Research on traffic zone partition method based on two-level partition theory

Yongchang Li¹ and Xuedong Yan¹*

¹MOT Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing, 100044, China

*Corresponding author’s e-mail: xdyan@bjtu.edu.cn

Abstract. As an important part of traffic planning, traffic zone division has drawn much attention for decades. Better than traditional methods of traffic zone division based on traffic surveys and road network structures, this paper proposes a two-level framework for dividing traffic zones by using massive car-hailing data from a big data perspective. Firstly, the grid model is established within the study area and massive travel data are matched into grids. Secondly, the Louvain community detection algorithm is used to divide the research area into different traffic medium zones (TMZs). Further, a K-prototype clustering algorithm based on partition is used to divide each TMZ into several traffic small zones (TSZs). Taking the Beijing Fifth Ring as an example, the study area is divided into 16 TMZs and 302 TSZs by using two-level partition theory. The proposed method of dividing traffic zones is proved to be efficient and accurate and can be easily applied to other areas, which helps planners to make more smarter traffic planning and city management.

1. Introduction

A complete urban traffic system is very large and complex, which makes the analysis and solution of traffic problems more difficult and complex. This complexity can be effectively reduced by dividing a complete urban traffic system into several traffic zones according to the characteristics of traffic flow direction and road network layout[1]. Domestic and foreign scholars have done a lot of research on the division of traffic zones. The research of foreign scholars includes the genetic algorithm proposed by F.Bacao[2] and R.M. Assuneao, which uses graph theory to divide social and economic zones[3]. Domestic scholars often use various clustering algorithms when dividing traffic zones. Song Liang used the method of fuzzy cluster analysis to divide traffic zones[4]. Cai Yi adopted K-means clustering method based on partition [5]. Zhu Qinpeng used IsoLPC community detection algorithm to divide traffic zones[6]. In addition, Xu Fang and Dai Bingkui proposed two-level partitioning of traffic zones, the core idea of which is to divide the study area into larger areas, and then subdivide the larger areas into smaller areas[7-8].

In this study, the idea of two-level partition theory is used for reference, and a grid model is established within the study area. Using floating car OD data, the research area is first divided into TMZs by Louvain community discovery algorithm, and then each TMZ is subdivided into several TSZs by K-prototype clustering method. The whole research area is divided into TMZs and TSZs in two steps. Compared with the traditional traffic zone partition method, this method has the advantages of high efficiency and accuracy, and it is based on the grid model and is not affected by administrative boundaries, so it can be easily extended to other areas.
2. Data description
The data used in this study are ride orders of Di-Di Express that occurred within the Beijing fifth ring for one week from June 5, 2017 to June 11, 2017. The original data contains many attributes such as starting and ending points.

The data are analyzed by nuclear density analysis, and the results are shown in figures. Figure 1 is the distribution of the starting points (O points) of orders in the early peak period, and Figure 2 is the distribution of the arrival points (D points) of orders in the early peak period.

![Figure 1. Distribution of Early Peak O Point.](image1)

![Figure 2. Distribution of Early Peak D Point.](image2)

In figures, green represents less order volume, while red represents more order volume. As can be seen from figures, the early peak O points mainly distribute in some residential areas, such as Guangyuan District near Beijing West Railway Station, Suojiafen District near Xizhimen, and residential groups around Dajiaoating, etc. The early peak D points mainly distribute in some business circles and science and technology parks, including Fengtai Science and Technology Park, railway station, Financial street, Zhongguancun, International Trade, Wangjing and Wangfujing. The OD aggregation points of early peak travel accord with people's cognition, which shows that the data is real and effective.

3. Traffic zone division method
This study divides traffic zones by two steps. The first step is to divide the whole research area into different TMZs by community detection method. Grids are then clustered in each TMZ and TSZs are subdivided in each TMZ. The advantage of using two-level partition is that the community detection algorithm is very fast for large-scale network and the result of the TMZs is good. Because each TMZ is small, K-prototype clustering not only has faster calculation speed, but also fully considers each grid attributes to make clustering more accurate.

3.1. Grid model establishment
Traffic grid model is derived from urban management grid modeling theory[9]. This study focuses on the OD point information inside the grid, so the study area is statically divided into grids of the same size. The final study area is the smallest rectangular area including the Beijing Five Rings. The latitude and longitude range of the rectangular is approximately 116.198432-116.546309 degrees east longitude and 39.753886-40.028000 degrees north latitude. The research area is divided into 100*100 grids.

