Unveiling Birch's Dominance in Trajectory Clustering: A Comparative Analysis

V S. PRAVEEN KUMAR[1], SAJIMON ABRAHAM[2], SIJO THOMAS[3], NISHAD.A[4], BENYMOL JOSE[5]

[1] *Department of Computer Science, Sas Sndp Yogam College, Konni, Kerala, India,*
[2]*School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India,*
[3]*School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India,*
[4]*Department of Higher Secondary Education, Government of Kerala, India*
[5]*Department of Computer Applications, Marian College, Kuttikanam, Kerala, India*

Abstract— The movement of an object and its associated data is of paramount importance for prompt interventions in challenging areas related to human mobility and the trajectories of moving objects. Spatio-temporal data constitutes the primary resource for developing applications in mobility-based management across all aspects of human existence and other objects. Mobility can be tracked when latitude, longitude, and time information are available. The inferences drawn from mobility data can be utilized for various purposes, particularly in applications where the distinctive features of moving objects hold significance. Mobility data finds utility in diverse studies and predictive applications, such as users' travel experiences, geomatic applications, and transportation system analysis. The importance of analyzing human mobility data spans from epidemic modeling to traffic prediction. There is a need for quantitative models that can encompass the statistical characteristics of individual human trajectories, urban planning, traffic monitoring, and location-based services, and to predict the spread of pandemic diseases. Incorporating Points of Interest in semantic regions allows for the augmentation of attributes in trajectory data, resulting in attribute-enriched trajectories. The SemTraClus algorithm [6] is employed for identifying and clustering semantic regions in spatio-temporal trajectories. This study entails a comparison of the performance of DBSCAN clustering in SemTraClus with other clustering methods, namely K-means, and BRICH. The evaluation of its performance and accuracy considers user participation weighting and the Silhouette score, all using the same dataset. The comparative study of clustering methods is conducted using a real trajectory dataset from the Geo-life project of Microsoft Research Asia [18].

Index Terms- Moving object trajectory, Point of Interest, Spatio-temporal data

I. INTRODUCTION

Compared to activity recognition, predicting activities is a more challenging task because it involves inferring future activities based on existing behaviors in the current phase [1]. Activity prediction relies solely on data features of trajectories of historical data, which may or may not include contextual information. Machine learning or Statistical techniques are applied to create predictions for future activities. 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 IoT(Internet of Things) and Big Data generated on the Internet, such as social network interactions and weather channels (e.g., Flickr, Facebook, Twitter, Foursquare), now it is possible to collect large volumes of mobility data of people, objects and animals, such as various types of vehicles, 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. A study proposes an algorithm named SemTraClus [6], which extracts revisited points, stay points, and user participation weights in different geographical areas.

To implement 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.

II. RELATED WORKS

Moving Object Data processing is emerging as an observable field of research. Various studies on mobility-based data cover various aspects of Big Data, including analysis of trajectory data, indexing, and retrieval. In this context, we project some specialized works in the field of extraction of Points of Interest.

In a study published in 2016 [7], patterns of human mobility are distinguished from space-time points saved on social networking sites. The outcome of this research is a semantically enriched dataset that opens up new possibilities for modeling human movement behavior.

We have also published a paper [5] proposing a Business Intelligence tool named "Predict-Move." This tool assesses the further customer movement from a Point of Interest (POI) to other business firms within large commercial establishments, enhancing customer services, and 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 centers in the management of transportation. 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, primarily focusing on directional variations. This has been tested with real trajectory data of oceanic fishing vessels, to detect the locations where vessels doing 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] developed a method for identifying the similarity of moving objects in a controlled path, using a combination of structural similarity and sequential mobility in trajectories. For managing road networks, they introduced an encoding technique.

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 separates and categorizes the semantic and contextual dimensions of traveling object data to provide concrete analysis.

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 Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The study analyzes and presents the advantages and disadvantages of each algorithm using the 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 measure its performance based on memory requirements, processing time, clustering quality, scalability, and stability.

