Comprehensive Analysis Based on Massive Taxi Trajectory Data

Longyan Tan

1School of Reliability and Systems Engineering, Beihang University, Beijing, China
*18375165@buaa.edu.cn

Abstract. With the increasing demand for taxis, a large number of taxi trajectory data have been used for in-depth analysis. However, most early studies only used limited data or simulations to analyze the route choice behavior or traffic congestion problems. To extract more useful information from taxi trajectory data, this paper analyzes massive raw dates collected in Chengdu from 4 different perspectives. To begin with, passenger demand is analyzed by counting the number of orders in each period. K-means clustering is used for regional planning and provides appropriate boarding points for passengers. Furthermore, this paper focus on quantity duration of order duration, which presents an approximate Poisson distribution. Finally, the ratio of driving distance to the linear distance between the start and end points is defined as the overtravel ratio. Through the comprehensive analysis of the taxi trajectory data, it can provide assistance to the travel demand of passengers and the scheduling of taxi drivers, and maximize resource efficiency. Moreover, the data collected in Chengdu are very representative and large, so the corresponding suggestions can also provide hints for public transport problems.

1. Introduction
As the taxi is one of the primary means of transportation, large-scale, high-quality, and consecutive spatiotemporal data have been collected by Global Position System (GPS). And with the development of ride-hailing services, these data are collected by some companies and researchers who provide beneficial suggestions to solve the transportation problems. The discovery of zones and people’s movement patterns supports a better understanding of modern cities and enables a more comprehensive strategy for urban planning. In terms of functional region discovery and movement patterns, many clustering algorithms have been developed. For example, Kim et al. developed a method that combined multiple [1], and Konjar et al. researched population flow of commute data to explore the economic interaction [2]. Deng et al. proposed a method combining clustering algorithms and field theory to describe interactions such as spatial and attribute relationships among spatial entities [3]. Chen et al. proposed a modified method based on previous research to simultaneously discover people’s zones and movement patterns, facilitating a better understanding of the complexities of cities and is conducive to rational urban planning [4]. Many other researchers use these datasets to monitoring abnormal events. These anomalous driving trajectories provide us an opportunity to extract driver or passenger behaviors and monitor adverse urban traffic events. For instance, Kuang et al. transformed taxi GPS data into a traffic flow matrix and used Shewhart control chart to detect anomaly traffic events, but anomalies reported were limited [5]. Chandola et al. proposed most of these methods can be grouped into four categories: distance-based, cluster-based, classification-based, and statistics-based categories [6]. Wang et al. analyzed these anomalous trajectories and discovered four anomalous behavior patterns to speculate on the cause of an anomaly based on statistical indicators of time and length [7]. However,
there are still a lot of challenges mentioned by Hwang et al. [8]. Besides, how to detect traffic congestion is also a topic issue. Traffic congestion has severe influences on the daily life of people. Due to typical recurrent mobility patterns of commuters and transport fleets, we can detect traffic congestion events on selected hours of the day, so called rush hours. For example, D’Andrea and Marcelloni designed a system that detects real-time traffic congestion and incidents on roads based on analyzing GPS trajectories of moving vehicles [9]. Kelet et al. introduced Shared Nearest Neighbor (SNN), followed by a density-based clustering to detect traffic congestion in different periods. Nevertheless, the data is delayed, and more clustering techniques should be tested to ensure accuracy [10]. Qiang et al. described a methodology using GIS applications that can visualize and identify trends in traffic conditions over time along corridors selected for analysis. This methodology can be applied to evaluate and monitor congested corridors in response to changes in land use patterns, policy decisions, and various temporary conditions affecting street operations. However, this methodology is limited in specific cases and lacks precision [11]. Route choice is also a hot topic among data science researchers. Analysis of the routing choice behavior provides theoretical support for route guidance and traffic assignment. For instance, Schussler and Axhausen collected travel data in Zurich area and calibrated the C-logit model and PS-logit model [12]. Li et al. collected data in Toyota City to explore the effect of travelers’ heterogeneous route choice. They concluded that route choice behavior is affected by travelers’ age, gender, vehicle displacement, and O-D’s characteristics [13]. Deng et al. proposed that drivers’ route choice behavior could be influenced by the distance [14]. The research based on taxi trajectory data has also made good progress in other aspects. Laha et al. discussed several incremental learning methods to solve the real-time destination/next pick-up location prediction problem. Yet, the distances in datasets that they considered were mostly short, so their results’ accuracy should be verified with datasets with long-distance [15]. Furthermore, Rodrigues et al. explored deep learning architectures for combining time-series and textural data for mobility demand prediction in eventful areas. This method based on text data lacks accuracy [16]. However, all the research above is based on relatively limited datasets which are not very representative. In addition, data preprocessing is not mentioned in these studies, so that invalid data have a significant influence on their results.

