Analysis of High-frequency Occurrence Areas of Beijing Taxis based on the K-means Method

Yingqi Xu*

Department of Brunel College, North China University of Technology, Beijing, China

* Corresponding Author Email: 21190010104@mail.ncut.edu.cn

Abstract. As the economy grows and of the pace of life accelerates, travelling by taxi has gradually become one of the most convenient and fastest choices for residents to go out. However, because of the randomness of people's travelling and the mobility of taxi driving, there exist an imbalance between the travelling demand of regional taxi users and the supply of taxis. In order to avoid the problem of uneven distribution of taxis and to achieve accurate prediction of taxi demand, this paper selects the vehicle number, latitude and longitude, and carrying status data of Beijing taxis as the basis of the study, and researches to explore the spatial distribution characteristics of the travelling demand of taxi users. The main contents of the paper are as follows. First of all, this paper initializes the GPS data of taxis, and obtains the boarding and alighting locations. The distribution of taxi regional travel demand is described by using K-means clustering method. Finally, the paper shows five high frequency locations. It is of great significance for taxi dispatching management, reducing empty rate, saving energy consumption and maximizing passenger travel demand.

Keywords: Taxi hot spots, K-mean clustering algorithm, travel demand modelling.

1. Introduction

With the expansion of the city scale and the acceleration of the pace of life, people are increasingly pursuing efficient transportation and travel modes. Taxi has the characteristics of all-weather operation, wide coverage and good data timeliness, and its driving time, driving route and boarding and alighting places are completely decided by passengers. However, due to the randomness of people's travel and the mobility of taxi driving. It leads to the problem of uneven distribution of taxis. Achieving accurate prediction of taxi demand is important for taxi dispatch management, reducing the idling rate, saving energy consumption, and maximizing the passenger travel demand. Nowadays, taxis are commonly equipped with GPS devices, and taxis generate massive amount of GPS track data during operation, which provides a possibility for taxi demand prediction based on deep learning methods.

The problem of hotspot area detection has always been the focus of research in the field of urban transport and travelling. More and more scholars obtain the hotspot areas of taxi users' trips through various types of spatial clustering algorithms, and further analyse the characteristics of users' trips in the hotspot areas and the problem of demand forecasting. There has been a lot of work on traditional machine learning prediction methods for trajectory destinations. Bayat et al used a Bayesian inference approach to predict cab destinations [1]. This method divides the traveling area of the cab into grids, after which the prior and posterior probabilities of each trajectory are counted based on the historical information, and ultimately Bayesian formulas are used to compute the probability that the destination is in each grid. Monrea et al does not predict the travel destination, it only predicts the next location where the object may travel, the model constructs a T-tree from historical trajectories, and then finds trajectories similar to the current trajectory from the T-tree, and the next location of this trajectory is used as the prediction result [2]. Wiest et al. used a Gaussian mixture approach to predict cab destinations, which establishes a probability distribution function between the travel trajectory and the travel destination, and then later predicts the destination of the cab through conditional probabilities [3]. Guillouet et al. used a clustering algorithm for the first time to destination prediction, the algorithm will first classify the trajectories into different categories according to their similarity, and then the trajectories that need to predict the destination will be similar to the clusters of

trajectories that have already been clustered to calculate the similarity, and the first few categories will be used to take the average of the weights as the destination [4]. Spatial clustering methods can be mainly divided into: clustering based on division, including K-means clustering,clustering based on density, including DBSCAN clustering, clustering based on statistical models, including kernel density clustering, etc. Hu used K-means clustering method to analyse the data after dimensionality reduction using Matlab, and by adjusting the parameter K-value selection, and finally obtained a stable travel hotspot area [5]. Zhao et al. improved the traditional K-means algorithm, and proposed a clustering model based on the self-organising mapping and K-mean fusion algorithm, which reduced the model's computing time [6]. Woong and others believed that the DBSCAN algorithm needs to repeat the calculation of the distance to all the objects in order to find the neighbouring objects when performing clustering, and because the intermediate clustering results need to be retained in the process of calculating, which makes the computational efficiency lower, so they proposed the CudaSCAN clustering method based on the DBSCAN, which further enhances the computational efficiency of the clustering [7]. Wang clustered the GPS track data of taxi in Xi'an according to the kernel density clustering algorithm, and classified the passenger hotspots into ranks and drew the spatial distribution map of hotspots [8]. In addition to the above spatial clustering algorithm based on GPS data, scholars also tried to use other information to solve the hotspot area detection problem. For example, Chen and others found that people preferred to be near the point of interest when getting on and off the bus, and this preference led to the phenomenon of high and low density clustering of travelling demand in the generation of spatial distribution, which led to the proposal of a detection method based on the peak of local density, and the results accurately identified the high travelling demand area [9]. Bi and others based on taxi GPS data and combined with land use data in the main urban area of Nanjing, using community detection technology to analyse the hotspots of taxi passenger travel, the results show that with the increase of residents' travel activities, travel hotspots will form clusters in urban space [10].

