Improvement of Short-distance Subsidy Policy for Airport Taxi Based on Cluster Analysis

Tao Yang*
School of communications engineering, North China Electricity Power University
Baoding071003, China

*Corresponding author e-mail: shuizjsztx02@163.com

Abstract. The income of airport taxis is related to the mileage. Due to the difference in passenger destinations, the mileage is long and short, and the taxi driver cannot pick or refuse to load. Therefore, the difference between short-haul passengers and long-haul passengers is significant. Based on such realistic problems, a relatively reasonable priority policy has been proposed to balance the income of taxi drivers. Under the existing policy based on the short-distance ticket of Shanghai Pudong Airport, combined with the large number of drop-off point data of taxis, the data is pre-processed and then visually analyzed on Bigemap, and then cluster analysis is carried out in matlab environment. The mileage distribution makes a more reasonable division of short-distance long-distance definition and proposes new ideas for solving the problem.

1. Introduction
Most passengers have to go to the city (or surrounding) destination after getting off the plane. Taxi is one of the main choices for passengers. The rental income of the airport is related to the mileage of passengers. The destination of passengers is far and near. Taxi drivers cannot choose passengers and refuse to carry them, but they allow taxis to travel multiple times. The management expects to give certain “priority” to certain taxis that return passengers once again, so that the income of these taxis is as balanced as possible.

2. Discussion on Range Division Based on Cluster Analysis

2.1. Problem hypothesis
(1) Make the assumption that when passengers who set off from the airport drop-off is close to the cluster point, it means that the driver has a higher probability of accepting the next order after a short empty load, This will reduce the loss caused by the short distance from the previous one received from the airport. And if the drop point deviates from the cluster point, or in the obvious suburban and remote counties, the subsequent dead time will remain relatively long. Time. Long-term idling will inevitably bring economic losses to taxi drivers.

(2) Assuming that the data record of the boarding point near Shanghai Pudong Airport can be approximated as being the case of drivers departing from the airport after queuing at the airport, then analyse the exit point of each record, and replace the mileage with the route distance between the two
points of the boarding and unloading point. The approximate mileage distribution of the taxi order can be obtained through the matlab visualization function.

2.2. Problem analysis
The subsidy scheme adopted by Pudong Airport is that if the taxi receiving the short-distance order chooses to continue to return to the airport for queuing, the airport will give it a certain priority. After research and analysis on the taxi subsidy scheme of Shanghai Pudong Airport, the model is in Shanghai. Based on the implementation of the subsidy program of Pudong Airport, the cluster analysis model and the mileage distribution of taxi orders are used to determine the long-and short-distance division criteria, and the division of long-and short-distance boundaries is made more reasonable, and the subsidy scheme is further optimized.

2.3. Model establishment and solution

2.3.1. Data preprocessing. In order to analyze the distribution of passengers' drop-off points in Shanghai. This article collected GPS data of more than 6,000 taxis in Shanghai within 30 days, In view of research needs, Focus on extracting the longitude and latitude data and passenger's status data required in the original records, and rejecting the data with incorrect or incomplete records, so the first task is to filter and clean the original data information to form a new data format.

| Company | Number | longitude | latitude | speed | direction | Passenger’s status | GPS status | time              |
|---------|--------|-----------|----------|-------|-----------|--------------------|------------|------------------|
| Q5      | 10788  | 121.4345  | 31.3422  | 32.3  | 242       | 1                  | 1          | 2018-04-01 12:52:16 |
| Q5      | 10028  | 121.3727  | 31.1594  | 30.4  | 339       | 1                  | 1          | 2018-04-01 12:52:16 |
| Q5      | 11446  | 121.3868  | 31.0559  | 39.5  | 79        | 0                  | 1          | 2018-04-01 12:52:16 |

After re-processing the filtered GPS data, it is not difficult to find that when the passenger’s status changes from 0 to 1, the corresponding latitude and longitude at this moment is the boarding point position. Similarly, when the passenger status changes from 1 to 0, the corresponding latitude and longitude is the position where people get off.

2.3.2. Data aggregation research. According to this basis, the geographic location of the getting on and off points is further extracted from the initial large amount Bigemap software for visual analysis, taking the daily data of 2018-04-20 as an example, as shown below, each sample point represents the distribution of the getting on and off points. The denser the point distribution, the more frequent the getting on and off:

| number | longitude | latitude | speed | Passenger’s status | time              |
|--------|-----------|----------|-------|--------------------|------------------|
| 10788  | 121.4345  | 31.3422  | 32.3  | 1                  | 2018-04-01 12:52:16 |
| 10028  | 121.3727  | 31.1594  | 30.4  | 1                  | 2018-04-01 12:52:16 |
| 11446  | 121.3868  | 31.0559  | 39.5  | 0                  | 2018-04-01 12:52:16 |
It can be judged from the figure that the getting on and off points in Shanghai all day show a clear agglomeration phenomenon, and mainly concentrated in the central areas of Shanghai, such as Huangpu, Jing'an, Changning, Putuo, Xuhui and other districts.

