advantages, including its simplicity, speed, and scalability. Nonetheless, it does come with certain limitations, such as the necessity to specify the number of clusters, sensitivity to the selection of initial centroids, and its tendency to converge to local optima.

3.4.1 Steps of K-Means in SemTraClus

- Step 1: Preprocess the dataset.
- Step 2: Apply the K-means algorithm to transform the data points, setting the criteria for 4 clusters and using a random state of 15.
- Step 3: Perform clustering on the dataset, generating 4 clusters. Any data points that do not belong to any of these clusters are treated as noise and subsequently removed.
- Step 4: Visualize the data points allocated to different clusters.

3.5 Application of BRICH in SemTraClus

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm is a popular choice, especially well-suited for handling large datasets. Numerous situations and experiments have demonstrated its efficiency in comparison to K-means and DBSCAN methods [20].

The algorithm operates as follows:

1. Initialization: Specify a clustering threshold and a maximum number of clusters, and initialize an empty tree.

2. Clustering: Begin inserting each data point into the tree, starting from the root. If a leaf node can accommodate the data point without exceeding the threshold, add it to the node. In cases where it would exceed the threshold, the node is split into two new subclusters.
3. Merging: Once all data points have been inserted, the algorithm proceeds to merge subclusters that exhibit similarity until the desired number of clusters is achieved.

3.5.1 Steps of BIRCH in SemTraClus

- Step 1: Preprocess the dataset.
- Step 2: Transform the data points using the K-means algorithm, specifying criteria for 4 clusters, and setting a random state of 15.
- Step 3: Perform clustering on the data, creating 4 clusters. Unlike other clustering algorithms, BIRCH does not consider any data points as noise and uses all data points in the clusters.
- Step 4: Visualize the data points allocated to different clusters.

3.6 Weightage Participation

Trajectory datasets provide valuable information about the movement of objects over time. These datasets find applications in various fields, including transportation and logistics, where tracking object movements is critical. Weightage participation by users can be a valuable approach within trajectory datasets, enabling users to assign weights or importance to different features or attributes in the dataset. In this study, we aim to delve into the concept of user weightage participation in trajectory datasets and uncover its potential benefits.

3.6.1 Steps for Calculating Weightage participation in SemTraClus Algorithm

The SemTraClus algorithm serves to identify Points of Interest (POI) from various trajectories, bringing together similar semantic locations into clusters. Each cluster comprises a series of connected locations associated with different individual users, essentially forming sub-trajectories connecting interesting locations or semantic points. These clusters are considered as semantic regions where enrichment can be applied. Semantic tagging is facilitated through a POI database, which stores and updates waypoints, landmarks, facilities, and other relevant information about each location [25].

Given that each cluster represents a semantic sub-trajectory involving multiple users, it becomes crucial to gauge the level of user participation within a specific geographical area. The priority of a semantic region correlates with the degree of interest displayed by different users in that region. To quantify this, a measure called "Weightage of Participation" (WP) is introduced. WP determines both the priority value of an individual trajectory within a semantic region and the overall priority of the semantic regions in the geographical area.

WP for a trajectory directly measures a user's interest in a semantic location. The calculation of WP for different movement trajectories is based on three factors: stay time, the count of location revisits, and the count of intersecting points. Each of these attributes exerts varying levels of influence in determining movement behavior.

A user's semantic trajectory encompasses various cluster points during a travel session. The degree of a user's participation in a cluster depends on two parameters: Spatial Density α and Temporal Presence β. Spatial density for a user trajectory Uj in a cluster Ci is defined as the ratio of the number of locations visited by user Uj in cluster Ci to the total number of semantic locations in the cluster. This spatial density, which reflects a user's presence in the identified semantic region, is given by:

$$\alpha(i,j) = (\text{No. of locations visited by Uj in Ci}) / (\text{Total no. of locations in cluster Ci})$$

Temporal presence β quantifies the extent of a user's stay duration within a semantic region. It is the ratio of the total stay time duration of a user Uj in cluster Ci to the total time spent by all users in cluster Ci, expressed as:

$$\beta(i,j) = (\text{Stay time duration of Uj in Ci}) / (\text{Total time spent by all users in Ci})$$

The WP of a user Uj in a cluster Ci serves as a metric to gauge the user's interest in that cluster. It is calculated as the averaged sum of Spatial Density α and Temporal Presence β, as shown below:

$$WP(i,j) = (\alpha(i,j) + \beta(i,j)) / 2$$

3.7 Comparison of various clustering algorithm using Various Methods

Here's are the various comparison methods used to compare the three clustering algorithms: K-Means, BIRCH, and DBSCAN.

- Silhouette Score
- Calinski-Harabasz
- Davies-Bouldin
- Average Cluster Size
- Detection of Noise Points (Applicable to DBSCAN only)
- Mean Latitude and Longitude
- Standard Deviation of Latitude and Longitude

4 Logical Framework of the Process Involved

Main Framework Steps:

1. Data Collection and Semantic Point Extraction: Gather the data and extract semantic points, including intersections, stay points, and revisited points.

2. Data Preprocessing: Preprocess the data by eliminating duplicates and null values, ensuring it is ready for the implementation of various algorithms.
3. Algorithm Selection and Implementation: Import different algorithms such as DBScan, K-Means, and BIRCH. Apply these algorithms to the dataset while setting a consistent number of clusters, with 4 clusters being used throughout each algorithm.
4. Results Visualization: Visualize the results produced by each algorithm to identify the clusters and their characteristics.
5. Cluster Accuracy Assessment: Evaluate the accuracy of the clusters using the Silhouette Score method.
6. Comparison and Result Visualization: Compare the results from different algorithms and visualize the outcomes for a comprehensive analysis.

5 Evaluation

- Microsoft Geolife trajectory data consist of 18670 trajectories of 182 user journeys that have 24876978 trajectory points with a total distance of 1292951 kilometers and a total duration of 50176 hours collected in a period of over 5 years (from April 2007 to August 2012).
- We have selected different tracks of 21 users which constitute 965 trajectories that have 1164069 trajectory points from the dataset.
- The algorithm has been implemented in python 3.10.2. All experiments are conducted in Intel Core i5 machine with 8GB RAM.