By reading the literature, the more common K-mean clustering method was used. The algorithm is intuitive in its thinking, fast in convergence, and better in clustering. The only parameter that needs to be adjusted is the number of clusters, and the algorithm is more interpretable. At the same time, the K-mean clustering algorithm also has the defect of iterative method. The clustering results may converge in the local optimum so as not to get the global optimal solution, the non-convex shape of the class cluster identification effect is poor, vulnerable to noise, edge points, isolated points, can deal with the limited type of data, for high-dimensional data objects clustering is not good, the selection of the K value of the bad grasp. In this paper, the data is firstly processed by dimensionality reduction to improve the accuracy of the model.

2. Methods

2.1. Data and Indicators

The dataset is from AMap, which counts the GPS data of Beijing taxis on 24 September 2019, and the data contains the vehicle number, latitude and longitude, whether it is carrying passengers or not, and the speed of the vehicle.

In the paper, K the number of clustering centers, is the number of categories of the sample after the clustering algorithm. The passenger carrying state is represented by a number, where 0 is the empty state as well as 1 is the carrying state. Therefore, when change of carrying status from 0 to 1, the site is defined as boarding location and when it changes from 1 to 0, the location is the alighting point.

2.2. Methodology

Clustering is a method of categorizing and aggregating data members of a data set that are parallel in a sense, clustering is a skill for research this intrinsic structure, clustering is unsupervised learning.

The k-mean clustering is the well-known algorithm for separating clusters. It is the most often used clustering method because to its ease of use and effectiveness.

The k-means (k-means clustering algorithm) is an iterative solution of cluster analysis algorithm. The procedures are to partition the data into K groups beforehand, choose K objects at random to serve as the initial clustering centers, and then calculate the distance between each point and the seed clustering centers, as well as assign each point to the closest cluster center. A cluster is represented by the cluster centers and the items associated to them. The cluster centers are adjusted for each sample assigned depending on the clusters' current items. This procedure is repeated until some termination condition is satisfied. Two examples of termination conditions are that no (or a minimal number of) items may be transferred to new clusters and that no (or a minimum number of) cluster centers may be changed once more, and the local minimization of the squared error sum is achieved [11, 12].

3. Results and Discussion

3.1. Data Preprocessing

The taxi GPS data selected for this paper contains multiple attribute fields per record, some of which are not meaningful for the research of this paper. For example, attributes such as the longitude, latitude and vehicle type are not very relevant to the research content of this paper. To lessen the demand for storage space and improve the research effectiveness of the ensuing experimental process, this paper eliminates the fields with little relevance in the GPS data, and only retains the effective fields to reduce the dimensionality. After the dimensionality reduction of the GPS data, the value of the data can be improved, and the loss of space storage resources can be reduced. Table 1 displays the data after dimensionality reduction.

Table 1. Vehicle data

Vehicle number	Longitude	Latitude	Passenger-carrying state
14027482	116.220	40.142	1
14244152	116.634	40.072	1
14246090	116.156	39.920	0
13229075	116.788	40.324	1
13896346	116.699	39.647	1

Since the data are captured randomly, many of them are in normal driving state and do not reflect the location of boarding and alighting well. So, the paper records the passenger status change of each vehicle in the process of processing the original data one by one, and when the passenger status field of a vehicle changes from idle to passenger, judge the location of the vehicle as the boarding point; when the status field changes from passenger to idle, judge the location of the vehicle as the alighting point at this time. The process is written in Python, and then the required vehicle number, latitude and longitude are extracted from the filtered records. We need the vehicle number to distinguish each record belongs to the vehicle, so that we can determine which boarding and alighting positions are generated by the same vehicle records, combined with the date of the record, we can launch these boarding and alighting points of the corresponding relationship. The latitude and longitude are necessary to calculate the hotspot areas.

