IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 9, 2020, accepted March 29, 2020, date of publication April 2, 2020, date of current version April 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985092

Multi-Mode Traffic Demand Analysis Based on Multi-Source Transportation Data

DAWEI LI[1,2], YUXIANG TANG[1], AND QIONG CHEN[3]
[1]School of Transportation, Southeast University, Nanjing 210096, China
[2]Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing 210096, China
[3]School of Architecture, Southeast University, Nanjing 210096, China

Corresponding author: Yuxiang Tang (tangyuxiang@seu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1600900, in part by the National Nature Science Foundation of China under Grant 71971056 and Grant 51608115, in part by the Six Talent Peaks Project in Jiangsu Province XNYQC-003, in part by the Open Project of the Key Laboratory of Advanced Urban Public Transportation Science, Ministry of Transport, China, in part by the Research Grants from the Research Grants Council of the Hong Kong Special Administrative Region under Project PolyU 15212217, and in part by the Hong Kong Scholars Program under Project G-YZ1R.

ABSTRACT Automatic vehicle identification (AVI) data, Integrated Circuit (IC) card data and Global Positioning System (GPS) data offer an emerging and promising source of information for analysis of traffic problems. Research on insights and information from AVI data for transport analysis has made little progress in developing specific applications especially. The emergence of multi-source data provides us with a new perspective for multi-mode transportation. This paper proposes a multi-mode traffic demand forecasting method based on AVI data, metro IC card data, and taxi GPS data. The paper extracts traffic origins and destinations (OD) information of travelers from the multi-source data and uses the extracted data for traffic zone division. Finally, a multi-mode traffic forecasting model is established on this basis. GPS data of taxi trips are selected as the clustering data and k-means algorithm was adopted to divide traffic zones in Shenzhen. Moreover, the research applies the principle of convex hull to outline the boundary of the cell. Additionally, this paper establishes the multi-mode transportation forecasting model by integrating the correlation between various transportation modes into the deep learning model for prediction. The results show that the multi-mode demand forecasting model has higher accuracy and better forecasting results comparing it with the single-mode demand forecasting model which is referring to the conventional four-step procedure. The result demonstrates that effective traffic and travel data can be obtained from multi-source data, providing an opportunity to improve the analysis of complex travel patterns and behaviors for travel demand modeling and transportation planning. Furthermore, the substantive contribution of this research is that it provides strong empirical evidence for the existence of correlation among multi-mode travels and travel demand.

INDEX TERMS Multi-mode transportation, multi-source data, k-means algorithm, multi-mode traffic demand forecasting.

I. INTRODUCTION

The traditional traffic demand analysis model is usually based on a single traffic pattern. With the increasing diversification of traffic demand, urban traffic planning is gradually transitioning from a single mode to a multi-mode transportation system that supports and influences each other [1]. The research on multi-mode traffic demand analysis is not only to explore the correlations among different transportation modes and the characteristics of traffic demand, but also to establish a foundation for the implementation of transportation planning and management based on multi-mode traffic.

With the diversified development of transportation modes, the structure of urban travel is increasingly complicated and diversified. Multi-mode transportation has become a hot research field. However, the analysis of multi-modal traffic demand has several major challenges.

1) With aspect of data, most data used in past research were from travel diaries, face-to-face interviews, and telephone interviews. Li *et al.* [2] pointed these data

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Xia.

are not accurate, it may lead to underestimate of travel because respondents may forget some travels.

2) Data sparsity is a problem. The single data source cannot provide the accurate information and even sometimes will miss value. For example, not all vehicles will contribute GPS data. In a period, only a few vehicles (such as taxis) on the road will contribute data. How to estimate the travel time of the road without GPS data coverage is a big difficulty [3].
3) The single-mode traffic demand forecasting method cannot adapt to the modern multi-mode transportation system, and the single data source cannot fully reflect citizen travel characteristics, how to collaborative computing of heterogeneous data is a new topic. Traditional machine learning problems are difficult to adapt to the complex modern traffic conditions, such as Natural Language Processing mainly analyzes text data, and image vision mainly based on image data [4], [5].

A. RELEVANT STUDIES

AVI technology, GPS, and radio-frequency-based technologies have made it convenient to collect real-time traffic information from a wide range of advanced sensor sources. This allows advanced traffic pattern analysis and forecasting for important transportation management and planning [6]–[11]. However, in the big data era, traffic information extraction is a biggest roadblock for analyst to promote traffic information service.

Dong *et al.* [6] pointed new traffic data acquisition methods could be roughly classified into four categories:

1) Sensor-based traffic information collection technologies, such as inductive loop detectors, microwave radar, laser radar, magnetometers, and ultrasonic detectors [6].
2) Video-based traffic information collection technologies, such as license plate recognition (LPR) systems [12], [13].
3) Radio-frequency-based traffic information collection technologies, such as radio frequency identification (RFID) dedicated short range communications (DSRC) and Call detail record (CDR) [6], [14], [15].
4) Location-based traffic information collection technologies, such as GPS-equipped floating cars [16]–[18].

AVI data, GPS data, and IC card data offer an emerging and promising source of traffic information for researchers to analyzing traffic phenomena [15]. AVI data, GPS data, and IC card data cover large geographic areas, yield large sample sizes. Therefore, many traffic information can be acquired in real-time. Their positioning accuracy is especially appropriate for the analysis of commute traffic [19]. Alvarez-Garcia *et al.* [20] proposed a forecasting model of line end point based on Hidden Markov chain by using GPS data. Papinski *et al.* [21] used GPS data to solve the problem of personal travel path selection. Castillo *et al.* [22] proposed an optimization method for the reconstruction and forecasting of travel matrix by using AVI data. Zhan *et al.* [15] used IC smart card to identify the travel pattern of citizens.

Collaborative computing of heterogeneous data can be used to identify OD and travel trajectories of commuters. They can also be used to divide Traffic Analysis Zones (TAZ) and extract activity data needed for trip generation and trip distribution. These are the first two steps of the conventional four-step approach which is used to build travel demand forecasting models for transportation management and planning. The conventional traffic zone division approach will apply geographical information, land use characteristics, economic characteristics, and social characteristics to divide the research area into many zones. Traffic division approaches are generally classified into following several approaches: graph theory [23]; mathematical programming [24]; graphical and numerical simulation [25]; clustering [6], [26] and so on.

These approaches have its own characteristic. For example, the graph theory approach depends on building a subgraph for traffic zone division, however, it is too difficult to solve and explain this problem. For these division approaches, they have a common dilemma for travel demand forecasting which is obtaining travel data for trip generation and distribution. Travel surveys need to take a lot of resources to carry out. Additionally, data from travel surveys have limited samples and may be inaccurate. However, Multi-source data cannot only provide a significant opportunity to improve the state of practice, but also make up for the problem of data sparsity mentioned before.

Furthermore, Traffic demand forecasting is an important part of traffic planning. The design of urban road network, the design of urban road network for traffic policy-making, and the management of traffic policy-making system are closely related to traffic demand forecasting.

Seyedabrishami and Shafahi proposed and applied the expert knowledge-guided algorithm to integrate knowledge into the fuzzy inference system of adaptive network to deal with the uncertainty of demand estimation [27]; Vliet *et al.* [28] proposed a new traffic demand forecasting approach which applied more traditional data collection approaches and make traffic demand forecasting be more effective; Zhou *et al.* [29] pointed out the important role of the back propagation network (BPN) in the travel demand forecast and analysis. They proposed new algorithm for conventional four-step procedure. However, these approaches cannot solve the challenges mentioned in Section I. They are limited to solve the problem of single mode traffic and the single data source cannot fully reflect commuters travel characteristics.

