International Journal of Information Retrieval Research
Volume 12 • Issue 1

Cluster-Based Cab Recommender System (CBCRS) for Solo Cab Drivers

Supreet Kaur Mann, Panjab University, India*
Sonal Chawla, Panjab University, India

ABSTRACT

An efficient cluster-based cab recommender system (CBCRS) provides solo cab drivers with recommendations about the next pickup location having high passenger finding potential at the shortest distance. To recommend the cab drivers with the next passenger location, it becomes imperative to cluster the global positioning system (GPS) coordinates of various pick-up locations of the geographic region as that of the cab. Clustering is the unsupervised data science that groups similar objects into a cluster. Therefore, the objectives of the research paper are fourfold: Firstly, the research paper identifies various clustering techniques to cluster GPS coordinates. Secondly, to design and develop an efficient algorithm to cluster GPS coordinates for CBCRS. Thirdly, the research paper evaluates the proposed algorithm using standard datasets over silhouette coefficient and Calinski-Harabasz index. Finally, the paper concludes and analyses the results of the proposed algorithm to find out the most optimal clustering technique for clustering GPS coordinates assisting cab recommender system.

KEYWORDS

Calinski-Harabasz Index, Clustering Techniques, Density-Based Clustering, Hierarchical Clustering, Partition Based Clustering, Recommender System, Silhouette Coefficient, Unsupervised Machine Learning

INTRODUCTION

Recommender systems are the software tools that recommend the user with a set of personalized suggestions which can be useful to the user. These suggestions help the user with the decision-making process (Ricci et al., 2011). Recommender systems are of much importance in cab services too. Recommender system for cab drivers has always been a major concern as cabs are the main source of transportation in the modern cities compared with the other transportation services like bus, train etc. A recommender system can be constructed using three approaches: Content-Based Filtering (Mooney & Roy, 1999), Collaborative-Based Filtering (Resnick & Varian, 1997) and Hybrid Filtering (Pazzani, 1999). Content-Based Filtering is based on the user's historical information and hence faces the Cold Start problem. Collaborative-Based Filtering makes an automatic recommendation to a user based on the taste and likings of several other users. Hybrid Filtering combines both of the filtering methods. Cab Recommender system uses Collaborative Filtering. Cab Recommender systems are useful to both the driver and the users (Yuan et al., 2013). It recommends the cab driver with the nearest passenger finding locations from where passengers can be found at a minimum travelling distance and thereby increasing their profit. It also helps the passenger to find a cab near them to save time (Wang et al.,

DOI: 10.4018/IJIRR.314604 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

2017). To recommend cab drivers with the next passenger finding location, it is essential to cluster the pickup Geolocations of the same area as that of the cab. There are several clustering techniques to cluster these geolocations. Clustering techniques can be broadly classified into three categories: Hierarchical Methods, Partition-Based Methods and Density-Based Methods (Wang et al., 2017).

Clustering evaluation can be performed either using extrinsic measure or using intrinsic measures. To evaluate a cluster using an extrinsic measure like adjusted rand index, Fowlkes-Mallows score etc. it is essential to have ground truth labels. Since, cab dataset does not have the ground truth labels so extrinsic measures cannot be used to evaluate the cluster performance for the Cab Recommender system. Whereas, to evaluate cluster performance using an intrinsic measure like Silhouette Coefficient (Peter, 1987), Calinski-Harabasz index, Davies-Bouldin Index etc., it is not required to have ground truth labels. Hence, Intrinsic measures such as Silhouette Coefficient, Calinski-Harabasz Index etc can be used to evaluate the cluster performance for Cab Recommender System. For this research paper, Silhouette Coefficient and Calinski-Harabasz Index are used to evaluate the cluster performance for Geolocation clusters as these score works fine with large datasets and are computationally faster than other intrinsic measures (Pedregosa et al., 2011).

Since clustering of geolocations is an unsupervised learning hence it becomes difficult to adopt a standard clustering technique that shall cluster the passenger pickup geolocations. Hence, there is a need for a framework to cluster the passenger geolocations which shall assist the cab recommender system to recommend most near passenger finding locations to the cab drivers.

Therefore, the research paper aims to design and develop an algorithm to generate clusters of passenger pickup geolocations using Hierarchical Clustering such as Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) and Clustering Using REpresentative (CURE), Partition-Based Clustering such as K-Means, Mini Batch K-Means and Spectral Clustering (Ng et al., 2002) and Density-Based Clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points To Identify Cluster Structure (OPTICS) for CBCRS. The proposed algorithm is rigorously evaluated over three unsupervised datasets of New York, Porto and Mexico Cities over the parameters of Silhouette Coefficient and Calinski-Harabasz Index.

BACKGROUND

Clustering techniques can be broadly classified into 3 categories: Hierarchical Clustering, Partition-Based Clustering and Density-Based Clustering. Hierarchical Methods (Murtagh, 1983; Bade & Nurnberger, 2006) treat each data point as a cluster and then join together these clusters to form larger clusters (Manning et al., 2009). Hierarchical Clustering either follows a bottom-up approach (agglomerative) or a top-down approach (divisive). The agglomerative approach of clustering is majorly used for recommender systems. Partition-Based Clustering divides the ‘n’ data points into ‘k’ pre-defined number of clusters, where $k <= n$ (Mazimpaka & Timpf, 2016). Each cluster has a centroid which represents every data point of that cluster. Density-Based methods divide the data points into clusters based on density (Ester et al., 1996). Densely populated data points form a cluster whereas sparsely populated data points may be treated as noise or outliers. If the number of data points in a specified diameter is greater than the minimum number of data points required to form a cluster then they form a cluster else they are treated as outliers.