5.1 Selected User-trajectory Details

5.2 Revisited points

We obtained the revisited points of users from the original dataset [18] for use in our clustering algorithms. The geographical locations of users in their respective areas are visualized in Figure 1.

5.2.1 Most Revisited Co-ordinates by users

Table 2 shows the location details and number of revisits of users. The table shows the details of users who have revisited the locations more than 4 times.

5.3 Intersection Points

We identified the intersection points of users from the original dataset [18] for use in our clustering algorithms. The geographical locations of users in their respective areas are depicted in Figure 2.

User	No.of trajectories
107	3
108	9
109	4
110	25
111	44
112	212
113	32
114	23
115	184
116	3
117	8
118	5
119	45
120	2
121	5
122	16
123	5
124	10
125	57
126	263
127	10
Total Trajectories	965

Table 1: you can find the details of 21 users along with their respective trajectory points.

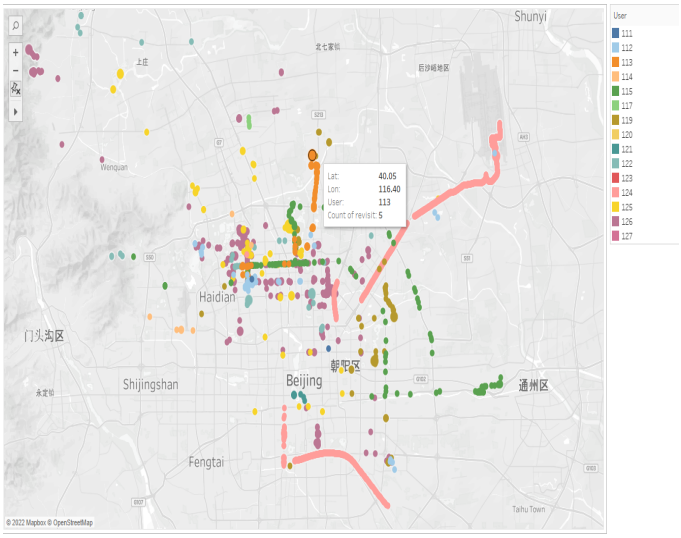


Figure 1: shows the revisited points of users in the trajectory dataset

5.4 Stay Points of users

From the original dataset [18], we identified the stay points of users for use in our clustering algorithms. The geographical locations of users in their respective areas are illustrated in Figure 3

user	latitude	longitude	Number of Revisits
125	40.0094	116.375	9
126	39.8217	119.478	8
126	39.8217	119.478	7
124	40.0519	116.61	6
113	40.0527	116.401	6
113	40.0527	116.401	6
113	40.0527	116.401	6
113	40.0527	116.401	6
113	40.0527	116.401	6
126	40.2123	116.272	6
126	39.8217	119.478	5
119	39.9538	116.493	5
122	39.9681	116.4	5
119	39.9271	116.471	5
126	39.8217	119.478	5

Table 2: we can clearly see that user 125 has the greatest number of revisited points followed by user 126

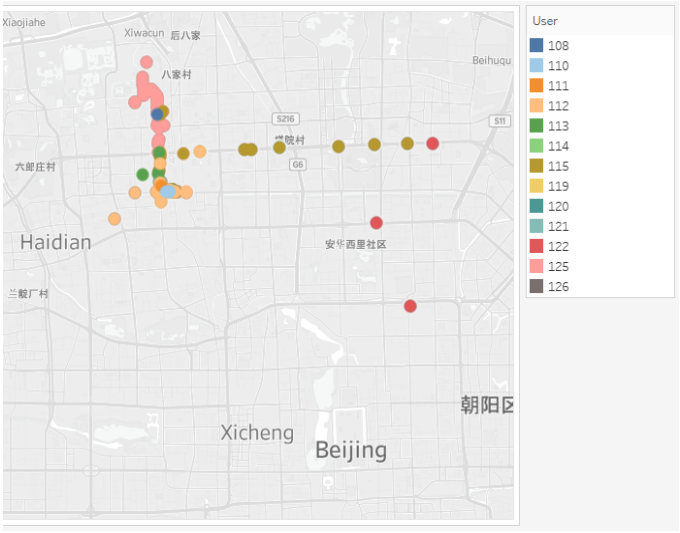


Figure 2: shows the intersection points of users in semantic region

5.5 Semantic point extraction and density clustering

- The SemTraClus algorithm extracts stay points, revisited points and intersecting points with the spatial and temporal threshold values 2 and 72 respectively.
- In Geo-life data set among the 1164069 trajectory points of 965 trajectories with 21 different users our algorithm extracts 8523 semantic locations which is shown in Table 3.

5.6 DBSCAN-Cluster Details

We have applied the DBScan algorithm in the selected user trajectories which divides the trajectories into 4 clusters. The

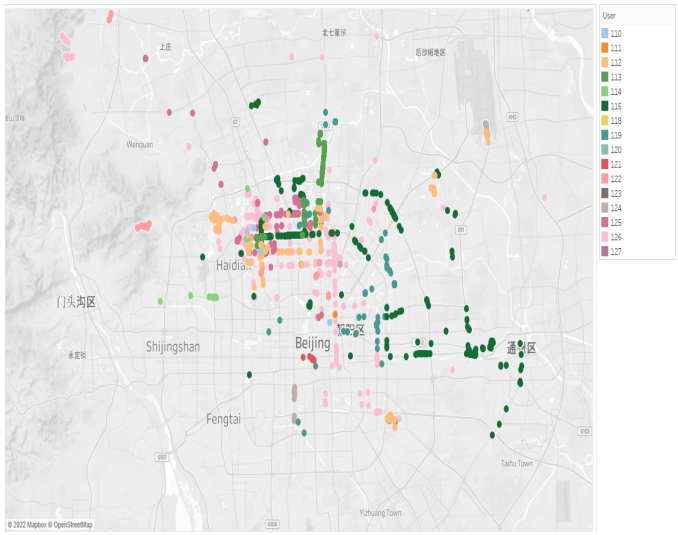


Figure 3: shows the stay points of users in semantic region

	User Count	21
	Trajectory points	1164069
Stay Points	Identified	27488
Stay Points	Valid	1460
Revisit	Identified	15239
Revisit	Valid	6899
Intersection	Identified	328
Intersection	Valid	164

Table 3

details of the trajectory points of users are shown in Table 4.