This study includes comparisons with other available methods, concluding that BIRCH stands as the most suitable clustering method for handling large datasets. In this study, we tried to prove the efficiency of the clustering based on the BIRCH algorithm using the trajectory dataset of 21 users from the Geo-life dataset.

III. METHODOLOGY

- 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. The clustering algorithms DBSCAN, K-Means, and BRICH are applied in the selected dataset and generate clusters for each algorithm. Besides, the weightage participation (WP) of users at different locations is extracted and compared using the evaluation criteria. Based on the Silhouette score [17] of the algorithm, the efficiency and validity of the clustering methods can be measured.

- Data Description

In this study, we selected the GPS trajectory dataset from the dataset collected as part of the Geolife project at Microsoft Research Asia [18]. This trajectory dataset consists of trajectory details of 182 users over more than five years, ranging from April 2007 to August 2012. Each GPS trajectory has time-stamped points which include information regarding longitude, latitude, and altitude. The dataset consists of a total of 17,621 trajectories, covering a distance of 1,292,951 kilometers and with a cumulative duration of 50,176 hours. These trajectories were stored using GPS phones and GPS loggers, which constitute a wide range of sampling rates. The 91.5 percent of the trajectories feature dense representation, with data points recorded at intervals of 5 to 10 meters or every 1 to 5 seconds.

This dataset collects a wide spectrum of users' movement, comprehending everyday routines like regular work, going home, leisure, and sports activities such as dining, hiking, cycling, shopping, sightseeing, jogging, etc. Researchers can utilize this trajectory dataset in a lot of domains like user activity prediction, including mobility pattern mining, location-based social networks, location recommendation, and location privacy assignment.

- Application of DBSCAN in SemTraClus

Clustering is a common machine-learning technique for grouping similar data points based on their similarities or inferences. Clustering algorithms aim to partition data points into different clusters to discover hidden patterns and structures in the data. Here we applied the density-based clustering algorithm DBSCAN [19] to the selected user trajectories. Its advantage lies in its ability to discover clusters with arbitrary shapes and sizes. The algorithm considers 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 and MinPt

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. The 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.

- Application of k-means in SemTraClus

K-means clustering is considered one of the most widely used and straightforward clustering algorithms due to its efficiency and accuracy.

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 closer cluster center, often referred to as the 'within-cluster sum of squares' (WCSS). Here we applied the density-based clustering algorithm K-means to the selected user trajectories as that of the application in the DBSCAN [19].

The algorithm performs as follows:

- Initialization: From the dataset randomly select k initial centroids.
- Assignment: Forming k clusters by allocating each data point to the nearest centroid.
- Update: Reevaluate the centroid of each cluster as the mean of all data points assigned to it.
- Repeat steps 2 and 3 until either the centroids no longer change significantly or a maximum number of iterations is reached.

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

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: Apply clustering and create four clusters. The data points that do not belong to any of these clusters will be removed from the context area and considered as noise.
- Step 4: Highlights the data points in the different clusters.

Application of BRICH in SemTraClus

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm is a common choice and best suited for handling large datasets. Here we applied the clustering algorithm BIRCH to the selected user trajectories as that of the applications in the previous two applications. Various experiments have illustrated its efficiency in comparison to K-means and DBSCAN methods [20].

i. Initialization: Initialize an empty tree by specifying a clustering threshold and a maximum number of clusters
ii. Clustering: Starting from the root inserting each data point into the tree. Include the data point into the node without exceeding the threshold. If it exceeds the threshold, the node splits into two new sub clusters.
iii. Merging: The algorithm proceeds to merge sub clusters only if all data points have been inserted until the desired number of clusters is achieved.

Steps of BIRCH in SemTraClus

- Step 1: Preprocess of the dataset.
- Step 2: Apply the K-means algorithm for transforming the data points specifying criteria for four clusters, and setting a random state of fifteen.
- 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.