Cluster Evaluation for unsupervised machine learning is performed using Silhouette Score and Calinski-Harabasz Index. Silhouette Coefficient is the metric to determine the similarity of an object to the cluster it belongs to. The value of the Silhouette Coefficient ranges from -1 to +1. The best value for SC is +1 which indicates that the object is correctly placed in the cluster. The value of 0 indicated that the object lies between the boundary of two clusters. The negative value indicates that the object is far away from its cluster. Silhouette Coefficient is calculated using the mean intra-cluster distance ‘a’ and the mean nearest cluster distance ‘b’ as given in equation(1) (Shutaywi & Kachouie, 2021)

$$SC = \left(b - a\right) / \max\left(a, b\right) \tag{1}$$

Calinski-Harabasz Index also called a variance ratio criterion is a ratio between the sum of cluster dispersion and the sum of inter-cluster dispersion. Calinski-Harabasz is described by equation (2) (Cengizler & Ün, 2017).

$$CH\left(k\right) = \frac{\frac{B_c\left(\mathrm{k}\right)}{\left(k-1\right)}}{\frac{W_c\left(\mathrm{k}\right)}{\left(n-1\right)}} \cdots \tag{2}$$

where n is the number of clusters and k is the number of class. B_c and W_c denotes between and cluster sum of squares respectively which are given as equation (3) and (4).

$$B_c = \sum\nolimits_{k=1}^{K} \left|C_k\right| \overline{C_k} - \overline{x}^2 \cdots \tag{3}$$

$$W_c = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{k,i}\, x_i - \overline{C}_k^{\,2}$$

An extensive background study was carried out to obtain in-depth knowledge about the cab recommender systems. Many researchers have used different methods to implement the recommender system like passenger mobility pattern (Yuan et al., 2013), cab driver pickup locations (Mittal, 2016), road network data (Deep et al., 2015), the spatiotemporal distribution of cab passenger demands or the cab GPS trajectory. In order to cluster the cab trajectories, the researcher have used various clustering algorithms such as OPTICS (Density-Based) (Ankerst et at., 1999; Braga et al., 2012; Yuan et al., 2013), K-Means (Partition-Based) (Hu et al., 2012; Hartigan & Wong, 1979, Wang et al., 2017), DBSCAN (Density-Based) (Ester et al., 1996; Shen et al., 2015; Szenasi, 2014; Wu et al., 2017), Mini Batch K-Means (Partition-Based) (Sculley, 2010) etc. algorithms.

RESEARCH METHODOLOGY

From the literature review, it can be inferred that though there exist several clustering algorithms yet they do not give satisfactory results when applied over an unsupervised dataset of geolocations for parameters of Silhouette Coefficient and Calinski-Harabasz score. Hence, there is a need for an algorithm to cluster the cab pickup geolocations which shall assist the solo cab drivers with the nearest next pickup location and shall work as a Cab Recommender System. Therefore, the objective of the research paper is to design and develop an algorithm to generate clusters of Cab pickup Geolocations using standard datasets concerning the Cab Recommender System based on an extensive literature review conducted.

The research methodology for this paper followed a two-phase approach.

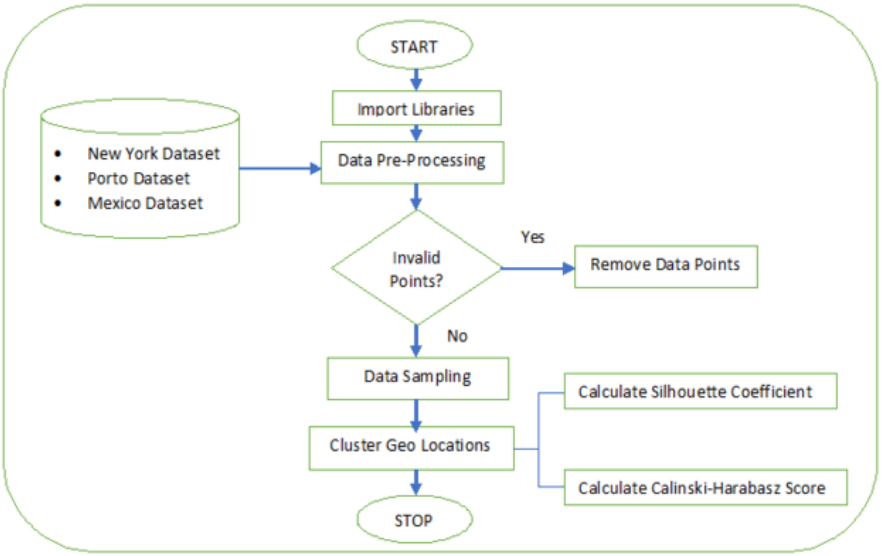
Phase I involved the design and development of an algorithm to generate clusters using Hierarchical Clustering, Partition-Based Clustering and Density-Based Clustering of passenger pickup geolocations for CBCRS.

Phase II involved the evaluation of the clusters formed by the algorithm over three standard unsupervised cab datasets of New York, Mexico and Porto cities using the parameters of Silhouette Coefficient and Calinski-Harabasz Index. The results were obtained and analysed to find an optimal clustering technique for Cab Recommender Systems.

Phase I: Design and Development of a Framework to Generate Clusters of GPS Coordinates

In this phase, the proposed algorithm was developed using Python 3.7 to cluster the passenger pickup geolocations of the dataset. The flowchart in Figure 1 depicts the sequential execution to cluster passenger pickup geolocations.

Figure 1. Flowchart for proposed Algorithm



Import Libraries

Various open-source libraries were imported from the python library to perform clustering like sklearn, NumPy, pandas, matplotlib. Sklearn is a library that includes clustering algorithms. NumPy is a library that is used to handle multidimensional arrays in python and it includes various functions of arrays. Pandas is a package used for data analysis and machine learning tasks. Matplotlib is a library function used to plot graphs in python as in MATLAB.

Data Pre-Processing

The datasets were pre-processed by removing null values if any, missing values, invalid data points of City, Negative fare or zero passenger entries.

Data Sampling

Since, it was difficult to plot large amount of data therefore the datasets were sampled as below.

New York Dataset: The 2014 Yellow Taxi Trip data (NYC Open data, 2014) includes the taxi trips of Yellow Taxi in New York city which were completed. This dataset consists of 19 attributes. This is a publicly available dataset at NYC Open Data. It contains taxi trips from 1st January 2014 to 31st December 2014. There are nearly 165 million data records in the dataset. The attributes used for this research study are vendor_id, pickup_datetime, dropoff_datetime, passenger, trip_distance, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude and total_amount. The dataset was divided into 12 files for each month of the year. Since March is the best time to visit New York and the number of entries in March month were maximum so the data was plotted for the month of March.