5.7 KMEANS-Cluster Details

We have applied the K-Means algorithm in the selected user trajectories which divides the trajectories into 4 clusters. The details of the trajectory points of users are shown in table 5.

5.8 BIRCH-Cluster Details

We have applied the BIRCH algorithm in the selected user trajectories which divides the trajectories into 4 clusters. The details of the trajectory points of users are shown in table 6.

5.9 Comparative graphs of clusters

Figure 4 displays a graph illustrating the number of semantic points obtained in each cluster when the DBScan algorithm is used for clustering. Likewise, Figure 5 presents a graph depicting the number of semantic points in each cluster for the K-Means algorithm. Additionally, Figure 6 provides insight into the number of clusters when implementing the BIRCH algorithm.

Users	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Grand Total
108	1				1
110	9				9
111	9	3			12
112	720				720
113	468				468
114	12				12
115	549	4			553
117	4				4
119	1495				1495
120	24				24
121	15				15
122	364				364
123		25			25
124	1995	727		45	2767
125	316				316
126	1721				1721
127			17		17
TOTAL	7702	759	17	45	8523

Table 4

Users	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Grand Total
108			1		1
110			9		9
111		3	9		12
112			720		720
113			468		468
114			12		12
115	3	4	546		553
117			4		4
119			1495		1495
120			24		24
121			15		15
122			374		374
123		25			25
124		726	1995	46	2767
125			317		317
126	392		1329		1721
127	17				17
TOTAL		758	7318	46	8534

Table 5

5.10 Comparison Chart of Brich, K-means and DB Scan

We conducted a comparison of clustering details among the DBScan, K-Means, and BIRCH algorithms. The clusters are labeled as 0, 1, 2, and 3. Figure 7 presents the distribution of revisited points, stay points, and intersection points of users within the various clusters formed by these algorithms

Users	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Grand Total
108	1				1
110	9				9
111	9	3			12
112	720				720
113	468				468
114	12				12
115	549	4			553
117	4				4
119	1495				1495
120	24				24
121	15				15
122	364		10		374
123		25			25
124	1995	726		46	2767
125	316	1			317
126	1721				1721
127	17				17
TOTAL	7719	759	10	46	8534

Table 6

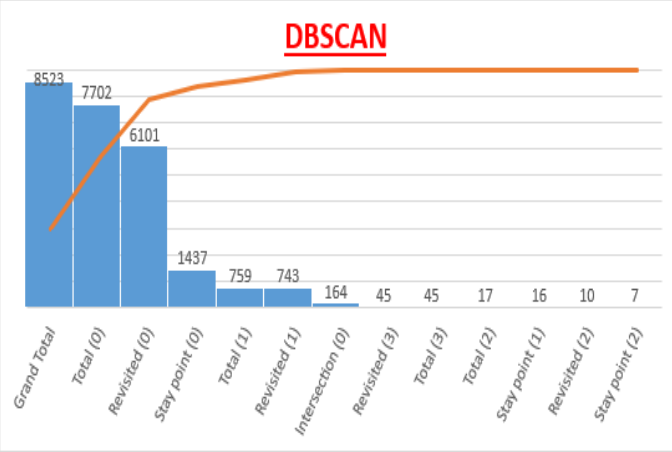


Figure 4

5.11 Visualization of clustering algorithms

After implementing the DBScan, K-Means, and BIRCH algorithms, we generated cluster-wise visualizations of trajectory points for users with stay points, intersection points, and revisited points. These visualizations for DBScan, BIRCH, and K-Means are depicted in Figures 8, 9, and 10, respectively

6 Weightage participation of users

6.1 Weightage participation - DB-Scan

The weightage participation of selected 21 users were calculated as mentioned in the section 3.6.1. The result of the weightage participation of users calculated in the clusters