B. OBJECTIVES AND CONTRIBUTIONS

To date, research on insights and information to be gleaned from multi-source data for transportation analysis has been slow, and there has been little progress on development of specific applications. This research preprocessed AVI data, GPS data, and IC card data in Shenzhen, and extract travel related information of different traffic modes to identify the travel spatio-temporal pattern of individual. Moreover, the research used an unsupervised classification approach to extract OD matrix for different modes of transportation. Finally, this paper also proposes a deep learning approach to predict the traffic demand in an aggregation way and was compared with single sparse data, which provides a new perspective for multi-mode traffic demand forecasting.

II. PROBLEM AND METHODOLOGY

For analyzing travel spatio-temporal pattern of individual, a primary task is to divide travel zones which is an important step in four procedure steps, however, dividing travel zones from multi-source data is a challenge in the following two aspects: 1) multi-source data include the data of travelers when they stay or move with all kinds of travel modes, e.g., taxi,bus, subway and private car. Vehicular trips must first be imputed from the raw data for further travel behavior analysis; 2) multi-source data are collected from multiple applications with different sample rates and at low-resolution. This hinders dividing travel zones. Once the different trip patterns are analyzed, the research then selects high resolution trips for dividing travel zones by k-means algorithm. Furthermore, the research builds a multi-mode transportation neural network to compare with the single mode transportation neural network and the result is used to provide strong empirical evidence for the existence of correlation among multi-mode travels and travel demand.

A. VEHICULAR TRIPS DETECTION

It is more complex to extract individual travel OD from IC data and AVI data in all data. Once the transaction type of IC card is swiped, it can be determined whether IC extracts data from subway or bus, and extract individual travel OD record to map station data crawled from AMAP. For aspect of AVI data, the research suppose that a user starts a new trip if there is no record for at least 7440 seconds before the current one, that is, $t_{current} - t_{previous} \geq$7440 seconds. Furthermore, the research drops out the trips with records. At this point, the records of each user have been partitioned into a sequence of trips, Finally, the research extract travel OD from it.

For aspect of GPS data, a number of methods have been proposed to derive the travel mode from trajectory data, most of them process the high-resolution GPS traces or utilizing sophisticated learning methods which require gold labels of travel modes for model training. For simplicity, the research identifies the trip as a taxi trip if its average speed is between 20 km/h and 100 km/h.

B. TRAFFIC ZONE DIVISION

Enriching the information of multi-source data help us to comprehensively understand the travelers travel behavior and characteristics. According to the analysis of multi-source data, this research selects high-resolution travel mode data to divide traffic zones. In order to make full use of the advantages of big data to divide multiple traffic zones with appropriate size, the research choose k-mean clustering method, and use the principle of convex hull to divide the boundary of the cell. Finally, the research marks the serial number for it.

C. THE EXISTENCE OF CORRELATION AMONG MULTI-MODE TRAVELS

On the basis of the divided traffic zones, the social and economic data of each traffic zones, such as GDP, population, number of students and employment, are used to predict the generation and attraction of travelers in each traffic zones. The travel demand of each traffic area is predicted through the combination of the traffic impedance between each traffic zones and the travel generation and attraction of each traffic area. The commonly used formula is as follows:

$$A_i = \left[\sum_j B_j \cdot U_j \cdot f(d_{ij})\right]^{-1} \tag{1}$$

$$B_j = \left[\sum_i A_i \cdot T_i \cdot f(d_{ij})\right]^{-1} \tag{2}$$

$$X_{ij} = A_i \cdot B_j \cdot T_i \cdot U_j \cdot f(d_{ij}) \tag{3}$$

where X_{ij} is travel demand from zone i to zone j, T_i is total amount of travel generation in zone i. U_j is total amount of travel attraction in zone j. A_i and B_j are operational parameters. d_{ij} is traffic impedance. $f(d_{ij})$ is traffic impedance function. a is model parameter.

The traditional method of trip distribution prediction mainly selects the generation, attraction and impedance between zones as the independent variables of the traffic demand forecasting model. Moreover, it is only limited to the forecasting of a single traffic mode. Vliet *et al.* [28] have pointed out that the accuracy of traffic forecasting can be improved by adding multi-modes transportation data. Although there is a certain correlation between travel demand and a variety of travel modes in reality, these conclusions have not been verified. This research uses the traditional traffic zones demand forecasting structure for reference to consider the single mode demand forecasting model based on the deep learning algorithm. For example, the research takes the generation and attraction of private cars and the impendence between traffic zones as the input of the model, and the travel demand of private cars as the output.

The research then establishes a multi-mode transportation model. Based on the original model, this research considers adding other transportation mode OD information to provide more features for the model to learn in order to improve the prediction accuracy. The prediction accuracy improvement which is compared to the original model can provide strong empirical evidence for correlation among multi-mode travels.

III. DATA DESCRIPTION AND PREPROCESSING

Our current research used data from Shenzhen on September 1, 2016, including IC card data, taxi GPS data and AVI data of Shenzhen. In this research, the study analyzed IC card

TABLE 1. Taxi Gps data format.

Field Name	Specification
YueXXX	license plate number
DATE	date of car trips
Time Stamp	magnetic field strength
Long	longitude of the vehicle position
Lat	latitude of the vehicle position
Event ID	ID of the initial event
STAT	record whether the vehicle have passengers
V	speed of the vehicle
DEG	direction of the vehicle

data totaled 3.71 million records (including only metro), taxi GPS data totaled 46.78 million records, and AVI data totaled more than 13 million records.

A. TAXI GPS DATA DESCRIPTION

Table 1 shows the format of raw GPS data obtained from vehicle terminal data collection. YueXXX is the unique identity code of a local license plate; DATE stands for date of car trips; Time Stamp is the exact time of the GPS record; Long stands for longitude of the vehicle position, which provides information on the vehicle; Lat stands for latitude of the vehicle position; Event ID is the ID of the initial event; STATE is used to record whether the vehicle have passengers; V stands for speed of the vehicle; DEG is the Direction of the vehicle.

These raw GPS data were pre-processed to improve system efficiency and performance of the query function. Preprocessing of data included data quality analysis, cleaning and conversion. For any GPS data, the study only chooses the valid data, including local license plate, date of car trips, time, longitude, latitude and whether to carry passengers.

In order to ensure the accuracy of data calculation and avoid the interference of wrong data or redundant data, it is necessary to clean the data. The trajectory data is very similar in form to the GPS data which shown in table 2. In this study, the approach of extracting GPS data information is divided into the following steps:

1) Vehicle information classification and extraction;
2) Vehicle trajectory can be simplified into OD information;
3) Exclude data from outside the research area;
4) Collate all taxi OD information.

The visualized results of the partial trajectories after processing are shown in figure.1.

As shown in figure. 2, the OD points of taxi trip were largely concentrated in LuoHu district, FuTian district and Nanshan district in the south of Shenzhen. Additionally, many taxi trips were concentrated in the southern region. In the subsequent division of traffic zones, this is a problem that needs to be paid attention to. Due to a small area in the southern region, many trips were concentrated. If the division of residential areas was not elaborate enough, it may lead to too many trips in the region.

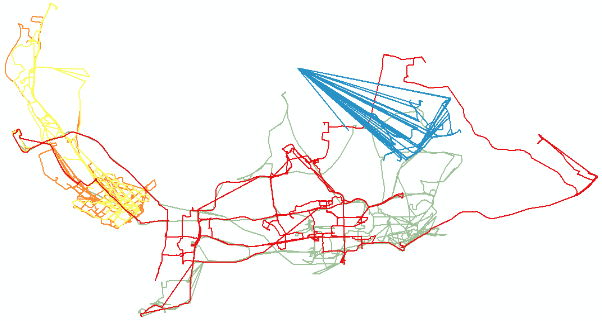

FIGURE 1. Visualized results of taxi trajectories.