Porto GPS Taxi Dataset: The dataset includes the taxi data for Porto city from 1st July 2013 to 30th June 2014 of 442 taxis running over the city. It contains 9 attributes. This is a publicly available dataset at Kaggle (Cross, Chris. 2017). It consists of 1.7 million entries for records. It contains a polyline which is a string of GPS coordinates taken from the cab after every 15 seconds. To cluster the pickup and drop-off locations, the pickup latitude and pickup longitude were extracted from the polyline. The timestamp in the dataset is Unix timestamp converted into Windows timestamp in the format "Year – Mon- Date Hour: Min: Sec". Since it was difficult to plot so many data points. So, the dataset was divided into 12 files for each month of a year. As May month us the best time to visit Porto and the number of bookings for the cab is maximum in May, so clustering was performed for the month of May.

Mexico Dataset: The Taxi Route of Mexico City includes the taxi routes for Mexico City from 1st June 2016 to 20th July 2017. It contains nearly 12 thousand records. This is publicly available data that can be downloaded from Kaggle (Navas, Mario, 2017). It consists of 12 attributes. The dataset was divided into 7 files one for each day of the week. Each file was again divided into 12 files based on the time from 0 to 11 hours of the day dividing into a total of 84 files.

Clustering

Clusters of the sampled dataset were formed. The clustering techniques used for this research paper are Hierarchical Clustering (BIRCH and CURE), Partition-Based Clustering (K-Means, Mini Batch K-Means and Spectral Clustering) and Density-Based Clustering Techniques (DBSCAN and OPTICS).

Cluster Evaluation

To evaluate the clusters resulting from the proposed framework, the Silhouette Coefficient and Calinski-Harabasz Index were calculated for each file. Average Silhouette Coefficient and average Calinski-Harabasz Index were calculated for each clustering technique and analysed.

Phase II: Evaluation of Clusters Formed Using Proposed Framework

The evaluation of clusters formed by the algorithm was done using two different parameters: The Silhouette Coefficient and Calinski-Harabasz Index. Silhouette Score determines the mean Silhouette Coefficient of each sample. Its value lies between 1 and -1. More the positive value better is the cluster. The value near 0 indicates overlapping clusters. The negative values indicated that the value is assigned to the wrong cluster. The higher the value of the Calinski-Harabasz Score depicts better cluster performance.

On the basis of flowchart in Figure 1, the framework as depicted in Figure 2 was proposed for the research to cluster passenger pickup geolocations.

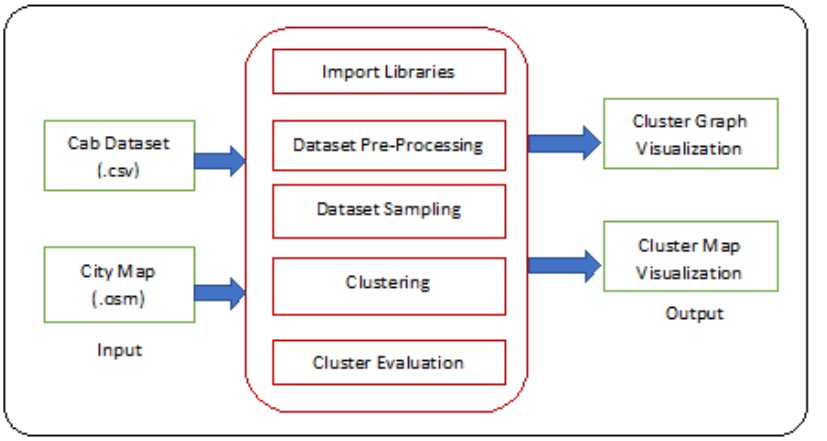
Input: The proposed algorithm accepts two files as input files: Cab Dataset (.csv) and City Map (.osm). For the research paper three cab datasets were used: the New York dataset, Mexico dataset and Porto GPS taxi dataset.

Output: Each clustering algorithm provided the output using two output files: Cluster-Graph visualization and Cluster Map Visualization. Cluster Graph Visualization file contains the data points and clusters being marked on Graphs with Lalitutde on X-Axis and longitude on Y-Axis. Cluster Map Visualization file contains the clusters and data points being marked on the city map of that particular latitude and longitude.

Based on above framework in Figure 2, the algorithm was framed as Algorithm 1.

Algorithm 1: Proposed Framework to Cluster passenger pickup Geolocations
Input: Dataset (New York Dataset, Porto GPS Taxi Dataset, Mexico Dataset) and City Map (New York, Porto, Mexico)
Steps:
1. *Import Libraries:* Import python libraries for clustering such as NumPy, pandas, sklearn, matplotlib etc.
2. *Dataset Pre-Processing:* Remove the null values, missing

Figure 2. Proposed Framework to Cluster Geolocations



values, Zero passenger entries, Negative fare entries and Out Bound Entries. The entries which do not belong to a specific city are removed.

3. *Dataset Sampling:* Since the dataset consist s of huge data points, hence it was essential to sample the dataset. The datasets were sampled and resulting files were formed as explained before.

a. *New York Dataset*: The dataset was divided into monthly data. The dataset was divided into 7 days of the week and 24 hours of a day. Resulting in 168 files.

b. *Porto GPS Taxi Dataset*: The dataset was divided into 12 months of data. The dataset was divided into 168 files for each weekday and each hour of the day.

c. *Mexico Dataset*: The dataset was divided into 12 files for each hour resulting in 84 files.

4. *Clustering:* Over the sampled dataset the clustering techniques were implemented. The clustering techniques used were:

- *Hierarchical Clustering*: BIRCH and CURE.
- *Centroid Based Clustering*: K-Means, Mini Batch K-Means Clustering and Spectral Clustering.
- *Density-Based Clustering*: DBSCAN and OPTICS

5. *Cluster Evaluation:* To evaluate the cluster resulting from the clustering techniques, the Silhouette Coefficient and Calinski-Harabasz Index were calculated for each file.

Output: Cluster-Graph visualization and Cluster Map Visualization.