The main contributions of this paper can be summarized as follows:
1. A large number of taxi trajectory data were collected in 2016 in Chengdu with a very effective data preprocessing method.
2. It shows the travel demand of passengers on weekdays, weekends, and holidays as well as at different times of the day.
3. This paper uses the K-means clustering algorithm to divide Chengdu into 20 representative areas and provide suitable pick-up points.
4. The quantity distribution of order duration is found to present an approximate Poisson distribution.
5. This paper defines the ratio of driving distance to the linear distance between the start and end points as the overtravel ratio, which can determine whether the drivers deliberately detour and provide suggestions about urban road planning.

The rest of this paper is organized as follows. In section 2, the authors describe the model used in this paper and the mathematical reasoning process of this model. The detailed experiments process and the results are exhibited in section 3. Section 4 concludes this paper and discusses future work.

2. Model Formulation
To begin with, due to the large quantity of dataset and invalid sampling points, this paper proposed several ways to preprocess the dataset. The most important one is to divide the trajectory dataset by time series. Afterwards, we can extract each individual order’ information such as starting time, ending time, starting latitude and longitude. Invalid sampling points are defined as the following 3 categories: 1. Orders with less than 20 sampling points or the duration of which is within 1 minute. 2. Orders with too close or too far distance between adjacent sampling points. 3. Drivers with less than 1000 sampling points per day. Then, the trajectory dataset is analyzed in the following 4 aspects.
To acquire the travel demands of passengers, the model will present the order distribution in each period. The results are very intuitive. Meanwhile, this paper recommends starting points by utilizing clustering algorithms. In the next section, this model proposes the order duration distribution as Poisson distribution. Ultimately, researchers can use this model to detect anomalies by analyzing the overtravel ratio. The overall framework of the proposed model is illustrated in Figure 1.

![Figure 1. Dataset Multiple-Aspects Analysis Model](image)

2.1. Order Distribution in Each Periods

In this part, we count the order distribution of orders in each period. Before computing, we have simplified the definition of the starting and ending points of each order. In view of that some orders are accepted by drivers online in advance, the starting points are not the sampling points that are collected firstly. On the contrary, there are usually some short distances between these two points in such orders. However, the influence of the short distance is not strong so we can ignore it. The next step is to choose the proper statistical time interval. In this paper, the smallest time interval is an hour. Ultimately, we can figure out orders distribution per hour in October 2016 in Chengdu.

2.2. Regionalization by K-means Clustering

The K-means is a variant of partitioning clustering, which aims to sub-group data objects of a dataset into disjoint clusters in a way that optimizes some criteria [17]. In K-means clustering, initial clusters are arbitrarily specified after which the algorithm calculates the centers of each cluster, and then generates new clusters by assigning each object to the closest cluster center. The algorithm continues until cluster centers do not change.

The first step in K-means clustering algorithm is to decide the value of K. Due to the limited computing resources, in this paper, K is set as 20. Afterward, we need to select initial centroids. Since a heuristic method is used in this paper, the location of initial centroids has a great influence on the final clustering result and running time, so we need to choose the appropriate k centroids, and it is best that these centroids are not too close.

The input is a sample set:

\[ D = \{x_1, x_2, x_3, \ldots, x_m\} \]  \hspace{1cm} (1)

The number of clusters is K and the number of iterations is N.