TABLE 2. Simplified trajectory information.

License Plate	Time	Long	Lat	Change State
YUE BXXX	110028	114.027946	22.609818	0
YUE BXXX	111958	113.968597	22.557151	1
YUE BXXX	120917	113.972786	22.540567	0
...	...	...	...	...
YUE BXXX	132247	113.917053	22.509232	1
YUE BXXX	110028	114.027946	22.609818	0
YUE BXXX	231325	114.109734	22.610067	1

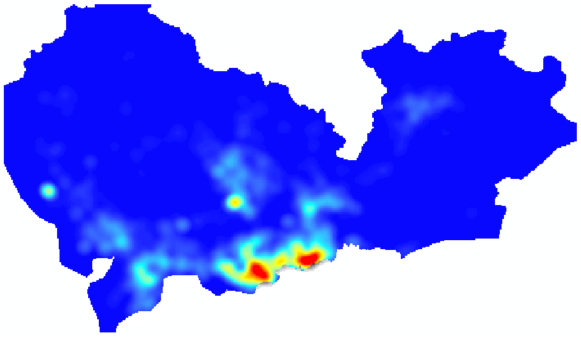

FIGURE 2. Nuclear density analysis results of taxi OD.

From the perspective of the trajectory line, the blue line shown in the figure.3 (a) was the trajectory line of medium and short distance trips, while the yellow and orange lines in the figure.3(b) were the trajectory line of long distance trips. It can be clearly seen that the taxi performs both short distance trips and long distance trips. However, in this research area, it mainly undertook medium and short distance trips. The

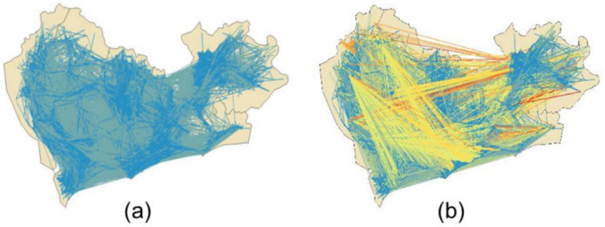

FIGURE 3. Trajectory line of taxi trips. (a) Short distance trips trajectory. (b) Long distance trips trajectory.

following research will further explore the characteristics of taxi trips.

B. IC CARD DATA DESCRIPTION

The IC card data in Shenzhen not only included the IC card data of metro trips, but also included the IC card data of bus trips. The research object of this paper did not include the transportation mode of bus. Therefore, the IC card data of bus were excluded from the IC data. The form of raw IC card is shown in table 3.

IC card data cleaning, the most important task, had these rules:

1) Reserving the following data fields for this research, including record encoding, card logic encoding, record time, event ID, station name, vehicle number and other data fields were removing when preprocessing.
2) Retaining data collected within the ShenZhen, our study area.
3) Keeping data record of ShenZhen metro from 7 o 'clock to 11 o 'clock.

C. PRIVATE CAR AVI DATA DESCRIPTION

The AVI data includes almost all types of motor vehicles in Shenzhen. Including special vehicles such as police cars and taxis. The purpose of this research is to extract private car trip records from the checkpoint data, explore the characteristics and rules of private car trip, and was compare with other transportation modes.

Different from the previous two kinds of data processing, this data processing approach is relatively complex. Because the traditional approach of extracting OD from vehicle identify data is to take the records of "origin" and "destination" of a vehicle as the origin and destination points respectively. However, this extraction approach would make it impossible to correctly identify multiple trips made by a vehicle.

As shown in figure.4, if the vehicle had passed a vehicle identify bayonet (I + 1), with a long stay, and then the vehicle passed the vehicle identify bayonet K, apparently interval time between the vehicle identify bayonets is far greater than the vehicle driving time interval t, according to the actual situation, the vehicle passed bayonet (I + 1) can be regarded as finished the trip at a time. However, the traditional extraction approach will include vehicle identify K in this trip, resulting in the previous completed trip cannot be identified.

TABLE 3. The form of Ic card data.

The Form of Data	Specification
70223XXX	record encoding
281197XXX	card logic encoding
300376	company code
31	transaction type
200	transaction amount
160	card balance
2016-09-01 02:12:23	record time
1	event ID
2016-09-01 02:12:22	unknown time 1
2016-09-01 02:12:24	unknown time2
Eastern Public Transport	company name
317	station name
CC565	vehicle number

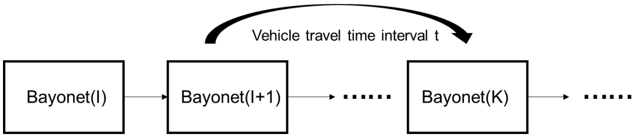

FIGURE 4. Vehicle trips record of bayonet.

In this situation, the research optimized the traditional extraction method and used the driving time between vehicle identify to determine whether a trip has been completed.

The method of extracting AVI data information was divided into the following steps:

1) Vehicle information was extracted by classification;
2) Vehicle trip time was used to distinguish trip destination:

The driving time was used to determine the destination of a trip; the research took the trip time of all vehicles between any two vehicle identify bayonets data record. Most of these records were the normal trip time of vehicles between the vehicle identify bayonets. However, many abnormal values were also included, which far exceed the normal trip time. These abnormal driving times were mainly caused by the long stay after completing a trip. In order to explore the distribution of trip time between vehicle identify bayonets, the research extracted all the driving time and made a box plot in figure.5.

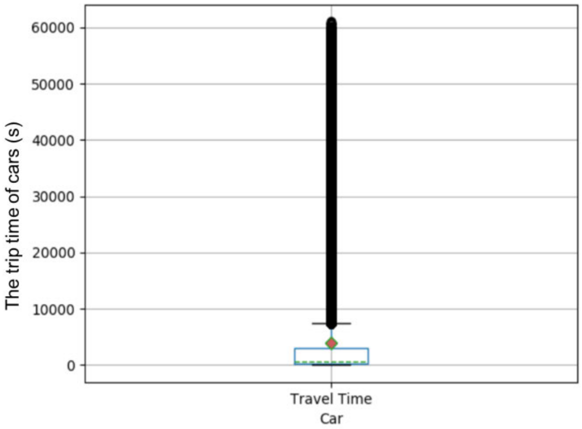

FIGURE 5. Car trip time box plot.

TABLE 4. Avi record of vehicle.

Serial Number	License Plate	Timestamp (second)	Time Interval	Detection Point ID
1	YUE BXXX	44343	0	2010075
2	YUE BXXX	44511	168	20606401
3	YUE BXXX	57069	12558	206A0422
4	YUE BXXX	59294	2225	39163
…	…	…	…	…
5	YUE BXXX	78622	6967	103A0895

In figure.5, the upper limit of the box diagram was 7440 seconds, and the trip time between vehicle identify bayonets greater than 7440 seconds is considered an outlier. Therefore, this research took 7440 seconds as the threshold to judge whether a trip is completed or not. If the time between two vehicle identify bayonets was greater than 7440 seconds, the vehicle was identified as having completed a trip at the upper vehicle identify bayonet. As shown in table 4, the driving time of the vehicle from the second vehicle identify bayonet to the third vehicle identify bayonet far exceeds 7440s, the vehicle was considered to have completed a trip from the bayonet I to bayonet II.

According to the comparison, 841228 trip records were extracted by the traditional extraction approach, while 1507470 trip records were extracted by the extraction approach based on the discrimination of vehicle driving time, which greatly improved the accuracy and completeness of data extraction compared with the traditional method.