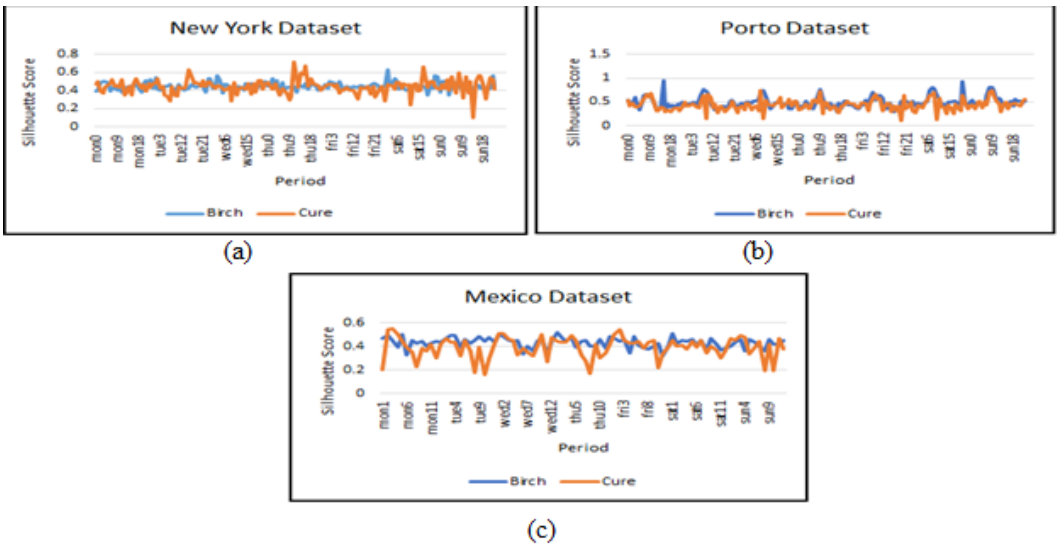
EXPERIMENTAL RESULTS

The proposed algorithm was implemented using Python 3.7, the following graphs were obtained and plotted for the various Silhouette Coefficient and Calinski-Harabasz Index of clusters formed using the different clustering techniques over the standard datasets. Each graph depicts the time period on the X-axis and the value of the Silhouette Coefficient and Calinski-Harabasz on the Y-axis.

Hierarchical Clustering

As seen in Figure 3(a), BIRCH clusters were robust for three periods Saturday 02:00 AM, Saturday 10:00 PM and Sunday 10:00 PM. Other BIRCH clusters were reasonable with the Silhouette Coefficient which varied between 0.35 and 0.62. Whereas the CURE clusters had a Silhouette Coefficient between 0.09 and 0.71. As depicted in Figure 3(b), it could be inferred that the values of the Silhouette Coefficient are more or less the same for BIRCH and CURE. However, the value for CURE lies between 0.73 and 0.10 and in the case of BIRCH, it lies between 0.95 and 0.29. As evident in Figure 3(c), the CURE algorithm has much lower values of the Silhouette Coefficient than BIRCH. The Silhouette Coefficient of CURE varies between 0.55 and 0.16 whereas the Silhouette Coefficient of BIRCH varies between 0.51 and 0.32.

Figure 3. Silhouette Coefficient for Hierarchical Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset



As seen in Figure 4(a), BIRCH clusters have shown peak values on Friday 09:00 PM and Sunday 06:00 PM. The values of the Calinski-Harabasz index lies between 3426.481 and 95.07167 for BIRCH and 1223.007 and 3.429926 for CURE for New York Dataset. As shown in Figure 4(b), BIRCH outperforms CURE. BIRCH has peak values on Saturday 10:00 AM and Sunday 09:00 AM. The range of Calinski-Harabasz lies between 6190.331 and 224.6094 for BIRCH and 3025.595 and 11.91751 for CURE. Whereas, as shown in Figure 4(c), though CURE has a maximum value at Monday 04:00 AM but the index value of both BIRCH and CURE are on the same scale with BIRCH having values between 339.7862 and 39.50915 and CURE having a value between 490.2045 and 16.95101.

Higher Value of average Silhouette Coefficient and average Calinski-Harabasz Index depicts a better cluster performance. As shown in Table 1, the average Silhouette Coefficient and average Calinski-Harabasz Index results of BIRCH shows better results than CURE for clusters of passenger pickup Geolocations over the unsupervised standard cab datasets of New York, Porto and Mexico when the number of desired clusters is set to 5 for CURE. The value of Average Silhouette Score for BIRCH is 0.448647, 0.494154 and 0.43333 as compared to average Silhouette Coefficient for CURE as 0.443054, 0.436439 and 0.39279 respectively for three cities. Average Calinski-Harabasz Index for BIRCH is 1347.968, 1030.36 and 129.1743 as compared to average Calinski-Harabasz Index for CURE as 751.3244, 533.5582 and 103.4119 respectively.

Figure 4. Calinski-Harabasz Score for Hierarchical Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset

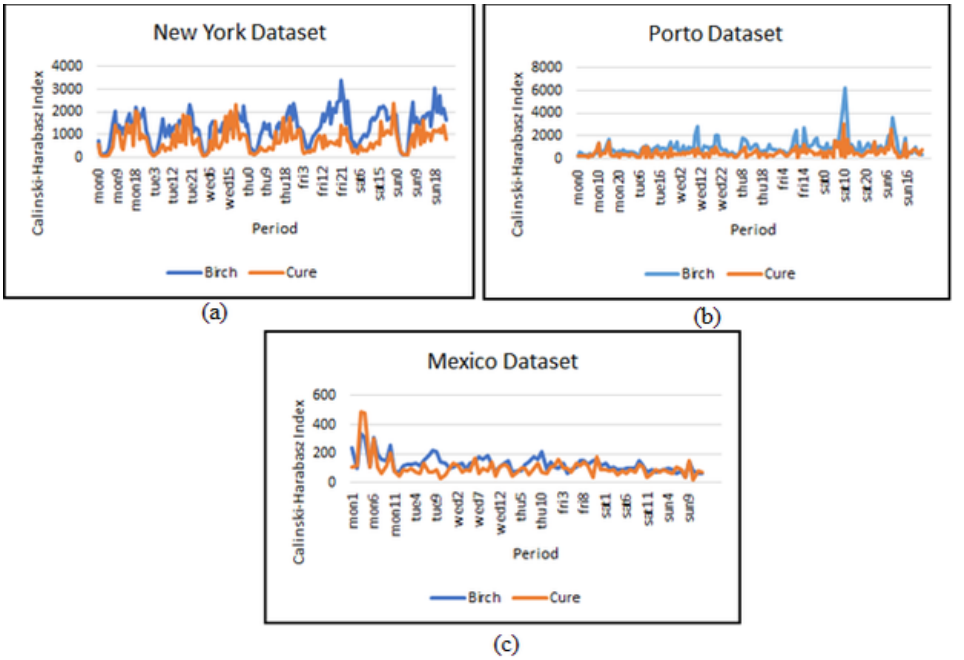


Table 1. Average Silhouette Coefficient and Average Calinski-Harabasz Index for Hierarchical Clustering