The output is clustering:

\[ C = \{C_1, C_2, C_3, \ldots, C_k\} \]  \hspace{1cm} (2)

Randomly selecting k samples from the data set D as the initial k centroid vectors:
For $t = 1, 2, 3 \ldots k$, to initialize cluster partition $C$ as:

$$C_t = \phi \ (t = 1, 2, 3 \ldots k)$$

(4)

For $t = 1, 2, 3 \ldots m$, we calculate the distance between sample data with the initial $k$ centroid vectors $\mu_j (j = 1, 2, 3 \ldots k)$:

$$d_{ij} = \|x - \mu_i\|^2$$

(5)

Then labeling $x_i$, which corresponds to the smallest $d_{ij}$ as $\lambda_{ij}$. Besides, we update $C_{\lambda i}$:

$$C_{\lambda i} = C_{\lambda i} \cup \{x_i\}$$

(6)

Finally, for $j = 1, 2, 3 \ldots k$, we re-calculate the initial centroid in each clustering until the clustering results no longer change.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

(7)

All the starting points can be divided into 20 clustering through $N$ times iterations. At the same time, regionalization is completed automatically. With well-defined regionalization, the taxi organization can dispatch taxis properly in peak hours in order to reduce the empty-load rate to maximize resource utilization and make more profits.

2.3. Drivers’ Daily Orders Distribution

In this part, the author researches the order duration. We can get that using the latest point minus the earliest point in each order. Due to the oversampling, the ending points were not the sampling points that were collected in the end. Nevertheless, researchers have found this error has a very limited impact on the results so we can ignore it. We try to approximate this distribution to a Poisson curve. The mathematical Eq. (8) of Poisson distribution is illustrated below:

$$P(x = k) = \frac{\lambda^k}{k!} e^{-\lambda} \ (k = 1, 2, 3 \ldots)$$

(8)

The trajectory data were all collected by GPS in urban area in Chengdu which indicates that it’s reasonable to replace order duration with order mileage. Combined with Chengdu taxi fare standards, we can figure out average daily income of drivers.

2.4. Overtravel Ratio

It’s acknowledged that the actual driving distance of the driver between the two locations must be significantly greater than the straight-line distance between the two. The degree depends on the rationality of the driver’s choice of road and the impact of road conditions and the rationality of road planning. In this paper, we define the ratio of driving distance to the linear distance between the start and end points as the overtravel ratio. Since the dataset cannot directly represent the distance between these two points, in order to better adapt to the dataset, we have made the following amendments to the definition of linear distance.

$$L_{\text{linear}} = 1000 \times ((\text{lat}_1 - \text{lat}_2)^2 + (\ln g_1 - \ln g_2)^2)^{1/2}$$

In Eq. (9), $L_{\text{linear}}$ to the linear-distance between starting and ending points. $\text{lat}_1$ and $\text{lat}_2$ means the latitude of the starting and ending points. While $\ln g_1$ and $\ln g_2$ refer to the latitude of the starting and ending points:

$$K' = \frac{\text{Samples}}{L_{\text{linear}}}$$

(10)
Suppose that all the factors that affect the value of $K$ have a multiplicative effect on $K$, we might as well define $K$ as:

$$K = a \times b \times c \times d$$ (11)

In Eq. (11), $a$ is the non-professional coefficient and $b$ is the road irrationality index. While $c$ indicates the secondary congestion impact coefficient. $d$ means the correlation coefficient.

3. **Experiment**

3.1. **Dataset**

The data consists of five columns: order number, driver number, timestamp, longitude, and latitude. The data with the same order number can be regarded as the sampling points collected in the same order. GPS collects data every 3 seconds, that is, latitude and longitude. The data utilized in this experiment is exhibited in table 1:

| driver         | indent         | timestamp       | longitude      | latitude      |
|----------------|----------------|-----------------|----------------|---------------|
| 19ff6c447f3bdb4e0f37 dfbe3c465fd5 | 4541eb113e79210ae7d 0d90e0e9236df | 1476325971 | 104.04585 | 30.65658 |
| 19ff6c447f3bda4e0f37 dfbe3c465fd5 | 4541eb113e79210ae7d 0d90e0e9236df | 1476325974 | 104.04638 | 30.65571 |
| 19ff6c447f3bda4e0f37 dfbe3c465fd5 | 4541eb113e79210ae7d 0d90e0e9236df | 1476325977 | 104.04666 | 30.65524 |
| 19ff6c447f3bda4e0f37 dfbe3c465fd5 | 4541eb113e79210ae7d 0d90e0e9236df | 1476325980 | 104.04977 | 30.65319 |
| 19ff6c447f3bda4e0f37 dfbe3c465fd5 | 4541eb113e79210ae7d 0d90e0e9236df | 1476325983 | 104.04618 | 30.65604 |

The time range of the original data covers from October 1, 2016 to October 31, 2016. The geographic range of the data covers from east 2.5 ring road to West 1 Ring Road, and from Binjiang middle road to north 3 Ring Road. The geographic range is illustrated below.