3.2. TMZ partition based on Louvain algorithm
In this study, a heuristic algorithm based on modular optimization is used to quickly extract community structure in the network, Louvain algorithm[10], which divides the grid in the research area into different TMZs. The process of the Louvain algorithm is as follows.

Step 1, each node in the network is in an isolated community.
Step 2, randomly select one node from all nodes and perform algorithm steps 3 to 4.
Step 3, for node i, find all its neighbor nodes, and calculate the size of the module gain $\Delta Q$ generated
if the node $i$ is moved from its current community to the community $C_j$ of its neighbor node $j$. Its calculation formula is as follow.

$$\Delta Q = \left[ \frac{\sum_{in} + k_{i,in}}{2M} - \left( \frac{\sum_{tot} + k_i}{2M} \right)^2 \right] - \left[ \frac{\sum_{in}}{2M} - \left( \frac{\sum_{tot}}{2M} \right)^2 - \left( \frac{k_i}{2M} \right)^2 \right]$$

(1)

Among them, $\sum_{in}$ is the sum of the weights of the inner edges of the community $C_j$, $\sum_{tot}$ is the sum of the weights of the connecting edges of the nodes in the community $C_j$. $k_i$ is the sum of the weights of the connecting edges of the node $i$, $k_{i,in}$ is the sum of the weights of the edges of the node $i$ and the inner edges of the community $C_j$, and $M$ is half of the sum of the weights of all the edges of the network.

Step 4, find the neighbor node $j'$ that produces the maximum modularity gain. If the maximum modularity gain $\Delta Q_{max}$ is greater than zero, let $c_i$ equal to $c_{j'}$, that is, move the node $i$ to the community where the node $j'$ is located.

Step 5, when all nodes cannot be moved, it indicates that the community partition has reached the optimal level at the current scale, and the network is aggregated to generate a new network. All nodes in the same community are mapped to one node in the new network, called super node; the internal connection of the community is mapped to the self-edge of the super node in the new network, and the weight is the sum of the internal edge weights; The weight of the connected edges between the nodes is the sum of the weights of the connected edges between the corresponding communities.

Step 6, after the new network is built, skip to step 1 and iterate over the calculation. The algorithm terminates until all nodes cannot be moved during an iteration.

3.3. TSZ partition based on K-prototype clustering algorithm

K-Prototypes algorithm[11] is an effective algorithm for clustering mixed numerical attribute and classified attribute data. The algorithm combines K-means with K-Modes algorithm, and controls the weight of numerical and classification attributes in clustering process by parameter $\gamma$. The formula for calculating the degree of dissimilarity is used to calculate the distance between the sample and the prototype. The distance between the sample $X_i$ and the prototype $V_i$ is defined as follows.

$$d(X_i, V_i) = \sum_{j=1}^{p} |x_{ij} - v_{ij}|^2 + \gamma \sum_{j=p+1}^{m} \delta(x_{ij} - v_{ij}) = d_1(X_i, V_i) + \gamma d_2(X_i, V_i)$$

(2)

Among them, $d_1(X_i, V_i)$ is the distance of numerical attributes, which is measured by the square of Euclidean distance; $d_2(X_i, V_i)$ is the distance of classification attributes, which is measured by Hemingway distance formula and $\gamma$ is the weight value of classification attributes.

The process of the K-prototype clustering algorithm is as follows.

Step 1, we select $k$ initial cluster prototypes from the dataset.

Step 2, we calculate the distance of the sample to each prototype according to the dissimilarity formula and divide it into clusters represented by the clustering prototype closest to it.

Step 3, for each cluster, we recalculate the clustering prototype according to the center adjustment algorithm.

Step 4, we calculate the distance of each data for the new cluster prototype. If the clustering prototype closest to one data is not the prototype of the cluster to which the current data belongs, we will reassign the objects of the two clusters.

Step 5, we repeat step 3 and 4 until there is no more data change in each cluster or it reaches the preset maximum number of iterations.