	Average Silhouette Score		Average Calinski-Harabasz Index	
	Birch	**CURE**	**Birch**	**CURE**
New York Dataset	0.448647	0.443054	1347.968	751.3244
Porto Dataset	0.494154	0.436439	1030.26	533.5582
Mexico Dataset	0.43333	0.39279	129.1743	103.4119

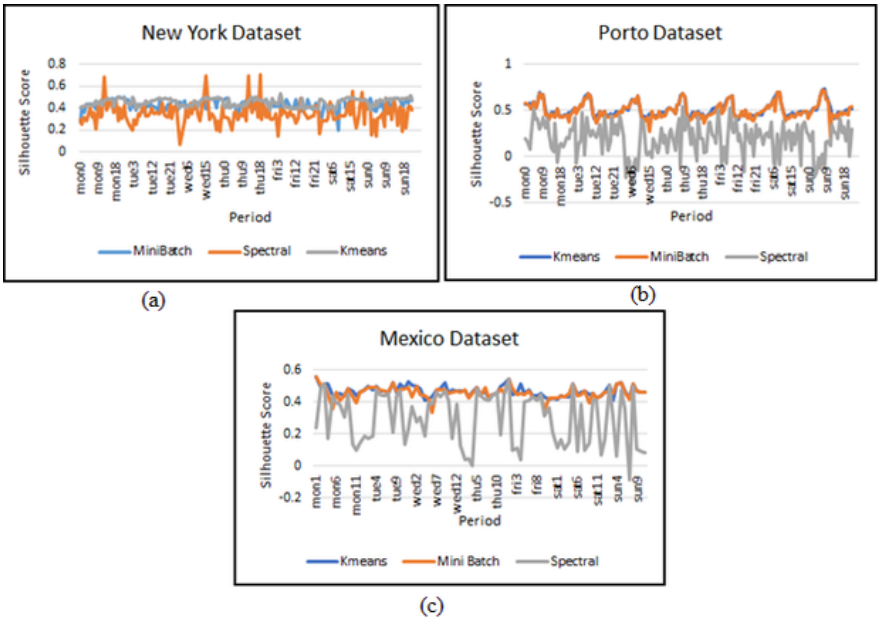
Partition-Based Clustering

In the case of Partition-Based Clustering, the algorithms used are K-Means, Mini Batch K-Means and Spectral Clustering. Since the results of Partition-Based Clustering algorithms depends on the number of clusters, so the number of clusters chosen for the research were 4 for all three datasets. Using the Elbow-Plot method the number of clusters was fixed to 4 for New York Dataset, Porto Dataset and Mexico Dataset.

Silhouette Coefficient was calculated for Partition Based clustering techniques: K-Means, Mini Batch and Spectral over three datasets. As depicted in Figure 5(a), it can be seen that Mini Batch clusters were robust for three periods Monday 09:00 PM, Wednesday 05:00 PM and Sunday 05:00 PM. Other Mini Batch clusters were reasonable with a K-Means coefficient. Whereas the Silhouette Coefficient for Spectral Clustering is low as compared to K-Means and Mini Batch clustering.

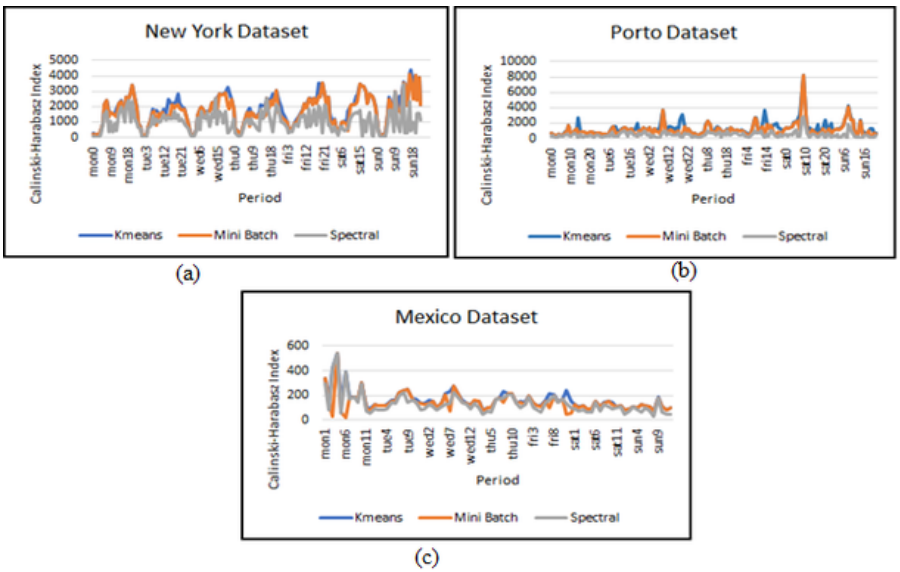
The Silhouette Coefficient value lies between 0.52 and 0.38 for K-Means, 0.52 and 0.19 for Mini Batch K-Means and 0.33 and 0.02 for Spectral Clustering. As evident in Figure 5(b) and Figure 5(c), it can be seen that the values of the Silhouette Coefficient are more or less the same for Mini Batch K-Means and K-Means whereas the value is too low for Spectral Clustering.