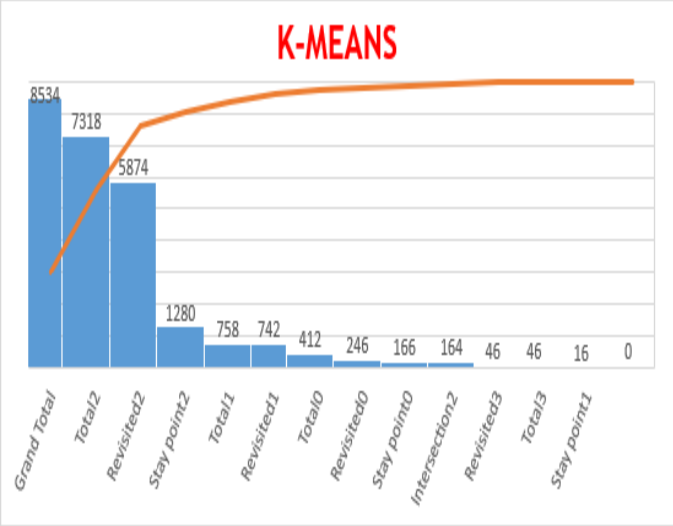


Figure 5

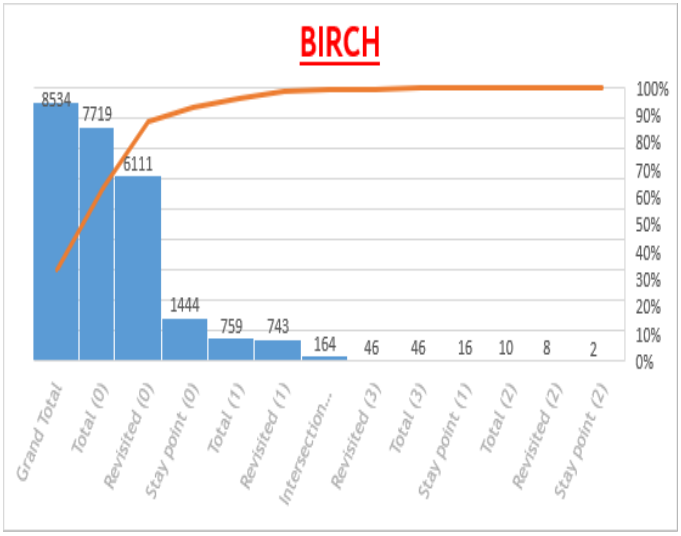


Figure 6

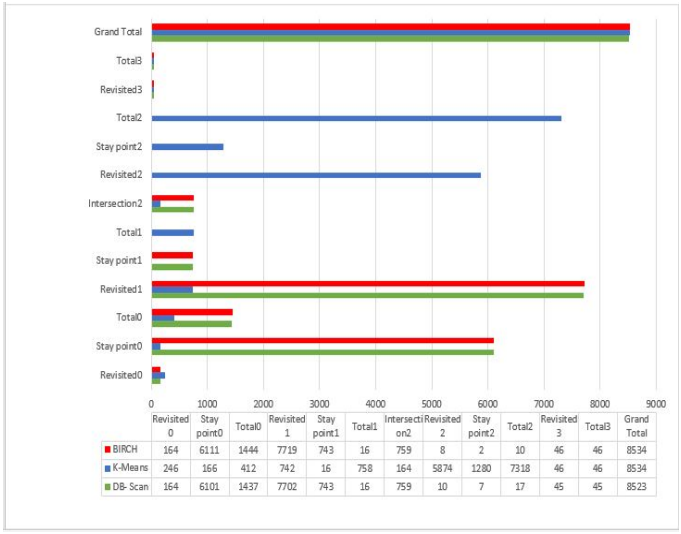
formed using DBScan algorithm are shown in the table 8.

6.2 Weightage participation – K-means

The weightage participation of selected 21 users were calculated as mentioned in the section 3.6.1. The result of the weightage participation of users calculated in the clusters formed using K-means algorithm are shown in the table 9.

6.3 Weightage participation – BIRCH

The weightage participation of selected 21 users were calculated as mentioned in the section 3.6.1. The result of the weightage participation of users calculated in the clusters formed using BIRCH algorithm are shown in the table 10.



	Revisited 0	Stay point0	Total0	Revisited 1	Stay point1	Total1	Intersection2	Revisited 2	Stay point2	Total2	Revisited 3	Total3	Grand Total
BIRCH	164	6111	1444	7719	743	16	759	8	2	10	46	46	8534
K-Means	246	166	412	742	16	758	164	5874	1280	7318	46	46	8534
DB- Scan	164	6101	1437	7702	743	16	759	10	7	17	45	45	8523

Figure 7: shows the comparison chart as well as the comparison table for each of DBScan, K-Means and BIRCH algorithm.

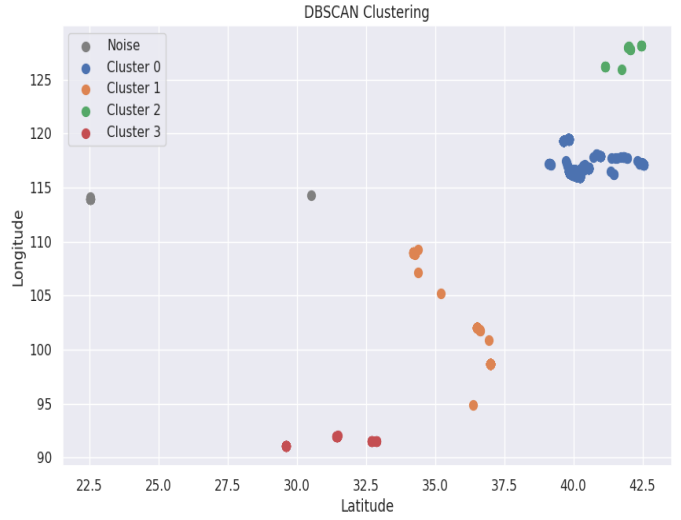


Figure 8

6.4 Comparison chart of Weightage participation

The weightage participation of the users in the various clusters created by using the algorithms DBScan, BIRCH and K-Means are shown in the tables 7, 8 and 9 and its comparison charts are shown in the figures 11 and 12. When we look at the weightage participation of various users in clusters formed using the three algorithms DBScan, K-Means and BIRCH, we can say that users in the clusters formed using BIRCH algorithm have more weightage of participation when compared to the clusters formed using, DBScan and K-Means. This can also be observed when comparing Tables 8, 9, and 10, and it is also prominent in Figures 11 and 12.

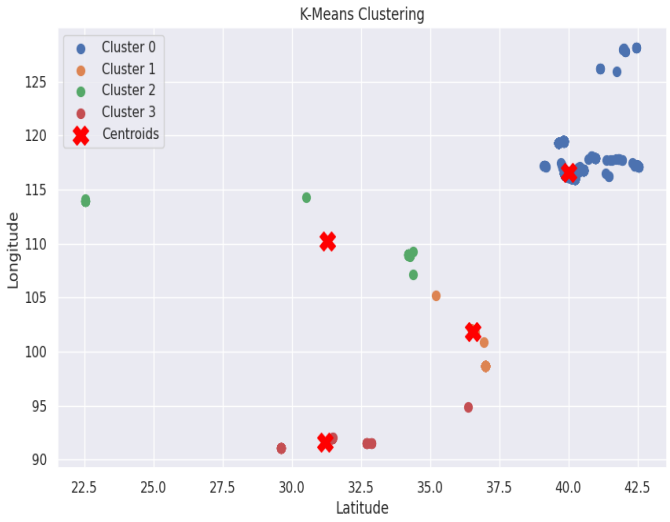


Figure 9

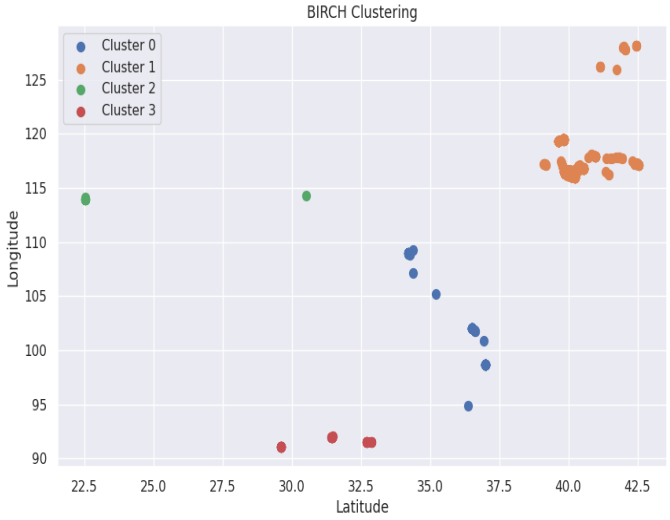


Figure 10

7 Comparison of various clustering algorithm using Various Methods

7.1 Silhouette

Silhouette refers to a method of interpretation and validation of consistency within clusters of data.

1. Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique.
2. The technique provides a score representation of how well each object has been classified.
3. Its value ranges from -1 to 1.

- A score of 1 indicates that clusters are well apart and clearly distinguished.

Users	Spatial density α	Temporal Presence β	Weightage of participation (WP)
108	0.00013	0	0.000065
110	0.001169	0.013623	0.007396
111	0.006432	0.774608	0.39052
112	0.093519	0.130615	0.112067
113	0.060787	0.140294	0.100541
114	0.001559	0.000516	0.001037
115	0.076571	0.239406	0.157989
117	0.00052	0	0.00026
119	0.194181	0.06536	0.129771
120	0.003117	0.009779	0.006448
121	0.001948	0.000205	0.001077
122	0.047279	0.011322	0.0293
123	0.032895	0.145091	0.088993
124	2.215704	0.001145	1.108424
125	0.040785	0.041293	0.041039
126	0.223406	0.426743	0.325074
127	1	1	1

Table 7: shows the weightage participation of 21 users after clustering using DBScan algorithm.

Users	Spatial density α	Temporal Presence β	Weightage of participation (WP)
108	0.000137	0	0.0000685
110	0.00123	0.017945	0.009588
111	0.005188	0.775731	0.390459
112	0.098388	0.172054	0.135221
113	0.063952	0.184805	0.124378
114	0.00164	0.00068	0.001160
115	0.087169	0.29164	0.189405
117	0.000547	0	0.000273
119	0.204291	0.086097	0.145194
120	0.00328	0.012881	0.008080
121	0.00205	0.00027	0.001160
122	0.051107	0.016102	0.033604
123	0.032982	0.145091	0.089036
124	1.230399	0.001508	0.615953
125	0.043318	0.054393	0.048856
126	1.133063	1.233167	1.183115
127	0.041262	0.007635	0.024449

Table 8: shows the weightage participation of 21 users after clustering using K-Means algorithm

- A score of 0 suggests that clusters are indifferent or the distance between them is not significant.
- A score of -1 implies that clusters are assigned in the wrong way.

Users	Spatial density α	Temporal Presence β	Weightage of participation (WP)
108	0.00013	0	0.000065
110	0.001166	0.013598	0.007382
111	0.005119	0.774602	0.389860
112	0.093276	0.130372	0.111824
113	0.06063	0.140034	0.100332
114	0.001555	0.000515	0.001035
115	0.076393	0.239117	0.157755
117	0.000518	0	0.000259
119	0.193678	0.065239	0.129458
120	0.003109	0.009761	0.006435
121	0.001943	0.000205	0.001074
122	1.047156	1.011301	1.029228
123	0.032938	0.145091	0.089015
124	2.214975	0.001142	1.108059
125	0.042255	0.041216	0.041736
126	0.222956	0.42595	0.324453
127	0.002202	0.001857	0.002029

Table 9: shows the weightage participation of 21 users after clustering using BIRCH algorithm

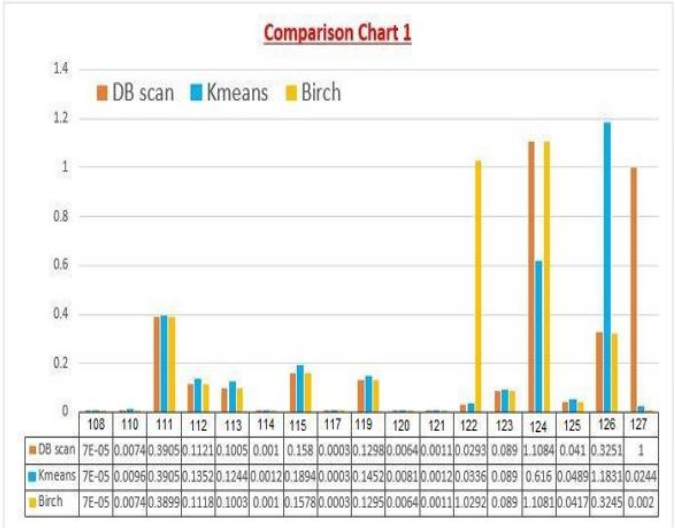


Figure 11: shows the comparison between the three algorithms DBScan, K-Means and BIRCH for the selected users.

Algorithms	Silhouette score
DB-Scan	0.949
BIRCH	0.962
K-MEANS	0.955

Table 10

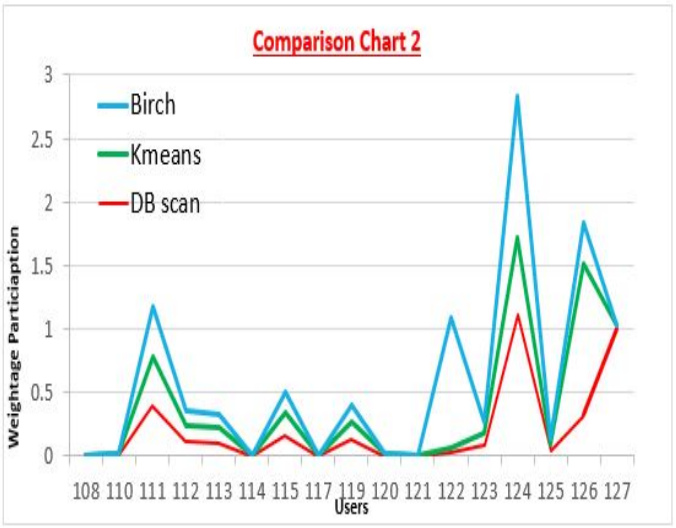


Figure 12: shows the comparison between the three algorithms DBScan, K-Means and BIRCH for the selected users.

