K-means Clustering Method Based on Kernel Density Estimation to Analysis Residents Travel Features: A Case Study of Chengdu

Siying Li¹, Zhili Liu¹ and Hui Zhao²

¹Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Bei-jing Jiaotong University, Beijing, China
²Beijing Transportation Institute
Email: 603711005@qq.com

Abstract. In order to study the spatiotemporal features of urban centre residents during peak hours during workdays and rest days, based on taxi on and off location information and urban points of interest data, a Geographic Information System (GIS) Kernel density estimation (KED) is used. Combined with the K-means clustering algorithm, the peak hours of residents’ travel and hotspot areas for boarding and alighting are identified, and the strength of the interaction between residents in each area is analysis using the structured Georgy Voronoi, and the spatiotemporal features of residents' travel are summarized.

1. Introduction
As a primary component of transportation for urban resident daily travel, taxi travel reflects urban travel features to a certain extent. (Fu et al., 2017) GPS trajectory data includes spatial-temporal information, for example, time, start points and end points of the journey of passengers. (Ling et al., 2018) Taxi trajectory data plays an important role in recognizing the features of travel behaviour. (Li et al., 2019) Using the improved DBSCAN algorithm and spatial analysis method to mine the residents’ travel regularities in the location information. (Xiong et al., 2017) Proposing a method that using kernel density estimation for initial parameter selection and density bias sampling, which can effectively accelerate the K-means clustering process. By the spatial-temporal distribution of the trajectory points, this study confirms the existence of job–house imbalance in Chengdu. This study seeks to answer the following important questions: How to identify the spatial-temporal features of travel regularity based on trajectory? How to classify different area based on its attributes? How to evaluate traffic flow among different areas.? To answer these questions, we first extract the spatial and temporal features of transit riders in study area that can represent the regularity of travel behaviour and get the distribution of taxi hotspot areas in peak hours. These features are calculated by the kernel density estimation (KED) based on continuous passenger behaviour from one-week of taxi order data in Chengdu. A spatial clustering algorithm, combined KED with K-means, is then utilized to categorize group study area into ten cluster based on points of interest data. Next, we conduct a superimposed analysis of the divided research areas and taxi order data, the ten clusters can reflect the intensity of transit travel through the OD Expected Line Graph.
In this study, the area of Chengdu city centre is taken as the research object, and a rectangular area with a size of about 15*15 square kilometres is selected, as shown in Figure 1. Data for this study were collected from GAIA Open Dataset November 1st to 7th, 2016, a total of 1,562,945 GPS data. An example of the data is shown in Table 1 below. The basic geographic data comes from the National Administrative Division Database and Road Vector Database of Open Street Map 1:400000 scale. The point of interest data comes from GAIA Open Dataset, which mainly includes name, address, and latitude and longitude information. After filtering and processing, a total of 87,730 pieces of data are selected.

Table 1. Taxi order data

| Order ID          | Start Timestamp | End Timestamp | Pickup Position Longitude | Pickup Position Latitude | Pickup Position Longitude | Pickup Position Latitude |
|-------------------|-----------------|---------------|----------------------------|--------------------------|----------------------------|----------------------------|
| eb9dd4095d9850e628 | 1477964797       | 1477966507    | 104.09464                  | 30.70397                 | 104.08927                  | 30.65085                  |
| 7cefd813775a6c    |                 |               |                            |                          |                            |                            |
| 387a742fa5a3fbe4a1f215ac58ea33a8 | 1477985585       | 1477987675    | 104.076509                 | 30.76743                 | 104.0637                  | 30.58951                  |

2. Methodology
To date various methods have been developed to extract GPS trajectory data and analysis urban traffic hot spots, mainly including the following categories: Spatial statistical methods, sample estimation methods, complex network methods, artificial neural network methods, data cube methods, clustering methods, etc. Using the kernel density estimation for selection of initial parameters can effectively accelerate the K-means clustering process (Xiong et al., 2017).