IV. TRAVEL CHARARCTERISTICS ANALYSIS AND TRAFFIC ZONE DIVISION

A. TEMPRORAL AND SPATIAL CHARARCTERISTICS ANALYSIS OF MULTI-MODE TRANSPORTATION TRAVEL

In urban transportation system, different modes of transportation undertake different transportation functions due to their transportation characteristics. Moreover, residents in multi-mode traffic environment tend to make combined and diversified travels. In the extracted data, metro card and private car trip was far more than taxi trip. This showed that in the transportation structure of Shenzhen, taxi trips account for a relatively small proportion, while metro and private cars trip account for a relatively large proportion. Based on the trip records of the three transportation modes, this section made a comparative analysis of the travel rules and characteristics of the three transportation modes.

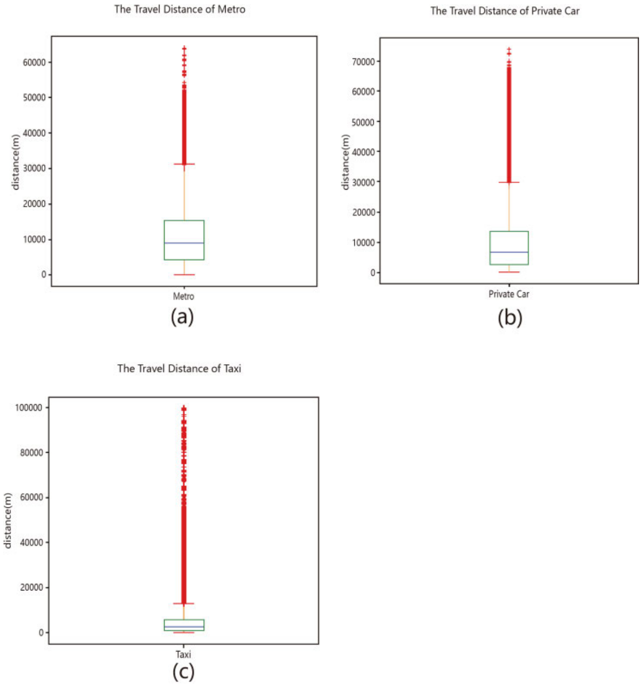

FIGURE 6. Box plot of different transportation modes. (a) Box plot of metro trips. (b) Box plot of private car. (c) Box plot of taxi.

The travel distance of transportation is often an important index to judge the transportation task undertaken by a mode of transportation. This section will use the results in Section 2 to analyze and compare travel distance distribution of various transportation modes.

Due to many outliers in the process of data processing, such as GPS data report point error or drift data interference, many vehicles trip data appeared thousands of kilometers abnormal trip. Considering the large amount of data, this research used the box plot for analysis and comparison. The research classifies the metro trip records, private car trip records and taxi trip records in Shenzhen. Since the geographic location information of the trip records was longitude and latitude coordinates, the research needed to convert the longitude and latitude to calculate the Euclidean distance from the original to the destination. Finally, it obtained the trip distance box plot of different transportation modes which shows in figure 6:

As shown in figure.6, the metro trip distance was generally less than 31 kilometers, with the upper and lower quartile being 15 kilometers and 4.5 kilometers respectively.

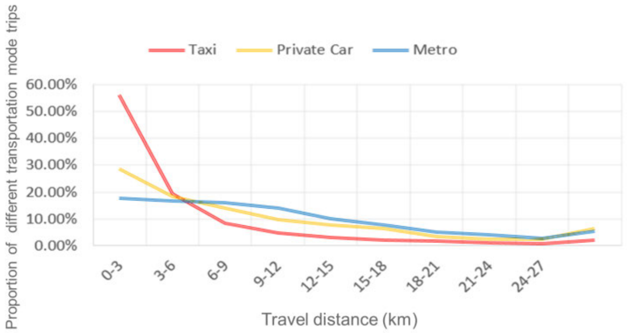

FIGURE 7. Multi-mode transportation trips distance proportion.

Considering that the distance in this research is not the actual trip distance and the distance of OD was the straight-line distance from the original and destination in the data, the actual trip distance values should be greater than this length. Generally, the short distance trips in big cities is about 4-6km. A small portion of the metro trips took short distance trip, while most of them take medium and long distance trips. Additionally, the trip distance of private cars was generally less than 30 kilometers and the upper and lower quarter points of the box plots were 13 kilometers and 2 kilometers respectively, indicating that the use of private cars in ShenZhen was relatively flexible. Private cars play an important role in short, medium and long distance trips. The difference between taxis and metros and private cars trips was the most obvious. According to the results of the box plots, most taxi trips were below 13.3 km, with the upper and lower quarter points being 6 km and 1 km. most of the taxi trips are short and medium distance.

However, the box plots showed only a general distribution of the data. If the research wanted to make a more detailed comparative analysis of different transportation patterns, it is difficult to rely on box plots alone. In this paper, the trend of the distance proportion in the total trip of different transportation modes was presented, as shown in the figure 7:

It can be clearly seen from the figure.7 that with the gradual increase of distance, the trip volume of the three transportation modes decreased relatively. However, the decline of taxi was much larger than that of private car and metro. This was consistent with the reality. Taxi trips were priced according to the trip distance. The larger the trip distance of taxi, the higher the price. On short and medium trip, taxis were favored by travelers because they are convenient and quick and relatively free to origins and destinations. The proportion of metro in short distance trip was relatively small. The proportion of metro in medium and long-distance trip was relatively large. However, private car trip distance was relatively stable and its trip volume was always in a moderate state with the change of distance.

In addition to distribution difference in trip distance, the distribution difference in time of the three transportation modes is also significant. As shown in the figure.8, the research made statistics on the number of trips in different traffic modes from 7 o 'clock to 23 o 'clock on September 1.

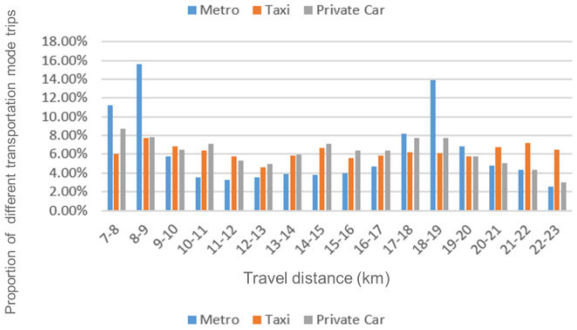

FIGURE 8. Multi-mode traffic trips distance proportion.

For the analysis of multi-mode spatio-temporal pattern, this research had the following important information, which would be used as the basis for the division of TAZ in the section 3.3 and traffic demand forecast:

1) From the perspective of space, the trip distance of taxi was shorter than that of the other two modes of transportation. It also focuses on short trips.

2) From the perspective of time, subway and private cars were affected by the trip time, while taxi trips were not affected by the time.

B. THE CLUSTERING METHOD OF TAZ DIVISION

The concept of TAZ is put forward from the field of traffic planning. Its purpose is combining the generation and attraction of traffic demand with the social and economic indicators of the region to show the spatial and temporal distribution of traffic demand flow between TAZ and or simulate the traffic state of the road network.

The division of TAZ is an important step in traffic planning. The traditional division of TAZ is mainly divided by administrative area or geographical restrictions and a series of division principles should be followed at the same time. According to the traditional division method, when division of traffic areas are intensive more, the accuracy is higher. However, the workload of traffic investigation, analysis and evaluation will also increase. It is difficult to balance the accuracy and workload. With the continuous development of transportation system technology, vehicle GPS, vehicle identify camera and other equipment continue to emerge, providing a more rapid and efficient data means for traffic analysis. It has become a hot issue in the field of transportation to study and predict the traffic behavior law by using the big data generated by traffic.