Read the CSV file and group the data according to the vehicle ID. Sort the information inside each group based on when it was sent. Record the data with the first and last occurrence of 1 in the status column in each group and extract the corresponding latitude and longitude.

3.2. Problem Analysis

First of all, research the boarding and alighting locations. Read the CSV file and group the data according to the vehicle ID. Sort the data in each group based on when the data was sent. Record the

first and last occurrence of 1 in the status column of each group, and extract the corresponding longitude and latitude information. Afterwards, categorize all locations. Use K-means algorithm to cluster the position coordinates of all groups and find the cluster center. Find the nearest data point to each cluster center and record the relevant information. Ultimately, generate two graphs representing the first and last occurrence of the clustering center graphs. Send the data time, latitude and longitude, print to a document.

The paper uses Euclidean distance to calculate the spacing between two points.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

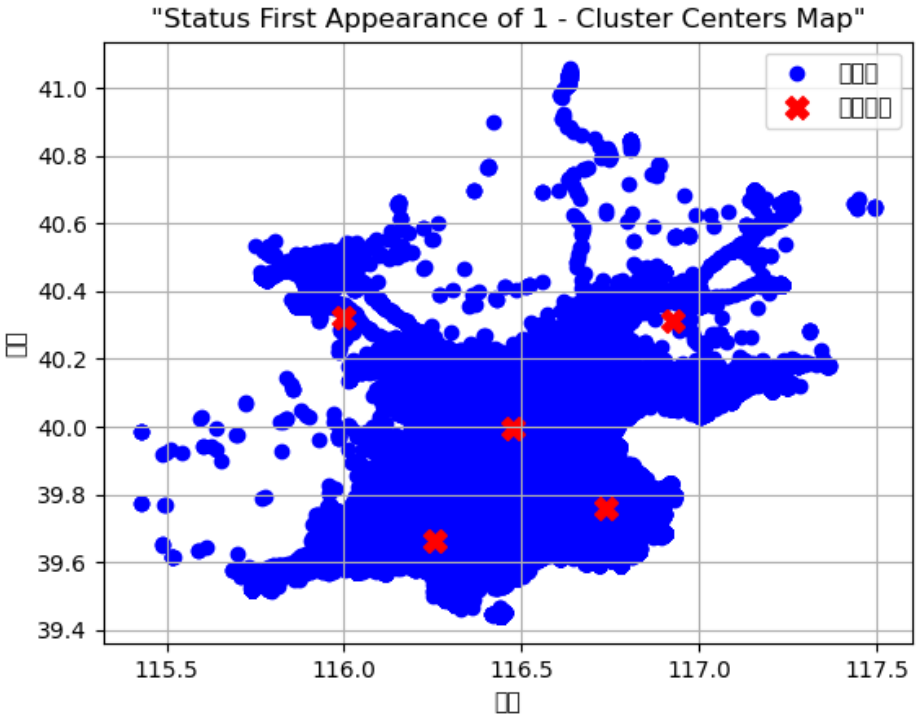

Fig. 1 Boarding location

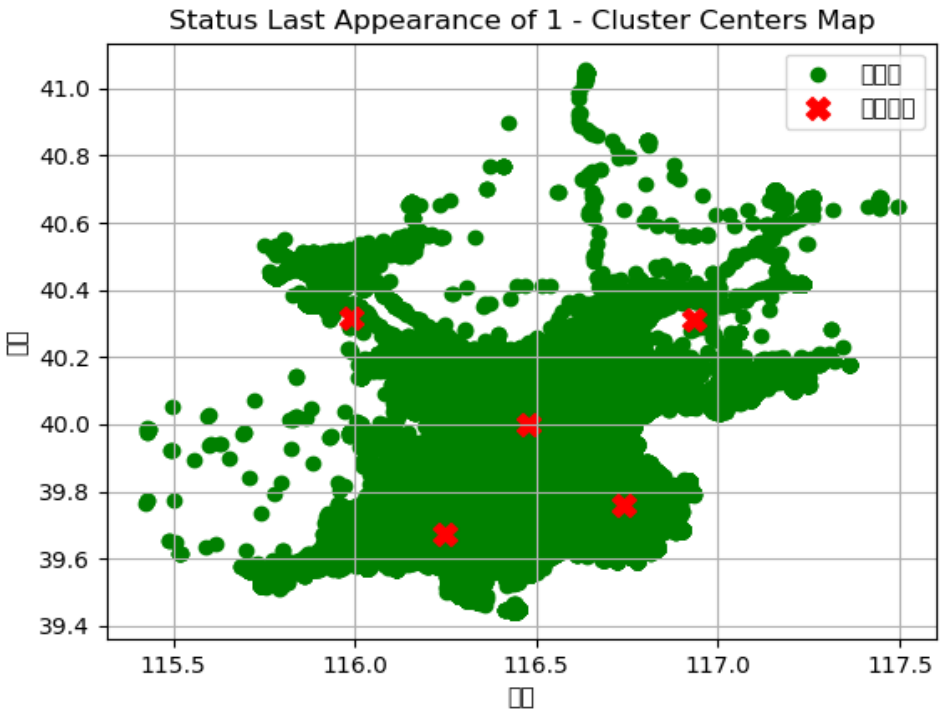

Fig. 2 Drop-off location

Figure 1 and 2 reflect five clustering centers for the boarding and alighting locations, respectively, for a total of ten clustering centers. After comparison, it is found that the clustering centers in the two graphs correspond to each other one by one, and the corresponding clustering centers are extremely close to each other. The reason is that after the passengers get off the bus, they will preferentially go to the nearest location to pick up the next passenger. Therefor in this paper, ten clustering centers are viewed as five for the study. Since the data comes from the Amap, the Gaode open center is used to locate the center. From left to right: Badaling Ancient Great Wall Natural Scenic Area, Yongding River, Wangjing SOHO Mall, Beijing Global Resort, Nanshan Ski Resort.

4. Conclusion