Figure 5. Silhouette Score for Partition Based Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset



As seen in Figure 6(a), K-Means clusters have shown peak values on Sunday 06:00 PM and Sunday 05:00 PM for New York Dataset. The values of the Calinski-Harabasz index lies between 4372.347 and 145.095 for K-Means, 4090.89 and 119.3877 for Mini Batch K-Means and 3503.092 and 33.36718 for Spectral Clustering in the case of New York Dataset. As shown in Figure 6(b), K-Means has peak values on Saturday at 10:00 AM for Porto Dataset. The range of Calinski-Harabasz lies between 8264.81 and 347.7258 for K-Means, 8239.594 and 127.3435 for Mini Batch K-Means

Figure 6. Calinski-Harabasz Score for Partition Based Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset



and 2871.384 and 26.2621 for Spectral Clustering in the case of Porto Dataset. Whereas, as shown in Figure 6(c), all three-clustering algorithm performs on the same scale giving the highest peak at Monday 04:00 AM for Mexico Dataset. The values of the Calinski-Harabasz index lies between 541.6556 and 18.865 for K-Means, 541.6556 and 18.13189 for Mini Batch K-Means and 541.6556 and 30.4384 for Spectral Clustering in the case of Mexico Dataset.

The average Silhouette Coefficient and average Calinski-Harabasz Index of K-Means shows better results over Mini Batch K-Means and Spectral Clustering for clustering passenger pickup Geolocations when the number of clusters is set to 4 as per the Elbow-Plot for Partition-Based Clustering over standard unsupervised cab datasets of New-York, Porto and Mexico as shown in Table 2.

Table 2. Average Silhouette Coefficient and Average Calinski-Harabasz Index for Partition-Based Clustering

	Average Silhouette Coefficient			Average Calinski-Harabasz Index		
	K-Means	**Mini Batch K-Means**	**Spectral Clustering**	**K-Means**	**Mini Batch K-Means**	**Spectral Clustering**
New York Dataset	0.4524455	0.434438	0.349337	1869.752	1704.474	1002.064
Porto Dataset	0.5106	0.496789	0.19274	1362.771	1110.594	502.079
Mexico Dataset	0.468929	0.458942	0.295171	162.0549	140.7774	130.515

The average Silhouette Coefficient of K-Means for three cities are 0.4524455, 0.5106 and 0.468929 as compared to the average Silhouette Coefficient of Mini Batch K-Means as 0.434438, 0.496789 and 0.458942 and of Spectral Clustering as 0.349337, 0.19274 and 0.295171 respectively.

The average Calinski-Harabasz Index of K-Means for three cities is higher with values as 1869.752, 1362.771 and 162.0549 as compared to average Calinski-Harabasz Index of Mini Batch K-Means as 1704.474, 1110.594 and 140.7774 and for Spectral Clustering as 1002.064, 502.079 and 130.515 respectively.

Density-Based Clustering

In the case of Density-Based Clustering, the algorithms that were implemented are DBSCAN and OPTICS with a minimum cluster size of 4. ‘

As seen in Figure 7(a), the Silhouette Coefficient values for DBSCAN over New York Dataset lies between 0.661 and 0.043 and for OPTICS the values lie between 0.661 and 0.054. As shown in Figure 7(b), it can be seen that the values of the Silhouette Coefficient are higher for DBSCAN than OPTICS over Porto Dataset. The Silhouette Coefficient values for DBSCAN lies between 0.89 and 0.05 and for OPTICS lies between 0.52 and -0.02. As evident in Figure 7(c), both DBSCAN and OPTICS perform more or less the same over the Mexico dataset.

As seen in Figure 8(a), DBSCAN and OPTICS clusters have shown peak values on Sunday 06:00 PM and Thursday 10:00 PM for New York Dataset. The values of the Calinski-Harabasz index lies between 1223.007 and 5.6075118 for DBSCAN and OPTICS over New York Dataset. As shown in Figure 8(b), OPTICS has peak values on Saturday at 10:00 AM for Porto Dataset. The range of Calinski-Harabasz lies between 500.8506 and 0.715458 and for DBSCAN the value lies between 801.7392 and 4.911969 for OPTICS in the case of Porto Dataset. Whereas, as shown in Figure 8(c), The values of the Calinski-Harabasz index lies between 25.25027 and 1.6078 for DBSCAN, 24.61491 and 1.59422 for OPTICS in the case of Mexico Dataset.

The average Silhouette Coefficient and average Calinski-Harabasz Index of DBSCAN shows better results over OPTICS for clustering passenger pickup Geolocations when the minimum number

Figure 7. Silhouette Score for Density Based Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset

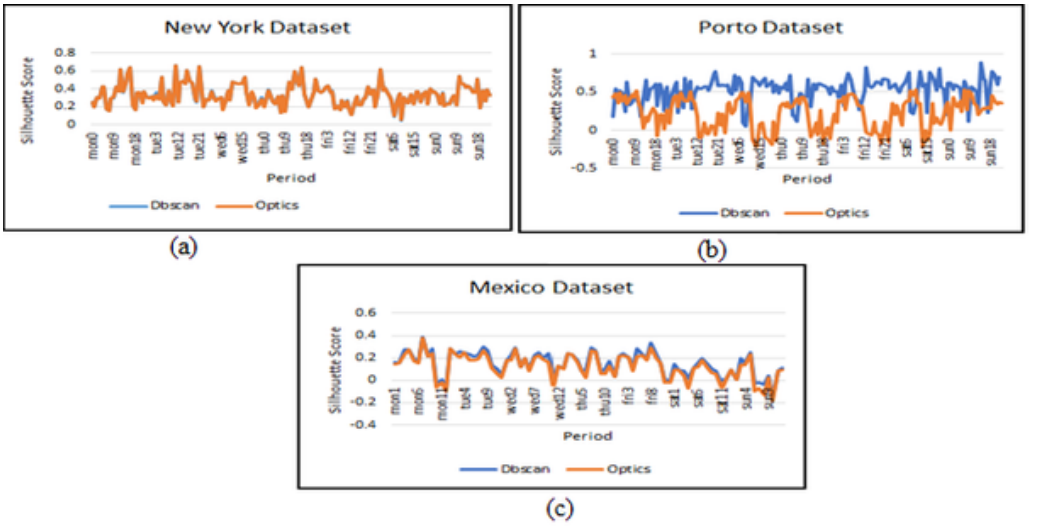


Figure 8. Calinski-Harabasz Score for Density Based Clustering a) New York Dataset b) Porto Dataset c) Mexico Dataset