2.1. Kernel Density Estimation Method
Kernel density estimation (KED) is a non-parametric estimation method of density-based spatial point patterns commonly used in urban travel hotspots. It estimates the density function of unknown point features based on existing point features. The kernel density value changes with the distance from the centre point. The closer the centre point is, the larger the density value is, the farther from the centre point, the smaller the density is. When the distance is equal to the bandwidth r, the density is zero (Zhao PX,2015). The function of the nuclear density shown in Eq. (1).
\[ O_i = \frac{1}{n\pi r^2} \sum_{j=1}^{n} k_j \left(1 - \frac{d_{ij}^2}{r^2}\right)^2 \]  

(1)

Where: \( O_i \) represents the kernel density of the point \( i \), \( K_j \) is the weight value of the point of \( j \), \( d_{ij} \) is the distance between the point \( i \) and \( j \). \( n \) is the number of points \( j \) within the radius of \( r \), which is centred on \( i \). Within the range of bandwidth \( r \), the weight value \( K_j \) of point \( j \) is the same.

In this study, the spatial density kernel density mapping analysis method described in spatial analysis is introduced into the location data clustering. The kernel density analysis method is used to express the trend of the spatial distribution of Chengdu residents’ travel in the form of equal density areas. Regions represent densely distributed urban populations, and the thresholds of different density levels determined by the natural breakpoint method. The spatial distribution of hot spots is obtained after all points are searched. The specific method is as follows: (1) define a fixed grid cell size, divide the entire area into countless small grids, and use it as the density calculation standard and output representation; (2) define a search radius as a continuous raster search step (3) Calculate the density value of each grid within the search radius through the kernel function; (4) Output the density value of each grid, Express different density levels with different colours, red means highest density, green means lowest density.

2.2. K-means Clustering Algorithm

The goal of K-means cluster analysis is to divide the observed objects into several sub-sets, each of which is a cluster that makes the object of the cluster object is as different as possible. The basic principle is a given data set \( x_i \) including \( n \) objects, and \( x \) is divided into \( k \) clusters, meanwhile the inner square error of the cluster is minimized, the corresponding target function shown in Eq. (2).

\[ \text{minDistance} = \sum_{i=1}^{n} \sum_{k=1}^{k} \delta_{ik} ||x_i - m_k||^2 \]  

(2)

Where, the objective function minimizes the distance from point \( x_i \) to the centre point \( m_k \) of each cluster, \( \delta_{ik} \) represents the membership of the data, if the data \( x_i \) belongs to the \( k \) cluster, the corresponding element value \( \delta_{ik} \) is 1; otherwise, the \( \delta_{ik} \) is 0.

The specific method is as follows: (1) initialize \( k \) objects; (2) randomly specify \( k \) points as the centre point of the initialization class; (3) for each sample, assign it to the cluster centre closest to it; (4) re-calculate various types of centres; (5) use the new class centre, go to step (2) and calculate the error; (6) if the error value converges, return to the class centre and terminate the algorithm, otherwise return to step (3).

2.3. K-means Clustering Based on Kernel Density Estimation

As shown in Figure 2, the more concentrated the original object, the larger the value of the density function. At the same time, the maximum value of the density function is also very close to the centre of the original object's concentrated area. Therefore, in order to reduce the number of iterations of K-means clustering and improve the efficiency of the algorithm, the maximum point of the kernel density function can be selected as the initial iteration centre of K-means.

![Figure 2. Kernel density estimation map](image-url)
3. Analysis of the Spatial-temporal Features

3.1. Time Distribution Features

According to the statistical analysis results within a week, taking the X-axis as time and the Y-axis as travel volume, the trend of time-share travel volume during the week is shown in Figure 2.