- Weightage Participation

Trajectory datasets provide valuable information about the movement of objects over time. These datasets are used in the applications of various fields like transportation and logistics, where tracking of object movements is critical. Weightage participation by users may be a valuable measure in the trajectory datasets, and the users may be assigned weights or importance to different features or attributes in the dataset. In this study, we focused on the concept of user weightage participation in the datasets concerned.
Steps for Calculating Weightage participation in SemTraClus Algorithm

The SemTraClus algorithm notifies Points of Interest (POI) from multiple trajectories and similar semantic locations are associated into clusters. Each cluster maintains a series of connected locations associated with various users, and creating sub-trajectories connecting interesting locations or semantic points. These cluster the semantic regions, where enrichment can be applied. Semantic tagging is achieved through a Point of Interest (POI) database, which stores and updates facilities, waypoints, landmarks, and other meaningful 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 interrelates with the degree of interest shown by different users in that region. A measure named "Weightage of Participation" (WP) is introduced for measuring the user's interests in the locations. WP identifies both the priority value of an individual trajectory a semantic region and the most relevant semantic regions in a particular geographical area.

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

A user's semantic trajectory comprises various cluster points during a travel session. The degree of a user's participation in a cluster is based on two parameters: Spatial Density (α) and Temporal Presence (β). The spatial density of a user trajectory Uj in a cluster Ci is specified 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, which shows a user's presence in the identified semantic region, which is derived by:

$\alpha(i,j)$ = (No.of trajectory locations visited by the user Uj in Ci) / (Total no. of locations in cluster Ci)

Temporal presence (β) measures the extent of a user's stay duration in a semantic region. 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, is expressed as:

$\beta(i,j)$ = (Stay time of user Uj in cluster Ci) / (Total time spent by all users in cluster Ci)

The WP (Weightage participation) of a user Uj in a cluster Ci shows a metric to identify the user's interest in that cluster. It calculated as averaged sum of Spatial Density (α) Temporal Presence (β).

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

IV. LOGICAL FRAMEWORK OF THE PROCESS INVOLVED

Main Framework Steps:

- Data Collection and Extraction of semantic points: collect the data and extract semantic points, including stay points, intersection points, and revisited points.
- Data Preprocessing: Preprocess the data by removing duplicates and null values for the implementation of various algorithms.
- Algorithm Selection and Implementation: Import various algorithms such as DBScan, K-Means, and BIRCH. Apply these algorithms to the dataset with a consistent number of clusters.
- Results Visualization: Visualize the results produced by each algorithm to identify the clusters and their characteristics.
- Cluster Accuracy Assessment: Evaluate the accuracy of the clusters using the Silhouette Score method.
- Comparison and Result Visualization: Compare the results from different algorithms and visualize the outcomes for a comprehensive analysis.

V. EXPERIMENTAL EVALUATION

- For the experiment, we have selected various trajectory tracks of 21 users which have 965 trajectories from the Microsoft Geolife trajectory dataset [18] and that have 1164069 trajectory points.
- The algorithm has been implemented in Python 3.10.2. All experiments are implemented in an eighth-generation Intel Core i5 computer machine with 8GB RAM.

A. Selected User-trajectory Details

We have extracted trajectory details for 21 users from the Geo-life dataset [18], as illustrated in Table 1