![Figure 2. Geographic Range of Dataset](image-url)
Sum of square error (SSE), which evaluates the discreteness of the cluster, is the sum of squares of the errors of the corresponding points of the fitting data and the original data, the Eq. (12) is as follows.

\[
SSE = \sum_{i=1}^{n} a \times (y_i - \hat{y}_i)
\]  

(12)

In Eq. (12), \(a\) is a constant. \(y_i\) is the original data while \(\hat{y}_i\) is the fitting data.

Silhouette Coefficient (SC) evaluates the distance between sampling points. The closer the data points in the same cluster, the further the distance between data points in different clusters, the greater the value of SC is.

\[
SC = \frac{b - a}{\max(a, b)}
\]  

(13)

In Eq. (13), \(a\) is the average distance of other samples in the same category, while \(b\) is the average distance of sample points in different categories.

3.3. Experimental Results and Analysis

The experimental results are divided into 4 sections which correspond to 4 parts of the analysis respectfully. In Section 1, we display passenger demand by counting the number of orders in each period. Section 2 proposes K-means clustering in regional planning and provides appropriate boarding points for passengers. Quantity duration of order duration is shown in Section 3, which presents an approximate Poisson distribution. In the final section, the distribution of overtravel ratio is demonstrated.

3.3.1. Order Distribution in Each Periods

In this section, the order quantity per hour on working days, holidays and weekends in 31 days of October is counted respectively, so as to obtain the order quantity - time trend chart of October which is displayed as follows:

![Order Distribution in October 2016](image-url)
Each subgraph reflects the distribution of orders for the day. The abscissa represents the time, which is evenly distributed from 0 a.m. to 24 p.m. and there are 24 corresponding intervals. The ordinate represents the normalized order quantity which relatively reflects the number of orders from the current moment to one hour after that moment.

It can be discovered from the order quantity distribution curve that the overall order quantity trend in October is roughly the same. From 0 a.m. in the morning that day, the order quantity gradually decreased, and reached the lowest point of the day around 4 a.m. or 5 a.m. in the morning. Starting at 6 a.m. in the morning, the number of orders began to rise rapidly. After ten a.m. the growth rate slowed down or the number of orders fell slightly. At around 11 a.m. the number of orders reached the first small peak of the day.

Normalized quantities of orders per hour on weekdays, holidays and weekends were averaged respectively to make the order quantity-time distribution curve, as shown in the figure below.

![Figure 4. Different Order Distribution in Weekdays, Weekends and Holidays](image)

The order distribution curve of holidays and weekends is similar, and the growth rate in the morning is slightly different. The growth rate of orders in the morning on working days is significantly higher than that on holidays and weekends, which is related to the early working hours of people's working days. The second and third peak hours of order quantity, 2 p.m. and 6 p.m., are consistent. The reasons may be attributed to the following two parts: people start to go to work at 2 p.m. in the afternoon, and office workers and students begin to go home at around 6 p.m. when the rush hour comes. After 8 p.m., the number of orders dropped significantly, which should be related to the fact that most people have already gone home. On the contrary, the number of orders on holidays from 4 p.m. to 10 p.m. is significantly higher than that on weekdays and weekends, reflecting the strong entertainment attributes in holiday nights. The above conclusions verify the stability of the order quantity data and actual behavior.

From the perspective of the driver, the distribution curve provides a clear instruction for the driver's work arrangement. Drivers can schedule their rides during the morning and evening rush hours, as well as the 2 p.m. to make sure there are enough orders to choose from and avoid the empty driving conditions. From the perspective of Didi company, this distribution can also be utilized as the guidance for
dispatching drivers, so as to arrange the number of drivers to go out of work at any periods to shorten the waiting time of passengers.