The research considered using GPS, IC card data and vehicle identify data, using k-means clustering method to divide Shenzhen traffic zones.

K-means algorithm is a clustering algorithm. The so-called clustering refers to the classification of relatively close data samples into the same cluster according to some similarity rules [25]. In general, the similarity between objects is measured by distance. The evaluation of clustering results is usually based on whether there is enough differentiation between classes in the clustering results and whether different objects are closely related under the same category.

K-means algorithm is very efficient for clustering big data and has good portability. However, the disadvantages are also obvious. Obviously, k-means algorithm needs to determine the size of k value which is the number of traffic zones. The results of selecting different k-value clustering are quite different. When the k-value is large, the zones will be too small, which will not only bring more workload to the follow-up data processing and analysis, but also lead to the fact that the travel volume of some zones is too small or appear many travel behaviors do not conform to the reality. A small value of k will lead to a lack of fine division of the area and too many travels volume in the area.

The k-means algorithm is mainly divided into the following steps:

1) Given the number of classification groups k;
2) Randomly selecting k initial clustering centers from data objects;
3) The distance of other objects to each center is calculated and classified into the class cluster with the smallest distance;
4) According to the clustering results, the center of each class is recalculated. The calculation process of the center is to calculate the average value of each dimension of all objects in this class;
5) All objects are reclassified according to the newly identified K clustering centers;
6) Repeating this process until the cluster center was essentially unchanged.

The selection of k value is directly related to the effect of clustering and has a great impact on the subsequent analysis. In this paper, elbow rule and contour coefficient were used to optimize the k value of clustering to improve the clustering effect. Elbow method was used to quickly and efficiently divide the appropriate quantity range of the traffic area. Furthermore, the number of the TAZ would be accurately and objectively evaluated by more in-depth and specific contour coefficient method.

Elbow method was calculated by taking the sum of the distortion degree of the category as the cost function. The sum of the distortion degree of the category is usually measured by the Sum of the Squared Error (SSE), which shown in the following:

$$SSE = \sum_{i=1}^{\mathrm{k}} \sum_{x \in c_i} dist(c_i - x)^2 \tag{4}$$

where k is number of clustering centers, is the ith cluster center and denotes Euclidean distance.

With the concept of error sum of squares, the researchers can find the kth center of mass mathematically. The objective function is set to be SSE. Moreover, in order to minimize it, the research take the derivative of it and set its derivative to zero:

$$\frac{\partial SSE}{\partial ck} = \frac{\partial \sum_{i=1}^{k} \sum_{x \in c_i} (c_i - x)^2}{\partial ck} = 0 \tag{5}$$

where k is number of clustering centers, is the ith cluster center.

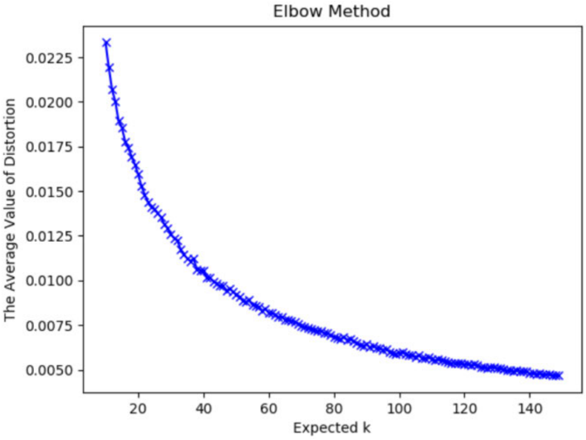

FIGURE 9. Result chart of elbow mothed.

The best center of mass to minimize SSE is the arithmetic mean of all objects in the cluster.

The elbow method also takes the error squared and SSE as the cost function. When the category increases gradually, SSE will gradually decrease. When k value of data with degree of differentiation reaches a certain critical point, the degree of distortion will be improved greatly, SSE significantly reduced and then gradually smooth. SSE relation between k is shown in figure.9, forming an ''elbow''. ''Elbows'' location near the critical point is the best k value in the range. In figure.9, when the number of zones was more than 100, the decrease of error SSE was very small and the whole image tends to be flat. The paper set the number of TAZ between 20 and 70.

As mentioned above, elbow method had its drawbacks. This approach took SSE as a cost function. However, when the k value of clustering increases to a certain degree or even extreme to a sample divided into a small area that SSE is 0 and the clustering effect is evaluated as the best, which is the extreme point of its evaluation. Silhouette coefficient approach better compensates for this. It considers the clustering cohesion and separation. Through the approach of Silhouette coefficient, the clustering results are evaluated by comparing the Silhouette coefficient. The value range of the silhouette coefficient is [−1,1]. When the value of the coefficient is closer to 1, it indicates that the object is more closely related to the homogeneous class cluster and has a higher discrimination with the surrounding class cluster, therefore, the evaluation effect is better. The calculation steps of Silhouette coefficient are as follows:

1) Calculating the average distance between object i and other samples of the class cluster $s_1, s_2 \ldots s_k$, a_i is called intra-cluster dissimilarity. When a_i is smaller, t

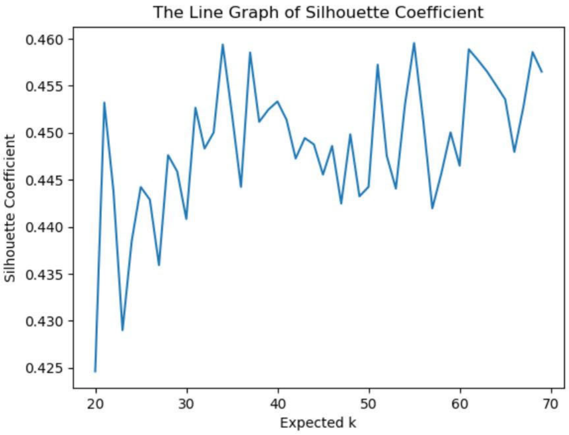

FIGURE 10. Silhouette coefficient graph.

the object is easier to be assigned to the class cluster. The mean of the intra-cluster dissimilarity of all objects in class cluster M is called the cluster dissimilarity of cluster M.

2) Calculating the average distance b_{ij} of all samples from object i to some other cluster N, which is called the dissimilarity between object i and N. Define the inter-cluster dissimilarity of object i: $b_i = \min\{b_{ij}, b_{i2} \ldots, b_{ik}\}$. The larger b_{ij} is, the less the object belongs to other class clusters.
3) According to the intra-cluster dissimilarity and inter-cluster dissimilarity of objects, the calculation formula of contour coefficient is defined as:

$$S(i) = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{6}$$

where a_i is intra-cluster dissimilarity, b_i is inter-cluster dissimilarity of object i.

According to the Section 3.2.3 mentioned, the range of the number of zones had been locked between 20 and 70 cells. Additionally, the Python language was used to classify the number of different cells to calculate their Silhouette coefficient values, as shown in figure.10:

As shown in figure.10, when the clustering value was around 21, 34, 51 and 63, the Silhouette coefficient was the largest and the cohesion and separation of classification performance can achieve a better effect. Considering Shenzhen city on September 1 on the same day the travel volume of each transport modes, if number of TAZ was too little, it would cause too much proportion in travel and further affect the subsequent model forecasting. Therefore, the research would select 34 traffic zones based on silhouette coefficient.