4. Algorithmic implementation and partition result

4.1. TMZ partition algorithm implementation and partition result
4.1.1. Implementation. The data required for TMZ partitioning includes point set data and edge set data, and the two types of data need to be extracted based on the original order data. The point set data includes grid number $g_i$, grid center point longitude $g_{i,lon}$, grid center point latitude $g_{i,lat}$; the edge set data includes start grid number $g_i$, end grid number $g_j$, and the ratio of the traffic volume between two grids to the total traffic volume $W_{i,j}$, which is calculated as follows.

$$W_{i,j} = \frac{q_{i,j}}{\sum_{i=1,j=1}^{ij} q_{i,j}}$$

(3)

Where $W_{i,j}$ represents the weight between node i and node j; $q_{i,j}$ represents the sum of the order quantities between the grid numbered i and the grid numbered j.

4.1.2. Partition result. After Louvain algorithm calculation, the nodes with high degree of connection are clustered to get different TMZs. The nodes and links are expressed in different colors to represent different TMZs, as shown in Figure 3. From the figure, we can see that different TMZs have typical location characteristics, such as Wudaokou, International Trade, Sanlitun, Xidan, Wangjing and so on. The TMZs are basically centered on these well-known business circles. According to the calculation results, the study area is divided into 16 TMZs, numbered 1-16, as shown in Figure 4.

4.2. TSZ partition algorithm implementation and partition result

4.2.1. Implementation. In this study, K-prototypes clustering is used to cluster TSZs in each TMZ.

(1) Clustering attribute

The input data include grid number $g_i$, grid center longitude and latitude $g_{i,lon}$ and $g_{i,lat}$, grid OD number $q_{io}$ and $q_{id}$, grid land attribute $g_{i,land}$. The longitude, latitude and OD number can be calculated and counted directly. Land use attributes are determined by dividing the number of O points by the number of D points in the grid during the early peak period. The formula for calculating land use attributes is as follows.

$$R_i = \frac{q_{i,o,morn}}{q_{i,d,morn}}$$

(4)

Among them, $R_i$ denotes the ratio of the number of O points to the number of D points in the early peak period of grid i, $q_{i,o,morn}$ denotes the number of O points in the early peak period of grid i, and $q_{i,d,morn}$ denotes the number of D points in the early peak period of grid i. $R_i$ greater than 1 indicates that there are more orders from the grid in the morning than arrive in the grid, and the grid can be defined as residential land attribute. $R_i$ less than 1 indicates that there are fewer orders leaving the grid in the morning than arriving in the grid, and the grid can be defined as the business land attribute. The grid with $R_i$ equal to 1 defines its land use attribute as mixed land use attribute.
(2) Attribute weight

There are three kinds of attributes in the K-prototype algorithm used in this study. The numerical attributes include the longitude and latitude of grid center, the number of grid OD, and the classification attribute is grid land attribute. Firstly, the average distance from all samples to each clustering center of each attribute in the initial state is calculated, and the dissimilarity distance of each attribute is divided by the corresponding average distance. In order to make the clustering results more aggregated in space, the weight coefficient theta is added to the longitude and latitude attributes in this study. The improved formula for calculating dissimilarity distance is as follows.

\[ d(x_i, V_j) = \text{theta} \left( \frac{1}{\text{Dis}_{\text{lonlat}}^{\text{aver}}} \left( \sum_{j=1}^{2} |x_{ij} - V_{ij}|^2 \right) + \left( \frac{1}{\text{Dis}_{\text{OD}}^{\text{aver}}} \right) \left( \sum_{j=3}^{4} |x_{ij} - V_{ij}|^2 \right) + \cdots \right) \]

Among them, the first and second attributes are longitude and latitude attributes, the third and fourth attributes are OD quantitative attributes, the fifth is land use attribute. \( \text{Dis}_{\text{lonlat}}^{\text{aver}} \) denotes the average longitude and latitude distance of the grid, \( \text{Dis}_{\text{OD}}^{\text{aver}} \) denotes the average distance of the OD number of the grid, \( \text{Dis}_{\text{land}}^{\text{aver}} \) denotes the average distance of the land use attribute of the grid, and theta denotes the added weight coefficient of longitude and latitude.