Table I

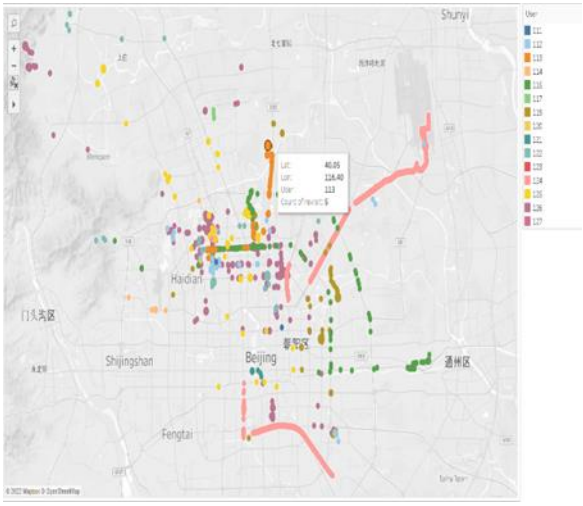
User	No.of trajectories
107	3
108	9
109	4
110	25
111	44
112	212
113	32
114	23
115	184
116	3
117	8
118	5
119	45
120	2
121	5
122	16
123	5
124	10
125	57
126	263
127	10
Total Trajectories	965

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

B. 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 Fig 1.

Fig 1. shows the revisited points of users in the trajectory dataset



C. 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.

Table II.

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

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

D. 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 Fig 2

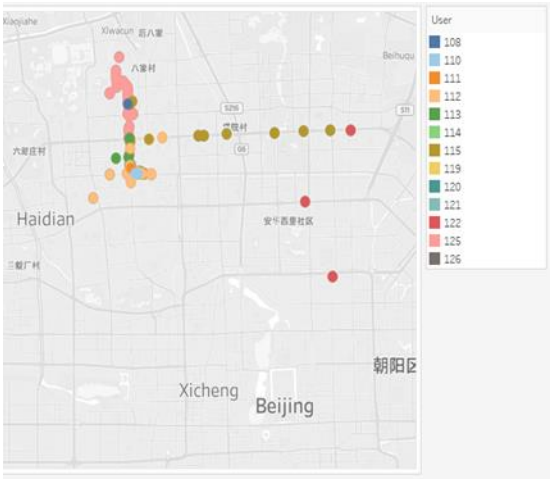


Figure 2 shows the intersection points of users in semantic region

E. 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 Fig 3.

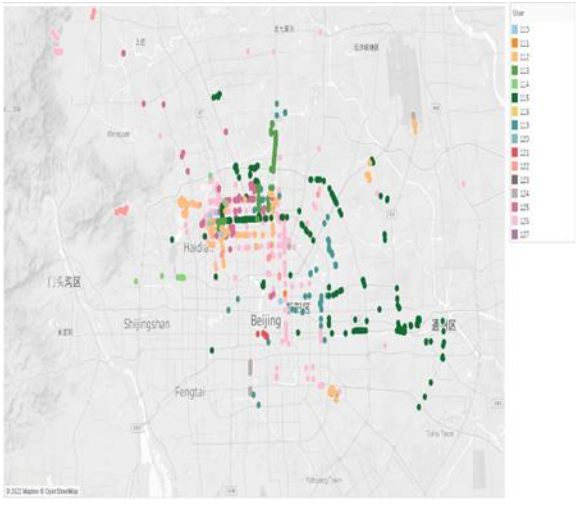


Fig.3 shows the stay points of users in semantic region

F. Semantic point extraction and density clustering

- The SemTraClus algorithm [6] intersecting points, revisited points and stay points, with the temporal and spatial threshold values 72 and 2 respectively.
- Our algorithm extracts 8523 semantic locations from the 1164069 trajectory points of 965 trajectories with 21 different users which is shown in Table III.

Table III.

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

G. DBSCAN-Cluster Details

We have applied the DBScan 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 IV.

TABLE IV.

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

H. 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 V.

Table V.

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	412	758	7318	46	8534

I. 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 VI

Table VI.

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

J. Comparative graphs of clusters

Fig. 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.