C. THE CLUSTERING METHOD OF TAZ DIVISION

When citizens travel by different transportation modes, their travels will have different characteristics in time and space. Therefore, according to their characteristics, GPS data was selected to divide the TAZ. The main reasons are following:

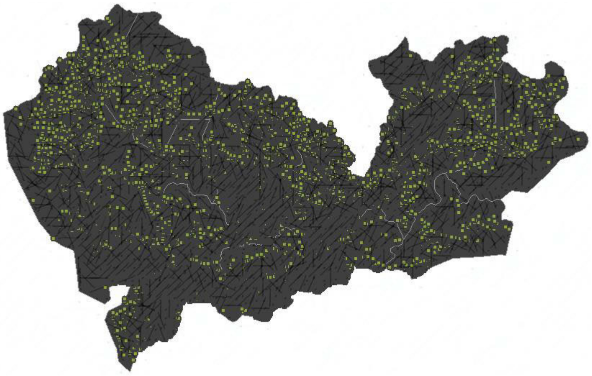

FIGURE 11. Distribution of taxi trip OD points.

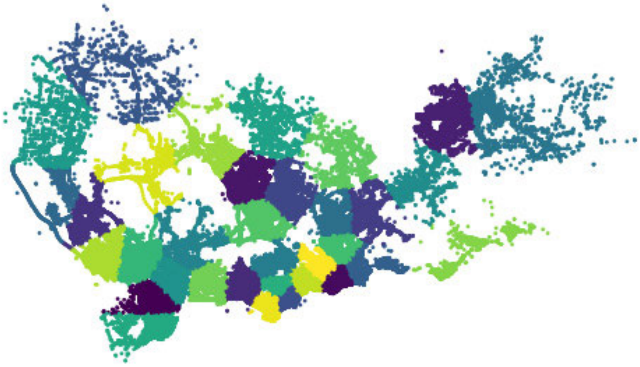

FIGURE 12. Results of k-means clustering.

1) When dividing the TAZ, researchers pay more attention to the trip generation of the TAZ. However, the trip generation of TAZ at different times is different. Therefore, if the travel information provided by the data is greatly affected by time, it will affect the TAZ division when researchers use data-driven method to divide. According to analysis in section 3.1. Taxi trip is quite different from the other two modes trip, which is reflected in stability time distribution.
2) The metro has fixed routes, and the nodes of routes are mainly composed of 137 metro stations. Although private cars trip freely, they record the trajectory by vehicle identify points. Therefore, their data coordinates are also fixed at more than 2000 vehicle identify points. By comparison, the OD of the taxi distribution from a certain extent reflects the urban spatial distribution of travel demand which is shown in figure.11. The trajectory of taxi is more random than other two modes, Therefore, it is more reasonable to use GPS data to divide traffic zone.

According to the elbow method and contour coefficient method used above, the number of traffic zones, i.e., the clustering k value, is set at 34. Cluster all taxi trip starting and ending points, and the clustering results are as figure.12:

By comparing division results which is shown in figure.12 with trip heat map which shown in figure.12. The heat map was simple according to the trip density analysis.

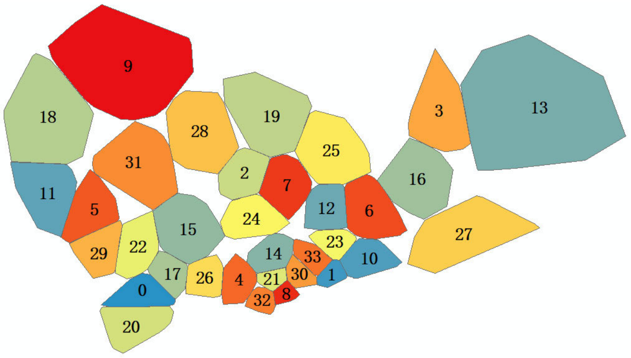

FIGURE 13. Result of boundary division.

TABLE 5. Distribution matrix of metro trips from 7:00-8:00.

D / O	0	1	2	…
1	654	764	88	…
2	232	867	102	…
3	387	141	528	…
…	…	…	…	…

It was found in heat map that the OD distribution of area was intensive in the southern area, the system divided it into the populated TAZ, smaller TAZ. In the northern area OD distribution was relatively sparse, the area divided in heat map is larger. The clustering results can fully reduce the large difference in OD distribution between the TAZ caused by the over-concentration and over-dispersion of trip density. At the same time, this division method also avoided the problem of excessive trips in the southern region. It was found that clustering results have a good effect, which can objectively consider the distribution differences of trip between different TAZ and solve the problem of large OD matrix value difference and large number of outliers in the forecasting model.

There are many methods for dividing zones. This paper proposed a boundary division method based on convex hull [26]. The convex hull refers to a set of points on a given plane, and the smallest convex polygon surrounding the set of points is obtained, which is the convex hull. In this paper, the class clusters formed by the divided samples were regarded as a set of points $s_1, s_2, s_3 \ldots s_k$, where k is the cluster number of the TAZ. The result is shown in figure.13.

D. OD IDENTIFICATION

The paper defined the TAZ as zones 0-33. The travel demand per hour is studied and the time of the travel demand generation (based on the departure time) was taken as the attribute to divide the time period. Considering the limitation of the metro opening time, it wass divided into 16-hour segments from 7:00 to 23:00 of the day of September 1, and OD was extracted for each hour segment. The OD distribution matrix of metro was obtained as table 5. Additionally, the OD distribution matrix of private cars and taxis were also similar to this.

TABLE 6. Model parameters.

Parameter	Value
Learning rate	0.0001
Batch size	256
Dropout Probability	0.5

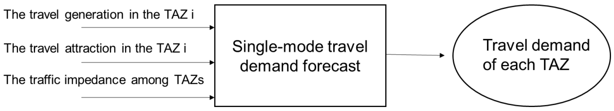
The travel generation in the TAZ i
The travel attraction in the TAZ i
The traffic impedance among TAZs
Single-mode travel demand forecast
Travel demand of each TAZ

FIGURE 14. The structure of the single mode forecasting model.

V. TRAFFIC DEMAND FORECASTING MODEL

A. EVALUATION METHODOLOGY AND PARAMETERS

To testify the existence of correlation among multi-mode travels and travel demand. Comparative experiment was also conducted. Multi-mode traffic demand forecasting model was established. The python libraries and google cloud were used to build our models. The research uses the NVIDIA Tesla P100, CUDA8.0 and cuDNN 5.1.

The experimental environments parameter settings are shown in table 6.

B. INTRODUCTION TO MULTI-MODE TRAFFIC DEMAND FORECASTING MODEL

In the traditional traffic zone demand forecasting methods mentioned in section II. Based on the divided traffic zones, the social and economic data of each traffic zone were collected. Moreover, the travel demand of each traffic zone was predicted by combining the traffic impedance of each traffic zone with the trip generation and attraction of each traffic zone. In this paper, the single-mode demand forecasting model was based on deep learning algorithm [30], [31], and the forecasting structure of traditional traffic zone demand forecasting was used for reference, taking private car forecasting as an example. The model was established by taking the private cars generation and attraction of TAZ, the distance between TAZ as the input of the model, and the trip demand of private cars as the output. The structure of the model is shown in figure14:

Since the data in this study was not large and the complexity of the model was moderate, two hidden layers were selected. For the selection of the number of neurons in the two hidden layers, if the number of neurons in the hidden layer is too small, the number of samples will not be satisfied; if the number of neurons in the hidden layer is too large, the generalization ability of the model will be poor. Therefore, the study used empirical formula to determine the approximate selection number of neurons in the hidden layer, and then select the neural network structure with the minimum error by selecting part of the data set for testing.