![Figure 3. Time distribution of taxi travel volume](image)

On the whole, the distribution trend of taxi trips is roughly the same throughout the week, and there are slight differences between weekdays and rest days. Compared with the rest days, the features of morning peaks are more obvious. The night peaks of rest days are more obvious than workdays. On the total number of sunrise trips, the rest days are higher than the workdays. As can be seen from Figure 3, the workdays show obvious peaks in the morning, noon, and evening on the time-shared trips, and the morning peak hours is from 8:00AM to 10:00AM, 12:AM to 2:00PM noon peak hours, 4:00PM to 6:00PM evening peak hours. The number of taxi trips during the morning rush hours is 31,341~35,961, the number of trips during the noon peak is 40823~45104, and the number of trips during the late rush is 38,501~44,317. Among them, Monday, Wednesday and Thursday are relatively stable, and the morning peak rental The number of car trips is the lowest and the highest on Friday; the number of morning peak taxi trips on rest days is 25,446 to 30,947 trips, which is 14% to 19% lower than the weekday morning peak trips, and the number of afternoon peak trips is 44,366 to 46,440 trips. The weekday peak increase of 3% to 8%, the evening peak trip volume is 43,583 to 47,327 trips, compared with the weekday peak increase of 8% to 13%, the average length of taxi trips is 15 minutes, and the commute time is mainly distributed at 10~20 minutes.

3.2. Spatial Clustering Feature

This article chooses two periods of morning and evening peaks corresponding to 8-10 o'clock, and 16:00-18 o'clock. Through the Spatial Analysis function in ArcGIS 10.2 software, using the KED to analysis the hotspot areas of residents during peak hours. After repeated experiments, the search radius was determined to be 500 meters and the pixel size was 100 meters. The hotspot areas are shown below. Red means the hottest and green means the lowest. The higher the hotspot level, the greater the amount of travel. The distribution and statistical results of hot spots are shown below.
Figure 4. Distribution map of pick-up points during morning peak hours on weekdays

Figure 5. Distribution map of morning peak drop-off points during morning peak hours on weekdays

Figure 6. Distribution map of pick-up points during evening peak hours on weekdays

Figure 7. Distribution map of morning peak drop-off points during evening peak hours on weekdays

Figure 8. Distribution map of pick-up points during morning peak hours on weekend

Figure 9. Distribution map of morning peak drop-off points during morning peak hours on weekend
In this study, using K-means clustering algorithm and the kernel density estimation method, and 10 cluster centres are obtained through repeated experiments, as shown in Table 3.

### Table 3. Cluster centre coordinates

| Cluster | Longitude  | Latitude  |
|---------|------------|-----------|
| cluster1 | 104.1402   | 30.6238   |
| cluster2 | 104.077    | 30.6148   |
| cluster3 | 104.0146   | 30.6365   |
| cluster4 | 104.0855   | 30.6995   |
| cluster5 | 104.1018   | 30.6437   |
| cluster6 | 104.0341   | 30.6723   |
| cluster7 | 104.0789   | 30.6595   |
| cluster8 | 104.1181   | 30.6743   |
| cluster9 | 104.0544   | 30.6355   |
| cluster10| 104.0307   | 30.7037   |
3.3. Spatial Interaction Features

Based on the analysis of travel laws from the time and space distribution dimensions, this study further excavates the dynamic laws of travel features from the spatial interaction dimension, and uses Georgy Voronoi to divide the study area into 10 regions based on the features of the cluster centre. The feature is that any position in the polygon is closest to the centre point of the polygon, and is far away from the centre point of the adjacent polygon. This paper selects the time period from 7 am to 9 am on weekday mornings to conduct a statistical analysis of the inflow and outflow of taxis in each area during the statistical period. It is found that the area is rented in three areas: 1, 4, and 9 The number of vehicles has decreased, and the distribution of these areas is dominated by residential areas. The number of taxis has increased in 7 areas of areas 2, 3, 5, 6, 7, 8, and 10. These 6 areas are distributed by employment land, business, etc. Area-based. The following figure shows the OD expectation line chart, which shows the direction of the interaction between residents' travel activities and the intensity of the interaction, which can roughly reflect the spatial flow features of residents' travel. It can be seen that the intensity of spatial interaction between Area 7 and Area 4 is the greatest, and residents' travel needs are the greatest. Therefore, urban management should consider strengthening the construction and management of public transportation in this area to improve the convenience of residents' travel (Li H, et al., 2019).