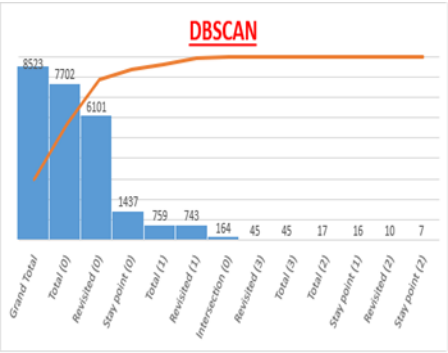


Fig.4

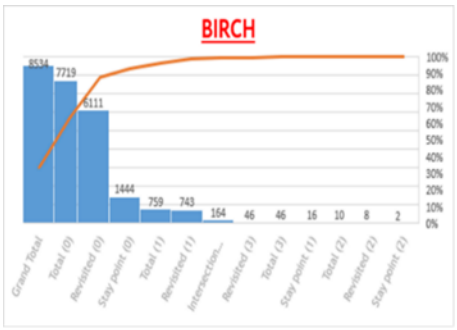
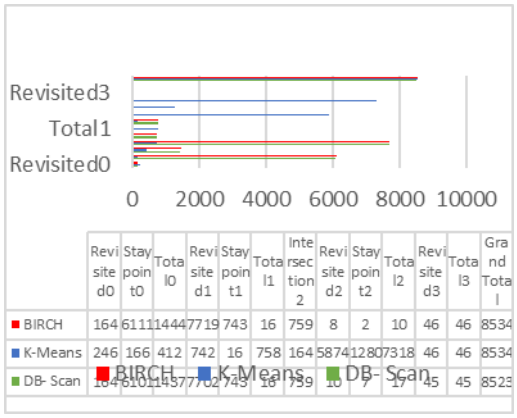


Fig.5



	Revisited0	Stay point0	Total0	Revisited1	Stay point1	Total1	Intersection2	Revisited2	Stay point2	Total2	Revisited3	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	184	6101	1457	7702	743	16	759	10	7	17	45	45	8523

Fig.6

K. 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

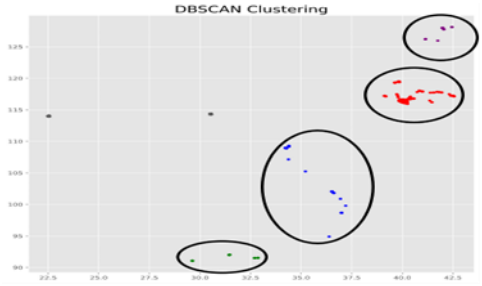


Fig 7

Fig7 shows the comparison chart as well as the comparison table for each of DBScan, K-Means and BIRCH algorithm

L. 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.

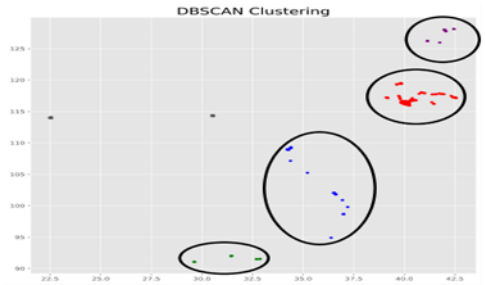


Fig.8

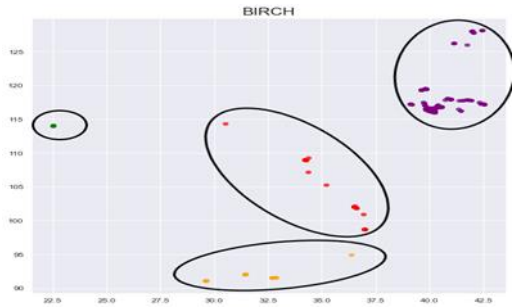


Fig.9

Fig.10

VI. WEIGHTAGE PARTICIPATION OF USERS

A. 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 formed using DBScan algorithm are shown in the table VII.

TABLE VII.

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 VII shows the weightage participation of 21 users after clustering using DBScan algorithm.

B. 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 VIII.

TABLE VIII.

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 VII shows the weightage participation of 21 users after clustering using K-Means algorithm

C. 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 IX.

TABLE IX.