2.3.3. K-means algorithm introduction and Matlab implementation. Due to the obvious agglomeration phenomenon of the data, this paper chooses to apply the K-means algorithm for cluster analysis. The K-means algorithm is a classical clustering algorithm in machine learning technology. In the process of clustering, the distance between two points is used as the evaluation index. The closer the distance is, the higher the similarity is. The closer the elements are. Combine them together to form a cluster and use these composed clusters as the ultimate goal.

The working principle of the K-means algorithm is as follows:

Figure 2. K-means algorithm schematic diagram
(The circle in the figure is the cluster center, and the square is the data to be clustered.)

1) Initialization. Enter the gene expression matrix as the object set X, input the specified cluster class number N, and randomly select N objects in X as the initial cluster centers. Set the iteration abort condition, such as the maximum number of loops or the cluster center convergence error tolerance.

2) Iterate. Data objects are assigned to the nearest cluster center to form a class according to similarity criteria. Initialize the membership matrix.

3) Update the cluster center. The data objects are then reassigned with the average vector of each class as the new cluster center.

4) Repeat steps 2 and 3 until the suspension condition is met.

And two questions about K-means algorithm:

1) It does not guarantee the best solution for finding a clustering center, but it guarantees convergence to a solution (no infinite iteration).

Solution: Run K-means a few times, each time the initial cluster center point is different, and finally select the result with the smallest variance.

2) It cannot tell how many categories are most suitable in a data set, such as the above illustration, the final result obtained by selecting different initial categories is different.

Solution: First set the number of categories to 1, then gradually increase the number of categories. Use the above method in each category. In general, the total variance will drop quickly until it reaches an inflection point; this means the variance does not significantly reduce by adding another cluster center, then save the number of clusters at this time.

Based on this principle, we calculated 15 cluster points through matlab programming, and plotted the cluster point coordinate data on the map in R language:

![Figure 3. Cluster center distribution (2018-4-20)](image-url)
The above is only a cluster analysis of Shanghai's one-day situation, and the cluster center of each day is further statistically summarized.

Table 3. Cluster point occurrence statistics

| District name       | Cluster point occurrences | District name       | Cluster point occurrences |
|---------------------|---------------------------|---------------------|---------------------------|
| Pudong New Area     | 127                       | Jiading District    | 28                        |
| Minhang District    | 62                        | Baoshan District    | 28                        |
| Putuo District      | 48                        | Hongkou District    | 15                        |
| Changing District   | 45                        | Jinan District      | 13                        |
| Xuhui District      | 40                        | Qingpu District     | 16                        |
| Huangpu District    | 32                        | Songjiang District  | 4                         |
| Yangpu District     | 32                        | Total               | 480                       |

Although the K-means algorithm has certain limitations, it must accumulate a large number of drop-off points at each cluster point. Therefore, each cluster point can be regarded as a regional active point in Shanghai. The regional distribution of active points provides a strong basis for the short-range of this paper.

2.3.4. Division of short-distance areas. Combined with the distribution of cluster points and the approximate mileage distribution of taxi orders, this paper makes a new improvement to the long and short divisions originally scheduled according to the fixed mileage, which is divided according to the central circle as shown in the figure below, according to the distribution of cluster points. The situation has been improved.

Figure 4. Initial division diagram
The adjusted area is divided as follows:

![Figure 5. Improved schematic](image)

3. Conclusion
According to the scenario given by the topic, this paper makes a new improvement to the division strategy of the short-distance range of the airport. Referring to the cluster distribution points of the getting on and off, the basic mileage division is adjusted to further balance the taxi driver's receipt income. Applying clustering algorithm to regional division and providing a new solution to the existing situation.

References
[1] Cao Liang. Research on Temporal and Spatial Changes of Urban Travel and Carbon Emission Based on Taxi GPS Data [D]. Shanghai Normal University, 2018
[2] Li Wei. Design and implementation of urban planning management system based on ArcGIS [D]. Xidian University, 2018.
[3] Yu Wentian. Analysis of spatial characteristics of rental vehicle-mounted passengers based on GPS data [D]. Jilin University, 2013.
[4] Wang Rui. A large-scale scale topographic map based on high-resolution remote sensing image [J]. Surveying and Spatial Geography Information, 2017, 40 (10): 125 - 128.
[5] Information on [http://www.bigemap.com/helps/](http://www.bigemap.com/helps/)