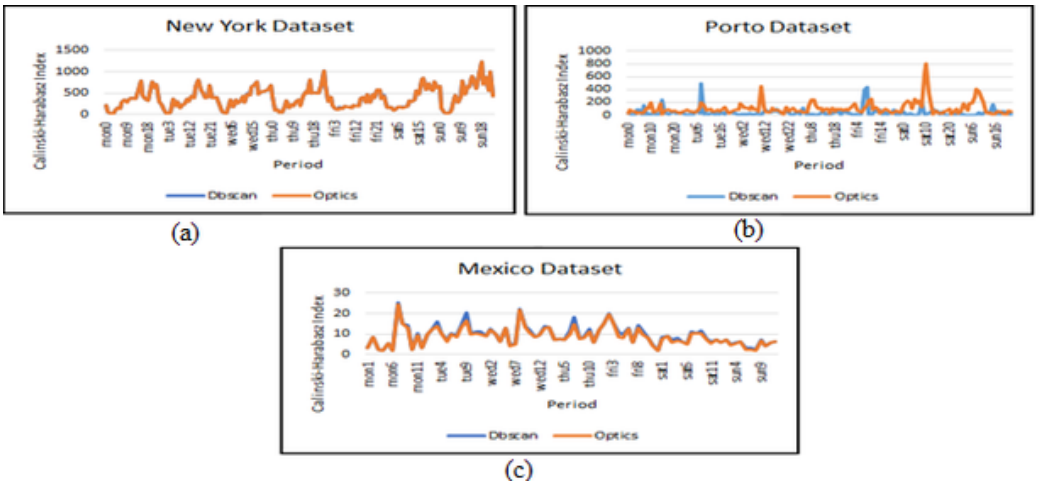


Table 3. Average Silhouette Coefficient and Average Calinski-Harabasz Index for Density-Based Clustering

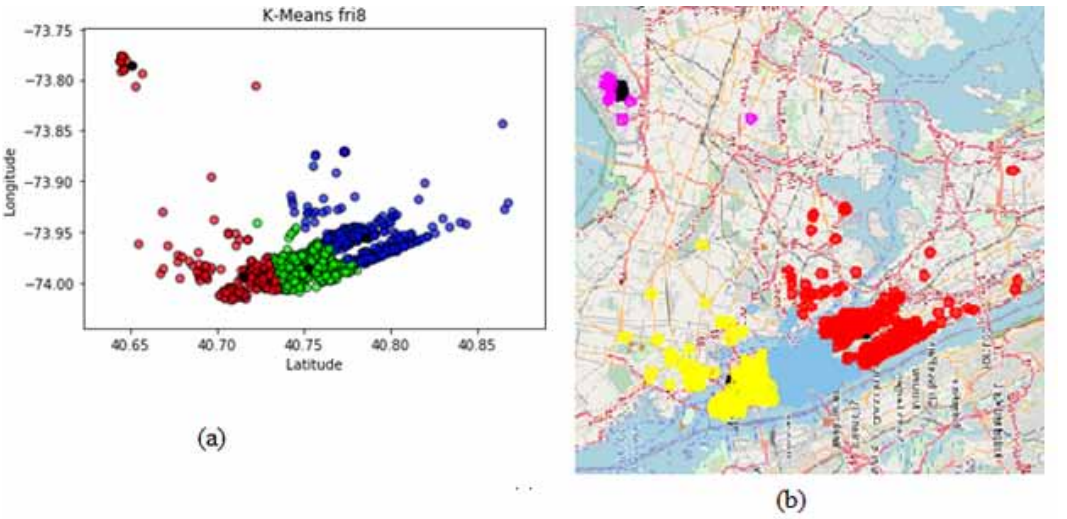
	Average Silhouette Coefficient		Average Calinski-Harabasz Index	
	DBSCAN	**OPTICS**	**DBSCAN**	**OPTICS**
New York Dataset	0.336022	0.334287	384.6645	383.651
Porto Dataset	0.521844	0.222097	40.87046	106.0731
Mexico Dataset	0.150233	0.126249	9.214427	8.684412

of datapoint to cluster is set to 4 for density-based clustering algorithms over standard unsupervised cab datasets of New York, Porto and Mexico as shown in Table 3. However, in case of average Calinski-Harabasz Index of Porto Dataset, DBSCAN does not show similar results as the dataset is smaller in size.

The average Silhouette Coefficient of DBSCAN is higher for three cities as 0.336022, 0.521844 and 0.150233 as compared to average Silhouette Coefficient of OPTICS as 0.334287, 0.222097 and 0.126249 respectively. The average Calinski-Harabasz Index for DBSCAN for three cities are 384.6645, 40.87046 and 9.214427 as compared to average Calinski-Harabasz Index for OPTICS as 383.651, 106.0731 and 8.684412 respectively.

The experiment generates two sets of output files for each sample: Cluster Graph Visualization and Cluster Map Visualization. Figure 9(a) shows the cluster graph visualization file for the New York dataset for Friday at 08:00 AM using K-Means clustering. Figure 9(b) is the cluster map visualization for the same file. The data points in a similar colour form a single cluster and the data points of different colours depicts they belong to different clusters. The centroids of the clusters are represented in Black colour. Since, K-Means algorithm does not mark the outliers so the clusters having number of points less than the minimum points required to form a cluster are not part of any cluster and therefor not marked.

Figure 9. (a) K-Means Cluster Graph Visualization (b) K-Means Cluster Map Visualization for New York Dataset on Friday at 08:00 AM



CONCLUSION AND FUTURE SCOPE

The research study aimed at identifying the optimal clustering technique to cluster passenger pickup geolocation that shall assists cab drivers to find the next passenger finding the location at the nearest distance to develop an efficient cab recommender system. This research paper proposes an algorithm to cluster passenger pickup locations using standard clustering techniques for an efficient cab recommender system. The evaluation of the algorithm was done using standard parameters of the Silhouette Coefficient and Calinski-Harabasz Index over three unsupervised datasets of New York, Porto and Mexico. In the case of Hierarchical clustering, BIRCH performance yield higher values of average Silhouette Coefficient and average Calinski-Harabasz Index than CURE for clustering passenger pickup geolocations over the standard datasets with the minimum number of clusters set to 4. In the case of partition-based clustering algorithms, K-Means generates better clusters as compared to

Mini Batch K-Means and Spectral Clustering over the standard datasets with the number of clusters set to a value of 4 using the elbow-plot method. For density-based clustering algorithm, DBSCAN generated better clusters than OPTICS with a higher average silhouette coefficient and higher average Calinski-Harabasz index over the standard dataset. Future work will focus on working with hybrid clustering techniques to generate better clusters that shall develop an efficient cab recommender system for solo cab driver.