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 IX shows the weightage participation of 21 users after clustering using BIRCH algorithm

D. 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 8, 9 and 10 and its comparison charts are shown in the figures 11 and 12. 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 8, 9 and 10 and its comparison charts are shown in the figures 11 and 12

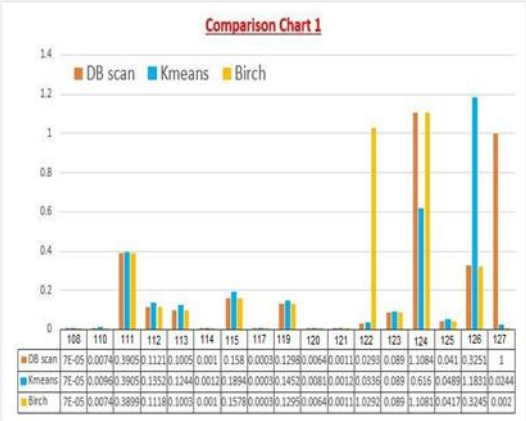


FIG.11

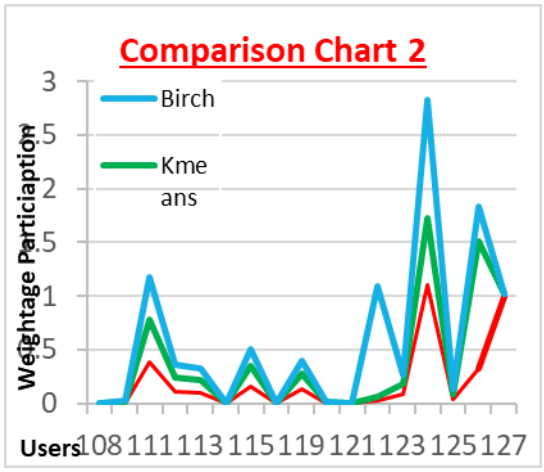


FIG.12

Figure 11 and 12 shows the comparison between the three algorithms DBScan, K-Means and BIRCH for the selected users.

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 has more weightage of participation when compared to the clusters formed using, DBScan and K-Means. This can also be observed when comparing table 8, 9 and 10 and it is also prominent in figures 11 and 12.

VII. COMPARISON OF VARIOUS CLUSTERING ALGORITHMS USING SILHOUETTE

Silhouette constitutes a method of validation and description of consistency in the clusters of data.

a) Silhouette score or Silhouette Coefficient is a metric used to measure the integrity of a clustering technique.
b) The technique provides a score representation of well classification of each object.
c) Its measured value ranges between -1 and 1.

- The score 1 depicts that clusters are well apart and distinguished.
- The score 0 depicts that clusters are indifferent or the distance between them is not significant.

- The score 1 implies that clusters are assigned improperly.

TABLE X

Algorithms	Silhouette score
DB-Scan	0.949
BIRCH	0.962
K-MEANS	0.926

- When comparing the Silhouette score of DB-Scan, K-Means and BIRCH we can clearly see that BIRCH outperforms DB-Scan and K-means in the SemTraClus Algorithm
- The Silhouette score of the algorithms BIRCH, DBScan and K-Means are calculated, and the results are shown in table X and the same results are visualized in figure 13.
- The BIRCH Algorithm gives better and more accurate results with the silhouette score 0.962 which is higher than that of DBScan(0.949) and K-MEANS(0.926) and weightage participation result is also prominent in BRICH.

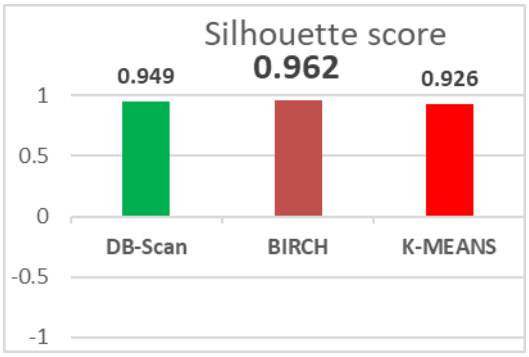


Fig.13

The two clustering algorithms, BRICH and K-Means were used to compare with the DBSCAN algorithm used in SemTraClus. The results showed that the BIRCH clustering algorithm has advantages over the other two algorithms because its clustering results were more reasonable and effective, and the Silhouette score values were higher and more stable. When dealing with user trajectories with similar spatial distribution characteristics, the BIRCH clustering algorithm can still distinguish the subtle differences and thus obtain better clustering results.