![Figure 12. OD Expected Line Graph](image)

4. Discussion

We further analysis travel behaviour in terms of their departure time, end time, travel time, traffic volume by time (Figs. 3) and the distribution of hotspot areas (Figs. 4 to 11).

The most of transit commuters depart from their homes around morning peak hours (8:00 AM-10:00 AM) and return during evening peak hours (4:00 PM-6:00 PM). In terms of traveling days in the one-week period, the most frequent number of traveling day for passengers is weekday, and the most obvious commuting characteristics, the lowest frequent number of traveling day for passengers is weekends.

Simultaneously, according to departure time, end time and travel hotspot areas, we recognize the place of residence and work. From Fig.4 to Fig.11 are heat maps in which the region with a dark color (from light green to dark red) implied high travel population density. In terms of workplaces, most commuters work in the core of Chengdu CBD, ChunXi Road (CXR), TaiGuLi (TGL), WanDa square (WDS), SM Square (SMS) and YinHai Centre (YHC). CXR and TGL are the key international financial and business centres in Chengdu.
5. Conclusions
In summary, this study proposes a series of data mining methods to identify analysis residents travel features based on taxi order data and the points of interest. The proposed method can recognize residences by mining spatial-temporal travel regularities over continuous travel behaviour, as well as extract the distribution of residence and workplace. This approach will significantly alleviate the burden of data collection and improve the efficiency of recognition.

1) The average travel time of a taxi is 15 minutes, and the commute time is mainly distributed between 10 and 20 minutes. The demand for taxi travel is relatively stable within a week. The trend of the travel volume on each day of the week is basically the same, showing three distinct peak hours, but the working days and rest days are slightly different, and the performance is in the morning peak hours of working days. Larger demand, wider spatial distribution, and higher heat levels.

2) The morning peak travel time is concentrated and the late peak travel time is scattered. The hot spots for getting on the bus during the morning peak are all residential areas, and the hot spots for getting off the bus during the late peak are distributed in residential and commercial areas, which better reflects the daytime Commuting travel features; the hot spots for getting on the bus during the morning rush hours are widely distributed and hot. Most of the pick-up and drop-off locations during the rush hours are in shopping malls and scenic spots, which reflects the strong demand for people's leisure and leisure travel. The findings of this study have a number of important implications for future practice, which guide the balanced distribution of work and residence to avoid excessive commuting.

3) This study explores the spatial interaction features of residents' travel by constructing the traffic OD expectation line map, and more intuitively discovers the spatial regularity of residents' travel. At present, in addition to taxis in cities, there are subways, buses, and shared bicycles. However, this article only studies taxi trajectory data and city points of interest data. Because the data source for this study is single, further research should be done to investigate multiple data sources to recognize the regularity of resident travel behaviour.

6. References
[1] Xin Fu, Maopeng Sun and Hao Sun. 2017. Taxi commuting recognition and analysis of space-time characteristics based on GPS data[J]. Journal of China Highway, 30(07): 134-143. (in Chinese)
[2] Tao Ling. 2018, City hot spot analysis based on POI data [D]. Kunming University of Science and Technology. (in Chinese)
[3] Hao Li, Xuzhi Wang and Wanggen Wan. 2019. Study on the Spatio-temporal Characteristics of Resident Trips Based on Location Data——Taking Shanghai as an Example [J] . Electronic Measurement Technology ,42 (19): 25-30. (in Chinese)
[4] Pengxiang Zhao. 2015. Research on extraction and analysis of urban hotspots based on trajectory clustering [D]. Wuhan University. (in Chinese)
[5] Kailing Xiong, Junjie Peng, Xiaofei Yang and Jun Huang. 2017. K-means clustering optimization based on kernel density estimation [J]. Computer Technology and Development, 27 (02): 1-5. (in Chinese)