Based on the above analysis, the specific plan to optimize the working hours of Didi drivers is as follows: the number of Didi drivers should be increased after 7 a.m. on weekdays and after 9 a.m. on weekends and holidays to meet the rapidly increasing demand for commuting. The number of orders in the afternoon has been maintained at a high level, so it is necessary to arrange enough drivers in the afternoon. On weekdays, Didi drivers are allowed to take an early break in the evening, as the commuting demand drops significantly after midnight. However, in holidays, the demand for online ride-hailing is still relatively large after 22:00, so it is necessary to ensure a certain number of Didi drivers.

3.3.2. Regionalization.

We can mark the starting point of all orders on the map, and visualize the distribution of the order quantity in different regions. Meanwhile, a cluster analysis is performed on the distribution. In this section, we utilize $K$-means clustering algorithm. The results are exhibited as follows:

![Figure 5. The results of $K$-means cluster](image)

Due to the large amount of data, only the data of 20-21 p.m. on October 13, 2016 are intercepted for analysis with a total of 12538 orders. All the points are divided into 20 classes which have been marked with different colors. The geometric center of each region has been marked in blue.

It can be seen that the closer to the urban area, the denser the number of orders. The denseness of orders in the lower left corner of the figure above is much greater than that in the upper right corner, especially in the region marked by the dark red dots. And the demand for orders near the third ring road is obviously less. Limited by computer configuration, the number of clusters cannot be too large. If we can increase the number of clusters such as 500 regions, each cluster contains a smaller range, which is more suitable for starting point recommendation.

According to the results of clustering analysis, it can provide the basis of scheduling optimization scheme for the taxi platform. For example, if the supply of a certain region exceeds the demand in a period of time, the drivers in this region can be prompted to other nearby regions with more order demands. On the contrary, if the supply in a certain class is less than the demand, the redundant drivers in the adjacent area can be dispatched to reduce the transportation pressure of the region and make more profits.
3.3.3. Order Duration Distribution. Here we focus on the diversity between different orders. Based on every individual order, we define the ending point timestamp minus the starting point timestamp as the duration of an order. The order duration distribution chart is displayed below.

Researchers have found out that order duration distribution can be fitted with Poisson distribution, which has been illustrated in Eq. (8).

Suppose that the average speed of taxi driving in Chengdu’s urban region is 40 kilometers per hour. Combined with the taxi fare $y$ in Chengdu and distance of order $d$:

$$
y = \begin{cases} 
8 & d \leq 2\text{km} \\
8 + 1.9 \times (d - 2) & 2 < d \leq 10\text{km} \\
23.2 + 2.85 \times (d - 10) & d > 10\text{km}
\end{cases}
$$

(14)

We counted the income of taxi drivers. The total revenue is about 3.6537 million yuan, the average daily income of each driver is about 200 yuan.

According to the above statistical results, it can be found that passengers take taxis mainly when they are on a short journey less than 5 kilometers, and there is also a certain demand for long-distance travel. Such demand statistics results are beneficial to Didi company to better determine the ladder pricing standard and increase the income of drivers.

3.3.4. Overtravel Ratio. In Eqs. (9) and (10) the overtravel ratio $K$ is defined and amended. In this section, the clustering algorithm is used to pick up the order data randomly. The distribution of $K$ is illustrated below:
This paper utilizes a quantitative method to control the value of $K$. As the data is randomly selected, the ability of the driver to choose a route could be ignored. Hence if the value of $K$ is greater than 8, we can believe that the urban road planning in this clustered region is not very reasonable.

Regarding the $K$ value as a random probability distribution, this kind of analysis has certain significance for municipal transportation planning, and is beneficial to guide the municipal system to conduct purposeful actions for places where the $K$ value is high. The comprehensive analysis provides a definite goal, avoids the inability to start, which is of positive significance to the optimization of urban roads.