CONCLUSION

Clustering is the prominent method for categorizing the semantic regions concerning the attributes of the data items. In this study, we compare the clustering characteristics of three algorithms DBSCAN, K-Means, and BRICH using the trajectory data of selected 21 users from Microsoft's geo-life trajectory dataset. The efficiency and validity of the clustering methods are evaluated with the weightage participation of the users and the silhouette score. The results depict that the BIRCH clustering algorithm has advantages over the other two algorithms. Its clustering results were reasonable and effective, and the Silhouette score values were stable and higher. Based on the evaluations done, it is proved that the BRICH clustering algorithm can be considered for clustering-based applications in trajectory data processing.

REFERENCES

[1] Li Xu , Mei-Po Kwan ReferencesMining sequential activity–travel patterns for individual-level human activity prediction using Bayesian networks

[2] Carlos Andres Ferrero. Luis Otavio Alvares, Vania Bogorny Multiple Aspect Trajectory Data Analysis: Research Challenges and Opportunities

[3] L. O. Alvares, V. Bogorny, B. Kuijpers, J. Macedo, B. Moelans, and A. Vaisman. A Model for Enriching Trajectories with Semantic Geographical Information. In GIS, page 22, 2007.

[4] Ronaldo dos Santos Mello , Vania Bogorny , Luis Otavio Alvares, Luiz H. Z. Santana , Carlos Andres Ferrero , Angelo Augusto Frozza1, Geomar Andre Schreiner, Chiara Renso MSTER: A Multiple Aspect View on Trajectories

[5] V S Praveen Kumar, Sajimon Abraham, Nishanth 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 & Sajimon Abraham (2019): SemTraClus: an algorithm for clustering and prioritizing semantic regions of spatio-temporal trajectories, International Journal of Computers and Applications.

[7] Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science, 30(5), 881-906.

[8] Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F.,& Vangenot, C. (2008). A conceptual view on trajectories. Data & knowledge engineering, 65(1), 126-146

[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," 2010 5th IEEE International Conference Intelligent Systems, London, UK, 2010, pp. 114-119, doi: 10.1109/IS.2010.5548396.

[10] Beber MA, Ferrero CA, Fileto R, et al. Individual and group activity recognition in moving object trajectories. J Inf Data Manage. 2017;8(1):50.

[11] Abraham S, Lal PS. Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations. Transp Res Part C: Emerg Technol. 2012;23:109–123.

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

[13] Khan, Kamran, et al. "DBSCAN: Past, present and future." The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, 2014.

[14] Liu, Qi, et al. "Differentially private and utility-aware publication of trajectory data." Expert Systems with Applications 180 (2021): 115120.

[15] Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: A New Data Clustering Algorithm and Its Applications. Data Mining and Knowledge Discovery 1, 141–182 (1997).

[16] https://doi.org/10.1023/A:1009783824328

[17] https://www.microsoft.com/en-us/download/details.aspx?id=52367

[18] https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished.

[19] https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/

[20] D. Hsu and S. Johnson, "A Vibrating Method Based Cluster Reducing Strategy", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 2, (2008), pp. 376-379.

[21] 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, 2016, pp. 1-7, doi: 10.1109/ICDSE.2016.7823946.

[22] https://www.sciencedirect.com/science/article/pii/S0169023X06000218?via%3Dihub

[23] https://en.wikipedia.org/wiki/DBSCAN

[24] https://en.wikipedia.org/wiki/K-means_clustering