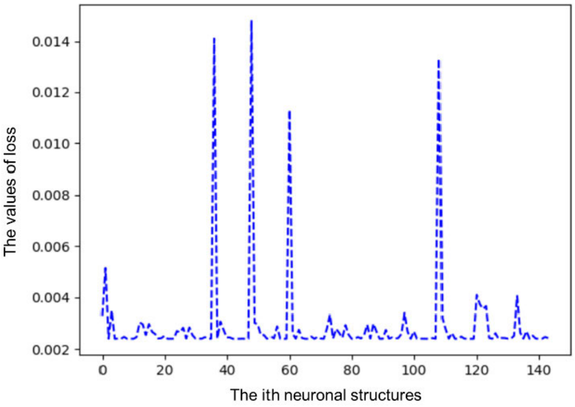

FIGURE 15. Test results of neurons with different structures.

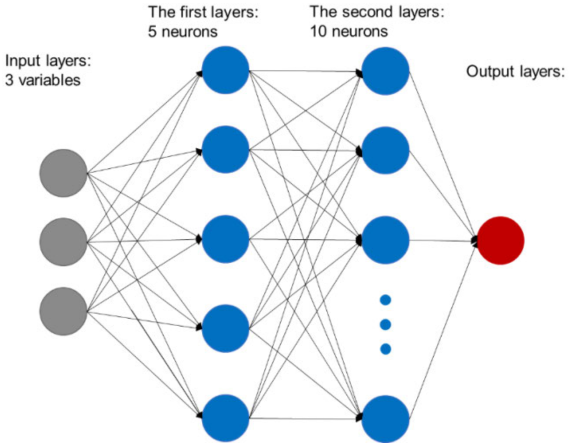

FIGURE 16. Single mode transportation neural network structure.

The empirical formula is:

$$t = \sqrt{m+n} + a \tag{7}$$

Where t is the number of hidden nodes s; m is the number of nodes in the input layer; n is the number of output layer nodes; a is a constant, usually 1 minus 10.

According to the empirical formula, the number of neurons in the hidden layer was set as 1 to 12, and the number of neurons in the two hidden layers was taken as the object to obtain the forecasting errors of all the different structure models through traversal, which was 144 times in total. Some samples were taken as test data and compared. According to the result, the research will choose the most appreciate neural network structure. The result is shown in figure.15 as follows:

As shown in the figure.15, After calculation, the 70th similar neural network model structure could achieve the minimum loss. Therefore, the 70th neural network structure with 5 neurons in the first hidden layer and 10 neurons in the second hidden layer was finally selected which is shown in figure.16. This single mode traffic demand forecasting model would be used to compare with the multi-mode traffic demand model in the next section.

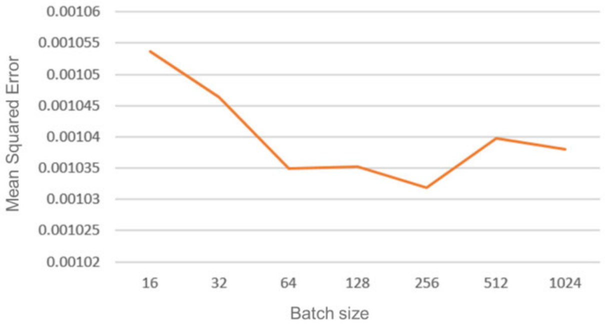

FIGURE 17. The result of batch size test.

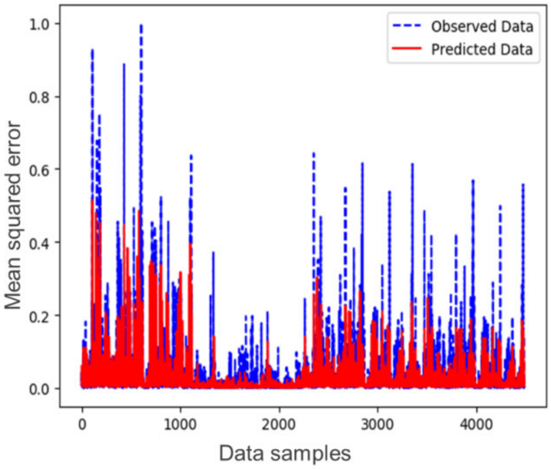

FIGURE 18. The result of traffic demand test.

The research selected part of the sample data to select different batch size through comparing their Mean Squared Error (MSE). The result of batch size test is shown in figure.17.

After comparison, the study finally chose 256 as the final batch size. In addition, the research also used the dropout method. Finally, the trained model was applied to the test set, and the result is shown in figure.18.

The calculated result of MSE (Mean Squared Error) was 0.002017, with a small error value, and good forecasting results have been achieved.

C. RESEARCH ON MULTI-MODE TRAFFIC DEMAND FORECASTING MODEL

In this research, the single-mode demand forecasting model has achieved good results. On this basis, this study further upgraded the model. There were some limitations to predict a travel pattern information based on its own information. The travel patterns of road network were complex and diverse. Furthermore, the single model ignored the correlation between travel patterns. In order to make it learn and improve the forecasting accuracy, this research considered adding information on other traffic patterns related to the

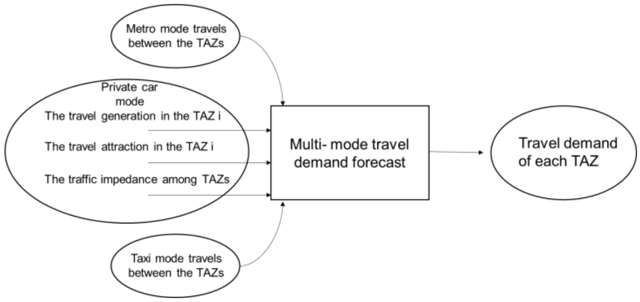
Metro mode travels between the TAZs
Private car mode
The travel generation in the TAZ i
The travel attraction in the TAZ i
The traffic impedance among TAZs
Multi- mode travel demand forecast
Travel demand of each TAZ
Taxi mode travels between the TAZs

FIGURE 19. The structure of the multi-mode forecasting model.

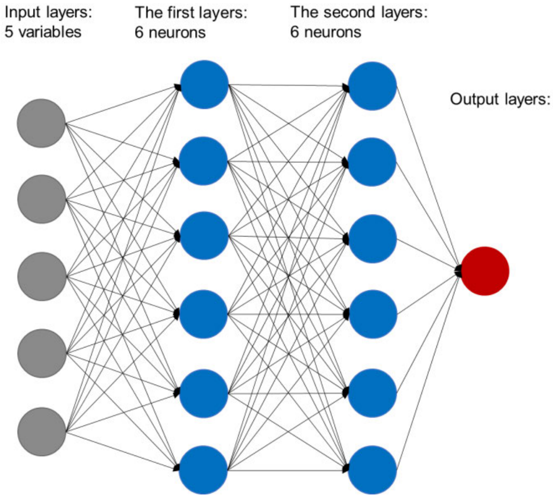
Input layers: 5 variables
The first layers: 6 neurons
The second layers: 6 neurons
Output layers:

FIGURE 20. Multi-mode transportation neural network structure.

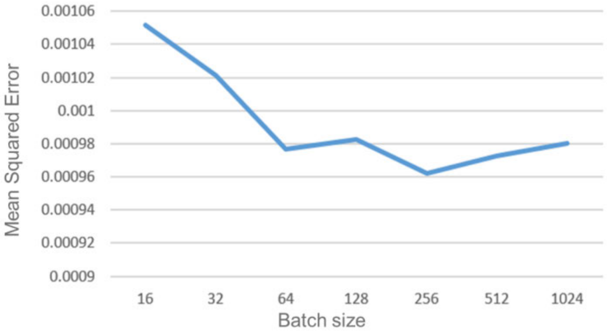

FIGURE 21. The result of batch size test.