4. Conclusion

The analysis of the model is based on a whole new massive dataset of taxi trajectories in Chengdu, which guarantees the credibility and university. Furthermore, this research is one of the first studies that has presented the method to use data mining on the trajectory collected by GPS in more than one aspect. Passengers’ demand is analyzed based on a spatiotemporal sequence in October 2016. By statistics and $K$-means clustering, the passenger demand in each hour and subregion can be accurately predicted, which can provide helpful suggestions to dispatch taxis to reach a higher level of resource utilization rate. Moreover, this article also proposes a method for predicting the quantity distribution of order duration, the results of which can be used to analyze the individual drivers’ operating hours and income situation. To a certain extent, it can also avoid the occurrence of drowsy driving. In addition, we put forward a new definition of overtravel ratio, which is significant to urban construction planning and abnormal events monitoring. After standardization and normalization of the data, we can define the journey with overtravel ratio under 8 as regular ones.

Based on the current work and problems, the methodology in this paper may be limited and could be improved in future work. For example, the dataset of taxi trajectories from other cities, such as Beijing, New York, and London, could be used to ensure the accuracy of this model. Furthermore, the authors’ computing resources are so limited that the precision of the results is relatively low. With greater computing resources, the models and results can be applied to economic use. By clustering the boarding points to more than 1000 regions, the taxi online platform can dispatch the cars to reduce the empty-loading ratio.
References

[1] Kim K, Oh K, Lee Y K, Kim S H and Jung J Y 2014 An analysis on movement patterns between zones using smart card data in subway networks *International Journal of Geographical Information Science* 28 pp 1781–1801

[2] Konjar M, Lisec A and Drobné S 2020 Method for delineation of functional regions using data on commuters *Proc of the 13th AGILE International Conference on Geographic Information Science Guimarães, Portugal* pp 1–10

[3] Deng M, Peng D L, Liu Q L and Shi Y 2011 A hierarchical spatial clustering algorithm based on field theory *Geomatics & Information Science of Wuhan University* 36 pp 847–851

[4] Chen Z, Gong X and Xie Z 2017 An analysis of movement patterns between zones using taxi GPS data *Transactions in GIS* 21 pp 1341–63

[5] Kuang W, An S and Jiang H 2015 Detecting traffic anomalies in urban areas using taxi GPS data *Mathematical Problems in Engineering* pp 1–13

[6] Chandola V, Banerjee A and Kumar V 2009 Anomaly detection: a survey *ACM Computing Surveys* 41 pp 27–43

[7] Wang Y, Qin K, Chen Y and Zhao P 2018 Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data *International Journal of Geo-Information* 7 pp 25–45

[8] Hwang J R, Kang H Y and Li K J 2005 Spatio-temporal similarity analysis between trajectories on road networks *Perspectives in Conceptual Modeling* 3770 pp 280–289

[9] D’andrea E and Marcelloni F 2017 Detection of traffic congestion and incidents from GPS trace analysis *Expert Systems with Applications* 73 pp 43–56

[10] Keler A, Krisp J M and Ding L 2017 Detecting traffic congestion propagation in urban environments – a case study with Floating Taxi Data (FTD) in Shanghai *Journal of Location Based Services* 11 pp 133–151

[11] Qing C and Hao W 2018 A Methodology for Measuring and Monitoring Congested Corridors: Applications in Manhattan Using Taxi GPS Data *Journal of Urban Technology* 25 pp 59–75

[12] Schussler N and Axhausen K W 2009 Accounting for route overlap in urban and sub-urban route choice decisions derived from GPS observations *Proceedings of the 12th International Conference on Travel Behavior Research* 590 pp 35–77

[13] Li D, Miwa T, Morikawa T and Liu P 2016 Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets *Transportation Research Part C: Emerging Technologies* 67 pp 31–46

[14] Deng Y, Li M, Tang Q, He R and Hu X 2020 Heterogenous trip distance-based route choice behavior analysis using real-world large-scale taxi trajectory data *Journal of Advanced Transportation* pp 1–16

[15] Laha A K and Putatunda S 2021 Real-time location prediction with taxi-GPS data streams *Transportation Research: Part C* 92 pp 298–322

[16] Rodrigues F, Markou I and Pereira F C 2019 Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach *Information Fusion* 49 pp 120–129

[17] Atsa’am D D, Oyelere S S, Balogun O S, Wario R and Blamah N V 2021 K-means cluster analysis of the west African species of cereals based on nutritional value composition *African Journal of Food, Agriculture, Nutrition & Development* 21 pp 17195–212