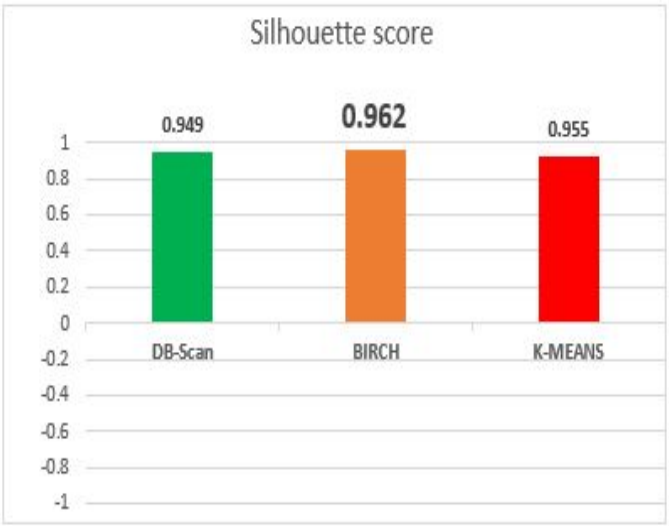


Figure 13

7.2 Calinski-Harabasz and Davies-Bouldin Index

- Variance Ratio Criterion, or Calinski-Harabasz Index, measures the ratio of the total of within-cluster dispersion to between-cluster dispersion in order to assess the quality of a grouping. To put it another way, it evaluates the degree of cluster separation and the density of the data points within each cluster. **Value Range:** Higher values on the non-negative index denote better-defined clusters. There isn't a set maximum. Better clustering is suggested by a higher Calinski-Harabasz score, which denotes that clusters are more distinct and well-separated from one another.
- The average similarity between each cluster and its most

Table 11: Our finding by applying the algorithms in geolife data set clustering are given below

Algorithms	Calinski-Harabasz Index	Davies-Bouldin Index
DB-Scan	86446.169469	0.131204
BIRCH	64518.503687	0.128266
K-MEANS	68196.469322	0.465701

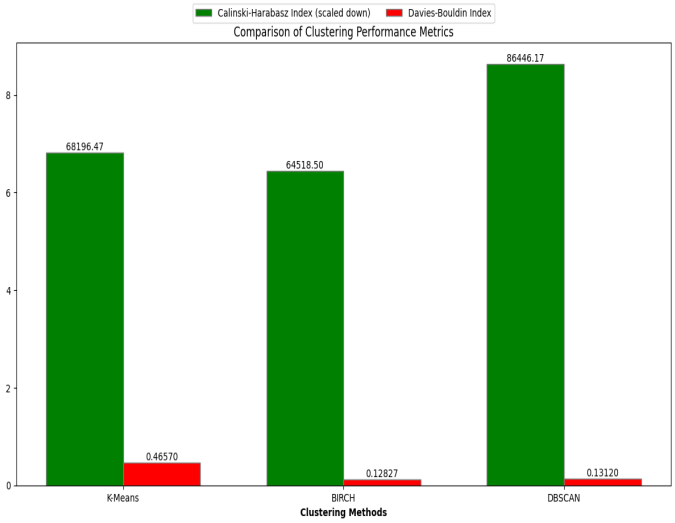


Figure 14

comparable cluster is determined by the Davies-Bouldin Index. This similarity is computed as the within-cluster dispersion to between-cluster separation ratio. **Value Range:** Lower values of the index, which goes from 0 to infinity, indicate better grouping. A lower Davies-Bouldin score denotes a more ideal clustering solution since it indicates that the clusters are compact and well-separated.

7.3 Average Cluster Size

Indicates the mean quantity of points within every group. **Value Range:** Depending on the clustering technique and dataset, varies. Interpretation: While smaller sizes reflect more clusters or smaller groups, bigger average sizes may indicate fewer clusters or broader groupings.

Algorithms	Average Cluster Size
DB-Scan	2130.75
BIRCH	2133.50
K-MEANS	2133.50

Table 12

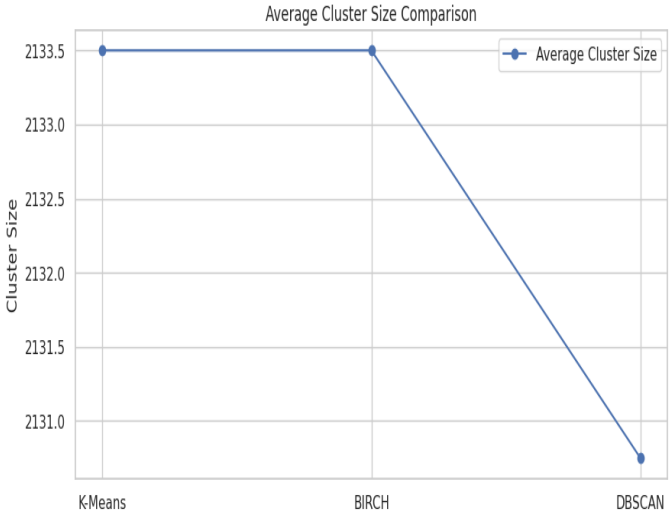


Figure 15

Algorithms	Mean Latitude and Longitude	
DB-Scan	37.365610	109.422861
BIRCH	32.702486	106.068466
K-MEANS	34.764107	105.080717

Table 13

7.4 Mean Latitude and Longitude

The average geographic coordinates of all map data points located within a given cluster. **Value range:** Determined by the maximal range of values of coordinates available in the dataset. Interpretation: These values determine the position of clusters on a map. Certain factors may cause the mean latitude and longitude to be applied to different clusters of population.