ACKNOWLEDGMENT

The authors of this publication declare there are no competing interests. This research was supported by Panjab University, Chandigarh, India.

REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. *SIGMOD Record*, *28*(2), 49–60. doi:10.1145/304181.304187

Bade, K., & Nurnberger, A. (2006). Personalized hierarchical clustering. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, (pp. 181–187).

Berkhin, P. (2006). *Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data*. Springer.

Braga, R., Tahir, A., Bertolotto, M., & Martin, H. (2012). Clustering user trajectories to find patterns for social interaction application. *Lecture Notes in Computer Science*, *7236*, 82–97. doi:10.1007/978-3-642-29247-7_8

Cengizler, C., & Ün, M. (2017). Evaluation of Calinski-Harabasz Criterion as Fitness Measure for Genetic Algorithm Based Segmentation of Cervical Cell Nuclei. *British Journal of Mathematics & Computer Science.*, *22*(6), 1–13. doi:10.9734/BJMCS/2017/33729

Cross, C. (2017). *Taxi Trajectory Data.* https://www.kaggle.com/datasets/crailtap/taxi-trajectory

Deep, A. S. (2015). Taxi Trip time prediction using similar trips and road network data. *IEEE International Conference on Big Data, IEEE Big Data,* (pp. 2892-2894).

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *International Conference on Knowledge Discovery and Data Mining*, Portland, OR, USA. (Vol. 2 pp 226-231).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100–108. doi:10.2307/2346830

Hu, H., Wu, Z., & Mso, B. (2012). Pick-up tree-based route recommendation from taxi trajectories. *Lecture Notes in Computer Science*, *7418*, 471–483. doi:10.1007/978-3-642-32281-5_45

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *Hierarchical Clustering*. Cambridge University Press.

Mazimpaka, J., & Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science.*, *13*(13), 61–99. doi:10.5311/JOSIS.2016.13.263

Mittal, Y. (2016). Finding optimal locations for taxi stands on city map. [Master's Dissertation]. IIIT Delhi.

Mooney, R. J., & Roy, L. (1999). Content-based book recommendation using learning for text categorization. In *Workshop Recommender Systems*. Algorithm and Evaluation.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. doi:10.1093/comjnl/26.4.354

Navas, M. (2017). *Taxi Routes of Mexico City, Quito and More.* https://www.kaggle.com/datasets/mnavas/taxi-routes-for-mexico-city-and-quito

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In: *14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01)*, (Vol. 15, pp. 849–856).

NYC Open Data. (2014) *2014 Yellow Taxi Trip Data.* [Dataset]. https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n/

Pazzani, M. (1999). A framework for collaborative, content-based, and demographic filtering. *Artificial Intelligence Review*, *13*(5/6), 393–408. doi:10.1023/A:1006544522159

Pedregosa, . (2011). Scikit-learn: Machine Learning in Python. *JMLR*, *12*, 2825–2830.

Resnick, P., & Varian, H. R. (1997). Recommender System. *Communications of the ACM*, *40*(3), 56–58. doi:10.1145/245108.245121

Ricci, F. (2011). *Introduction to Recommender Systems Handbook. Recommender Systems Handbook*. Springer.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi:10.1016/0377-0427(87)90125-7

Sculley, D. (2010). Web-scale k-means clustering. *Proceedings of the 19th international conference on World wide web (WWW '10)*, Association for Computing Machinery, New York, NY, USA, (pp 1177–1178). doi:10.1145/1772690.1772862

Shen, Y., Zhao, L., & Fan, J. (2015). Analysis and Visualization for hot spot-based route recommendation using short-dated taxi GPS traces. *Information (Basel)*, *6*(2), 134–151. doi:10.3390/info6020134

Shutaywi, M., & Kachouie, N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy (Basel, Switzerland)*, *23*(6), 759. doi:10.3390/e23060759 PMID:34208552

Szenasi, S. (2014, June). Clustering Algorithms in Order to Find Accident Black Spots Identified by GPS Coordinates. 14th SGEM GeoConference on Informatics, GeoInformatics And. Remote Sensing, (pp. 17–26).

Wang, R., Chow C., Lyu Y., Victor C., Kwong S., Li Y. and Zeng J. (2017). TaxiRec: Recommending Road Clusters to Taxi Drivers Using Ranking-Based Extreme Learning Machines. *IEEE Transactions on Knowledge and Data Engineering*, (pp. 1-1. 10.1109).

Wu, L., Hu, S., Yin, L., Wang, Y., Chen, Z., Guo, M., Chen, H., & Xie, Z. (2017). Optimizing cruising Routes for Taxi Drivers using the Spatio-temporal Trajectory Model. *SPRS International Journal of Geo-Information.*, *6*(11), 373. doi:10.3390/ijgi6110373

Yuan, N. J., Zheng, Y., Zhang, L., & Xie, X. (2013). T-Finder: A recommender system for finding passengers and vacant taxis. *IEEE Transactions on Knowledge and Data Engineering*, *25*(10), 2390–2403. doi:10.1109/TKDE.2012.153

Supreet Kaur Mann received her MCA with distinction from Thapar University, Patiala in 2010. She completed her MTech-CSE with distinction from Lovely Professional University, Phagwara in 2014. Currently, She is Assistant Professor at Department of Computer Science and Applications, Panjab University, Chandigarh. She is pursuing her PhD in Computer Science under the guidance of Prof. Sonal Chawla.

Prof. Sonal Chawla is MCA with distinction from Thapar Institute of Engineering and Technology, Patiala and PhD in Computer Science from Panjab University, Chandigarh. Currently, She is Professor at Department of Computer Science and Applications and Honorary Director of Centre for IAS Studies, Panjab University, and also Fellow of Panjab University Senate. She is an avid researcher and an accomplished academician having more than 50 research papers in National/International refereed journals. She is a recipient of AICTE Career Award for Young Teachers and has successfully completed projects from various Government Agencies like UGC, AICTE etc. She is the Editorial Board/ Advisory Member of many Journals of repute besides being on the panel of State and Central Universities and Institutions in different capacities.