4.2.2. Partition result. In this study, TSZs obtained by K-prototype clustering method are shown in Figure 5. Different TMZs are differentiated by different colors. The interior of each TMZ is subdivided into TSZs according to the principle of similar attributes. The number of TSZs in each TMZ is shown in Table 1. There are altogether 16 TMZs and 302 TSZs in the whole research area.

| TMZ | Area (km²) | Number of cells included |
|-----|------------|--------------------------|
| 1   | 109.53     | 43                       |
| 2   | 71.91      | 28                       |
| 3   | 43.83      | 17                       |
| 4   | 64.71      | 25                       |
| 5   | 45.99      | 18                       |
| 6   | 34.38      | 13                       |
| 7   | 19.17      | 7                        |
| 8   | 19.26      | 7                        |
| 9   | 14.49      | 5                        |
| 10  | 111.24     | 44                       |
| 11  | 28.53      | 11                       |
| 12  | 16.56      | 6                        |
| 13  | 48.24      | 19                       |
| 14  | 27.45      | 10                       |
| 15  | 75.69      | 30                       |
| 16  | 49.50      | 19                       |
| sum | 780.48     | 302                      |

Figure 5. Distribution of TSZs in each TMZ.

Traffic zone is the basic space unit to study the generation and distribution of traffic, so the division of traffic zone is the basic work of traffic investigation and planning[12]. The most important principle of traffic zone partition is homogeneity, that is, the characteristics of land use, population, economy and society in zones should be as consistent as possible[13]. The clustering method used in this study is to cluster the similar grid into the same TSZ, which conforms to the principle of traffic zone partition.

5. Conclusion

This paper applies the idea of two-level partition theory, uses OD travel data of floating car, and builds grid model in the research area. Louvain community detection algorithm and K-prototype clustering algorithm are used to divide the research area into TMZs and TSZs. According to the above method, the Beijing Fifth Ring Area can be divided into 16 TMZs and 302 TSZs. Compared with the traditional method, this method is fast and efficient, and the grid model can be easily applied to other areas.
The contributions of this study are as follows.

1. The grid model is established, and the grid is taken as the basic analysis unit. The complex map matching algorithm and high precision vector map are avoided, which make the division of traffic zones easier to operate.

2. The grid model contains a large number of grids, so community detection algorithm is used to divide the TMZs. For extremely complex network maps, community detection methods have great advantages in computing accuracy and speed.

3. K-prototype clustering based on the divided TMZs can not only greatly improve the speed and accuracy of clustering, but also fully consider the attributes and weights of the grid, making the results of clustering TSZs more accurate and effective. Moreover, this clustering method can cluster a large number of mixed attribute variables and has the characteristics of easy expansion. The research method in this paper can provide theoretical reference for other scholars.

4. In this paper, the idea of two-level partition theory is extended, and the results of two-level partition of TMZs and TSZs are obtained. In future research, different levels of traffic problems can be analyzed according to different research contents. For example, when analyzing the regional traffic operation state, we can use the TMZ as the analysis unit; when forecasting the traffic volume, we can use the TSZ as the analysis unit.

References

[1] Yuqiang Lv, Yong Qin, Limin Jia. (2010) Research on dynamic partition method of traffic district based on taxi GPS data clustering analysis. Logistics technology, 29 (9): 86-88.

[2] Bação F, Lobo V, Painho M. (2005) Applying genetic algorithms to zone design. Soft Computing, 9(5):341-348

[3] Neves M C. (2006) Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7):797-811.

[4] Liang Song. (2011) Theoretical Analysis and Classification Method of Transportation District. Chang'an University.

[5] Yi Cai. (2013) Division of Nanjing traffic district based on taxi GPS data. Science and Technology Innovation Report, (23): 218.

[6] Qinpeng Zhu. (2017) Laplacian Central Peak Clustering Algorithm and Its Application in Traffic Zone Division. Zhejiang University of Technology.

[7] Fang Xu, Chunhua Qian. (2017) Urban Loop Traffic Zone Division Based on Two-level Zoning. Forestry Construction, (3): 74-77.

[8] Bingkui Dai, Ji Hu. (2010) Research on double-deck division method of regional traffic district. Railway transportation and economy, 32 (11): 86-89.

[9] Xinjuan Zhao. (2006) Research on model of resource management for grid based on computational economy. Computer Engineering & Design, 6-7:1199-1202.

[10] Blondel V D, Guillaume J L, Lambiotte R. (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics, 2008(10):155-168.

[11] Wei Chen, Lei Wang, Ziyun Jiang. (2010) Hybrid attribute data clustering algorithm based on K-prototypes. Computer application, 30 (8): 2003-2005.

[12] Qiaofeng Guo. (2010) Brief analysis of traffic district division. Scientific and technological innovation, 2010(28): 270.

[13] Chunfu Shao. (2004) Traffic Planning Principle. China Railway Publishing House, Beijing.