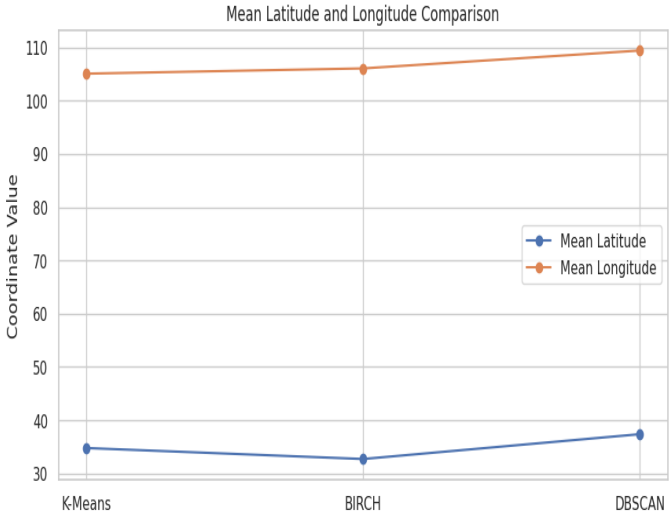


Figure 16

Algorithms	Std Dev Latitude and Longitude	
DB-Scan	0.600423	0.899680
BIRCH	1.085125	0.749838
K-MEANS	1.719273	1.166711

Table 14

Figure 17

7.5 Std Dev Latitude and Longitude

Calculates the spread of data points from the central location of a cluster, in terms of geographic coordinates. **Value Range:** Values are positive, indicating that smaller values represent less dispersion and larger values reflect even more spread out clusters. Interpretation: Lower standard deviations imply tightly packed clusters, whereas higher values indicate dispersal in a wider geographical area.

Based on the given metrics, here's a comparison of the three clustering algorithms: K-Means, BIRCH, and DBSCAN.

1. Silhouette
   BIRCH attained the highest silhouette score of 0.962, indicating that it best separated the clusters and had more compact clusters. K-Means comes very close to BIRCH, scoring 0.955, indicating it performs satisfactorily for clustering. DBSCAN has the lowest Silhouette Score 0.949, implying that its clusters are not quite as well-separated or as compact as those of the other two methods.
2. Calinski-Harabasz
   DBSCAN yields the highest value of 86,446, which indicates that it has produced the most cohesive and well-delineated clusters. K-Means takes the second-highest value of 68,196, meaning that reasonably well-defined cluster separation occurs. BIRCH has the lowest Calinski-Harabasz Index (64,518), which may imply less rigorous cluster boundaries when compared with the other methods.
3. Davies-Bouldin
   BIRCH has the lowest Davies-Bouldin index (0.1283), suggesting that it had the most compact clusters, with the best separation. DBSCAN came in next with a score of 0.1312, similar but a little worse in its performance than BIRCH. K-Means has the highest Davies-Bouldin index (0.4657), indicating that its clusters are less compact and have a lesser separation than those of the other two.
4. Average Cluster Size
   The average cluster sizes for both K-Means and BIRCH algorithms are 2133.50, indicating that their clusters are evenly distributed. With 2130.75, DBSCAN has a somewhat lower average cluster size, which may indicate that it excluded certain points as noise.
5. Detection of Noise Points (Applicable to DBSCAN only)
   DBSCAN managed to detect 11 noise points, which are indicative of its capability to detect outlier behavior, in sharp contrast to other K-Means and BIRCH algorithms that do not provide explicit noise incorporation.
6. Mean Latitude and Longitude
   BIRCH exhibits lower mean latitudinal and longitudinal values than the other two methodologies. DBSCAN has the maximum means, showing that its cluster centers would be much differently positioned from those derived using K-Means and BIRCH.
7. Standard Deviation of Latitude and Longitude
   Among the three algorithms, DBSCAN exhibits the lowest standard deviation for latitude (0.6004), and hence this translates to a more consistent location of its clusters. K-Means shows the most significant latitude standard deviation (1.7193), indicating wide dispersal. For longitude, BIRCH has the lowest standard deviation (0.7498), whereas K-Means has the widest (1.1667).

8 Limitations

The location-based services are now becoming promising areas of research. The availability of datasets used for creating location-based services is very limited. Another limitation of this study is that, the access of other attributes in connection with the location-based data is a challenging task. The privacy-preserving information of the users is confidential and which cannot be accessed without their consent. Clustering of trajectory data with more attributes makes it more meaningful, but the collection and processing of multi-attribute data requires more effort to complete its processing.

9 Future work

We can create a more accurate predictive system based on the location-based data and associated semantic aware attributes. Suppose we get the location-based information and social media interactions of the user under a particular consent-based domain, we can develop a novel system to predict the next activity or movement of the user with proper information.

10 Conclusions

From the metrics discussed, it can be seen that different clustering algorithms have strengths in different purposes of clustering. BIRCH has the highest silhouette value at 0.962 and the lowest Davies-Bouldin index at 0.1283, thus this algorithm is best suited for applications where closely grouped but distinct clusters are of higher priority. K-Means, as much as it is performing in an acceptable manner with close silhouette score of 0.955, is certainly not comparable to BIRCH or DBSCAN considering the compactness since their Davies-Bouldin Index is higher at 0.4657. Instead, DBSCAN has succeeded well in noise point detection considering it does not consider 11 points as outliers and also gains the highest Calinski- Harabasz score: 86,446 depicting that it can form

contiguous clusters that can handle some level of outliers but the compactness is compromised marginally. One other factor is that users in the clusters formed using BIRCH algorithm has more weightage of participation when compared to the clusters formed using, DBSCAN and K-Means. K-Means and BIRCH are distributed similarly in terms of average cluster size, while the slightly lower average cluster size of DBSCAN implies that some points are excluded as noise. DBSCAN also has consistency in latitude, which is indicated by the lowest latitude standard deviation at 0.6004, which may be beneficial for datasets requiring geographical stability in clusters. However, considering the high silhouette score, more weightage participation, and minimal Davies-Bouldin index with no sensitivity of performance to noise, BIRCH is the most acceptable algorithm, having a tight and well-separated clusters thus becoming a reliable one with an application where the prioritized structure and coherence inside clusters are concerned.

Acknowledgments

References

1 L. Xu and M.-P. Kwan, ”Mining sequential activity–travel patterns for individual-level human activity prediction using Bayesian networks,” [periodical/source].