In this paper, the GPS data of Beijing taxis on 24 September 2019 are used, preprocessed and then classified using the K-means algorithm to obtain five clustering results: the Badaling Ancient Great Wall Natural Scenic Area, the Yongding River, the Wangjing SOHO shopping mall, the Beijing Universal Resort, and the Nanshan Ski Resort. These five places have high demand for taxis. Appropriate deployment can be done to supply passenger demand and at the same time reduce the empty taxi rate. However, there exist shortcomings in this study. On the hand, the algorithm itself has limitations and the result may be a local optimum rather than a global optimum. And the algorithm is sensitive to noise and outliers, and is highly disruptive to cluster center assignment. On the other hand, the data itself has a certain bias, slightly different from the real data, the study did not carry out back-correction, so the results of the study may also have a certain deviation from reality.

References

[1] Bayat S, Naglie G, Rapoport M J, et al. Inferring Destinations and Activity Types of Older Adults from GPS Data: Algorithm Development and Validation (Preprint). 2020.

[2] Monreale A, Pinelli F, Trasarti R , et al. WhereNext: A location predictor on trajectory pattern mining. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009.

[3] Wiest J, Hoffken M, Kresel U, et al. Probabilistic trajectory prediction with Gaussian mixture models. Intelligent Vehicles Symposium (IV), 2012.

[4] Besse P C, Guillouet B, Loubes JM, et al. Destination Prediction by Trajectory Distribution-Based Model. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(8): 2470-2481.

[5] Hu Lanlan. Recommendation of high-yield hotspot areas for GPS-based taxis. Wenzhou University, 2015.

[6] Shasha Zhao, Yi Xiao, Yueqiang Ning, Yuxiao Zhou, Dengying Zhang. An Optimized K-means Clustering for Improving Accuracy in Traffic Classification. Wireless Personal Communications, 2021.

[7] Woong-Kee Loh, Hwanjo Yu. Fast density-based clustering through dataset partition using graphics processing units. Information Sciences, 2015, 308.

[8] Wang, Li Xuan. Research on urban taxi passenger travelling characteristics based on KDE and GWR. Chang'an University, 2020.

[9] Chen X J, et al. Urban hotspots detection of taxi stops with local maximum density. Computers Environment and Urban Systems, 2021.

[10] Bi S, et al. Analysis of Travel Hot Spots of Taxi Passengers Based on Community Detection. Journal of Advanced Transportation, 2021.

[11] Fu R, Zhang Z, Li L. Using LSTM and GRU neural network methods for traffic flow prediction. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE, 2016: 324-328.

[12] Hong W. Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. Neurocomputing, 2011, 74(12): 2096-2107.