2 C. A. Ferrero, L. O. Alvares, and V. Bogorny, ”Multiple aspect trajectory data analysis: Research challenges and opportunities,” [periodical/source].

3 L. O. Alvares, V. Bogorny, B. Kuijpers, J. Macedo, B. Moelans, and A. Vaisman, ”A model for enriching trajectories with semantic geographical information,” *GIS*, vol. 22, 2007.

4 R. dos S. Mello, V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner, and C. Renso, ”MSTER: A multiple aspect view on trajectories,” [periodical/source].

5 V. S. Praveen Kumar, S. Abraham, and N. A., ”A proposal for an efficient business intelligence tool using spatio-temporal and geo-tag data for strengthening the decision support system,” *8th Pan IIM World Management Conference*, IIM Kozhikode, India, 2021.

6 A. Nishad and S. Abraham, ”SemTraClus: An algorithm for clustering and prioritizing semantic regions of spatio-temporal trajectories,” *International Journal of Computers and Applications*, 2019.

7 K. Siła-Nowicka, J. Vandrol, T. Oshan, J. A. Long, U. Demšar, and A. S. Fotheringham, ”Analysis of human mobility patterns from GPS trajectories and contextual information,” *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 881-906, 2016.

8 S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, ”A conceptual view on trajectories,” *Data and Knowledge Engineering*, vol. 65, no. 1, pp. 126-146, 2008.

9 J. A. M. R. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny, ”DB-SMoT: A direction-based spatio-temporal clustering method,” *5th IEEE International Conference on Intelligent Systems*, London, UK, pp. 114-119, 2010, doi: 10.1109/IS.2010.5548396.

10 M. A. Beber, C. A. Ferrero, R. Fileto, et al., ”Individual and group activity recognition in moving object trajectories,” *Journal of Information and Data Management*, vol. 8, no. 1, pp. 50, 2017.

11 S. Abraham and P. S. Lal, ”Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations,” *Transportation Research Part C: Emerging Technologies*, vol. 23, pp. 109–123, 2012.

12 I. Portugal, P. Alencar, and D. Cowan, ”Developing a spatial–temporal contextual and semantic trajectory clustering framework,” *arXiv preprint arXiv:1712.03900*, 2017.

13 K. Khan, et al., ”DBSCAN: Past, present and future,” *5th International Conference on the Applications of Digital Information and Web Technologies* (ICADIWT 2014), IEEE, 2014.

14 Q. Liu, et al., ”Differentially private and utility-aware publication of trajectory data,” *Expert Systems with Applications*, vol. 180, pp. 115120, 2021.

15 T. Zhang, R. Ramakrishnan, and M. Livny, ”BIRCH: A new data clustering algorithm and its applications,” *Data Mining and Knowledge Discovery*, vol. 1, pp. 141–182, 1997. doi: https://doi.org/10.1023/A:1009783824328

16 *Geo-Life GPS Trajectory Dataset*, Microsoft, available at: https://www.microsoft.com/en-us/download/details.aspx?id=52367

17 ”Silhouette Coefficient: Validating clustering techniques,” available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c: :text=Silhouette

18 ”GeoLife GPS trajectory dataset user guide,” Microsoft Research, available at: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/

19 D. Hsu and S. Johnson, ”A vibrating method based cluster reducing strategy,” *Fifth International

Conference on Fuzzy Systems and Knowledge Discovery*, vol. 2, pp. 376-379, 2008.

20 I. A. Venkatkumar and S. J. K. Shardaben, ”Comparative study of data mining clustering algorithms,” *2016 International Conference on Data Science and Engineering* (ICDSE), Cochin, India, pp. 1-7, 2016, doi: 10.1109/ICDSE.2016.7823946.

21 Available at: https://www.sciencedirect.com/science /article/pii/S0169023X06000218via

22 ”DBSCAN,” Wikipedia, available at: https://en.wikipedia.org/wiki/DBSCAN

23 ”K-means clustering,” Wikipedia, available at: https://en.wikipedia.org/wiki/K-meansclustering

Praveen Kumar V.S He is working as an Ast.professor in a Government Aided College and has 25 years of teaching experience in UG programme and 15 years in PG. His area of interests are Spatio-temporal data mining and Artificial Intelligence for Human Rights.He has published six papers in International journals.

Dr. Sajimon Abraham. (MCA, MSc. (Mathematics), MBA, PhD (Computer Science)). He has been working as Faculty Member in Computer Applications and IT, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India. He currently holds the additional charge of Director (Hon), University Center for International Cooperation. He was previously working as Systems Analyst in Institute of Human Resource Development, Faculty member of Computer Applications in Marian College, Kuttikkanam and Database Architect in Royal University of Bhutan under Colombo Plan on deputation through Ministry of External Affairs, Govt. of India. His research area includes Data Science, Spatio-Temporal Databases, Mobility Mining, Sentiment Analysis, Big Data Analytics, E-learning, Data Base and Data Mining, Data Management, Web Clickstream Analysis, Business Analytics, and he has published 110 articles in National, International Journals and Conference Proceedings.

Mr. Sijo Thomas serves as a research scholar affiliated with the School of Computer Sciences at Mahatma Gandhi University in Kerala, India. In addition to his academic pursuits, he assumes the role of a consultant software architect. His software solutions are utilised by prominent state universities, private universities, and autonomous colleges in India. In addition, he has academic experience of more than 5 years as a faculty of science in colleges. Mr.Sijo Thomas holds a Master’s degree in computer applications from Mahatma Gandhi University in Kerala, and he has also achieved a Master of Philosophy degree from Bharathidasan University in Tamil Nadu, India.

Dr. Nishad A (M.C.A, M.Tech), is a Senior Higher Secondary Teacher in a Government Higher Secondary School His area of research includes Bigdata Analysis, Moving Object Data Mining and Trajectory Clustering. He has published more than 12 papers in International and National journals and conference proceedings.

Dr. Benymol Jose (MCA) She is working as an Associate Professor in a Government Aided College and has 25 years of teaching experience in UG programme. Her main research focuses on Unstructured data and NoSQL databases, Big Data Analytics, Data Mining. data mining and Artificial Intelligence .She has published twelve papers in International journals.