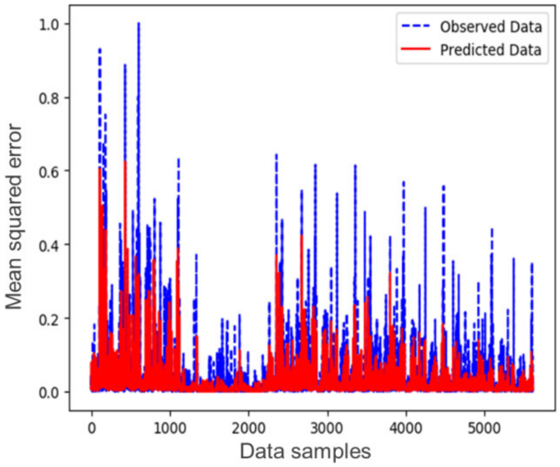

FIGURE 22. The result of multi-mode traffic demand test.

predicted object based on the original model to provide more features for the model [28].

In this section, the dimensions of input data were increased compared with single-mode traffic demand forecasting. The research took private cars as the forecasting object. Moreover, the travel demand data of taxi and metro trips were added to the input nodes of the network as input values. Additionally, the correlation between various traffic modes was added to the model. The structure of the multi-mode forecasting model is shown as figure.19:

Neuron structure also adds two dimensions to the single-pattern forecasting model. Similarly, the empirical method and the traversing method were used to obtain the optimal neural network structure model. The first hidden layer was determined with 5 neurons, and the second hidden layer was determined with 6 neurons, as shown in figure.20:

Based on the determination of neuron structure, the research selected the batch size for model which was shown in figure.21:

The Loss function still chose MSE, which iterates faster. Moreover, batch size is selected to 256, which is a reasonable batch size. Finally, the model was trained, and the trained model was tested with a test set. The test results are shown in figure.22:

The calculated result of MSE is 0.00176, and the value was smaller and more accurate than the single-mode transportation forecasting. Certainly, it is difficult to accurately compare the performance of the two models based on this index alone, and more evaluation indicators would be introduced in the following sections for comprehensive evaluation and comparison. It had certain limitations to predict a traffic pattern only by the information of the transportation mode itself. The traffic patterns of road network were complex and diverse, and the single forecasting model ignored the correlation between them. Based on the original model, this research considered adding the information of other transportation modes related to the predicted object to provide more features for the model, which also proved indirectly that the fusion of multi-mode traffic data was beneficial to improve the accuracy of traffic demand forecasting.

D. MODEL EVALUTION AND COMPARISON

The single mode traffic demand forecasting model only considers the generation and attraction of the traffic zones and the traffic impedance as the influencing factors of the travel demand, while the multi-mode traffic demand forecasting model comprehensively considers the influence of the other

TABLE 7. **Model results comparison table.**

Evaluation Indexes	Single Mode Traffic Demand Forecasting Model	Multi-Mode Traffic Demand Forecasting Model
MSE	0.002017	0.00176
MAPE	14.83086	12.67376
MAE	0.01709	0.01661
MSLE	0.00191	0.00138

two modes of traffic demand. This section will compare the two prediction models.

In the process of model fitting, the loss function was MSE. Loss is a kind of error function composed of actual value and forecasting. In addition to using loss function to fit the model, the function could also be used to evaluate the model. Commonly used evaluation indexes include: MSE (mean squared error), MAE (mean absolute error), MSLE (mean squared logarithmic error), and MAPE (mean absolute percentage error).

Different evaluation indexes have their advantages and obvious defects. For example, MSE index will expand the impact of larger values, making the samples with larger values play a more important role in model evaluation. MSLE is an improvement based on MSE, which weakens the influence of the maximum error band by logarithm function.

Different evaluation indexes have their own advantages and disadvantages. This research combined different evaluation indexes to evaluate the models more comprehensively. The evaluation results are in VIII:

It can be seen that the accuracy of multi-modal traffic demand forecasting model hads been improved in all aspects, and the results of different indicators showed the advantages of multimodal traffic demand forecasting model in demand forecasting.

This research showed that the correlation between different traffic modes was helpful to optimize the demand forecasting model. The original single mode traffic forecasting model could use more data features for learning, which greatly improves the prediction accuracy because it increased other related traffic modes for travel demand prediction. Furthermore, it provided strong empirical evidence for the existence of correlation among multi-mode travels and travel demand.

VI. DISCUSSION

Traffic zone division and travel demand forecasting play important roles in the accurate calibration of travel demand forecasting models. Traffic zones are an important component of traffic surveys, growth forecasting, trip generation and trip distribution. Several common rules in the conventional traffic zone division approach [29]. These include: 1) taking rivers and mountain as the natural and administrative boundaries; 2) maintaining the consistent characteristics of social, land-use and economic activity in one traffic zone; and 3) selecting appropriate traffic zone sizes. In general, zones are larger in the suburbs where traffic volume is low and smaller in urban/ Central Business District areas where traffic volume is high. However, the conventional method of traffic zone division is not a theory-driven approach that cannot reflect the latest or even real-time traffic patterns and consistent characteristics within a traffic zone. Our data-driven approach had shown that that GPS data, IC card data and AVI data are emerging and promising data sources for traffic zone division. The proposed method offered improved flexibility in selecting the desired number of zones and reasonable division accuracy.

Moreover, our research also proposed a deep learning method to predict the traffic demand in an aggregation way and compare with single sparse data, which provided a new perspective for traffic demand forecasting and provide strong empirical evidence for the inexistence of correlation among multi-mode travels.

Our future research will address the following limitations and challenges:

1) Our current approach did not take into consideration geographical boundaries (rivers, railways) and administrative boundaries. In the future, we will develop a refined method of traffic zone division which will combine geographical boundaries, administrative boundaries and geographic information systems (GIS) through a post-processing.
2) Multiple transportation modes used by a same traveler affects feature data extracted from the multiple data.
3) The forecasting model lacked a certain interpretability, and it could not give a clear explanation of what kind of correlation exists between various data.
4) Due to the limited available data, the research combined the spatial and temporal characteristics of various data to select the data of taxis for dividing the TAZs. In the future, the research will collect more useful and relevant information, such as citizen acceptance of different transportation modes, relevant traffic policies and commuters travel taste heterogeneity, to divide the community, which will be more reasonable [32]–[35].

VII. CONCLUSION AND RECOMMENDATION

This paper proposed a data-driven to extract travels' OD information from multiple data, and to use the extracted data for analyzing the characteristics of multiple transportation modes. Moreover, the research proposed a k-means clustering algorithm to classify TAZ. It based on characteristics of multiple transportation modes to divide the ShenZhen into a total of 34 traffic zones using another k-means clustering algorithm. Furthermore, it proposed a machine learning method to predict the traffic demand in an aggregation way and compare with single sparse data, which provided a new perspective for traffic demand forecasting and provide strong empirical evidence for the inexistence of correlation among multi-mode transportations. At present, there are a few researches of multi-mode traffic demand forecasting. This research put forward some ideas on this research field. However, there are still many deficiencies. The following are some aspects worth further expanding and deepening in this research:

1) In this research, AVI data, taxi GPS data and IC card data were selected as data sources for multi-mode traffic demand forecasting, and metro, taxi and private car were selected as three transportation modes. Future researches can be further deepened in the selection of transportation modes, such as establishing a more comprehensive multi-mode traffic demand forecasting model with bus and other transportation modes and exploring the complex non-linear spatial and temporal relations among regions and multiple transportation modes.
2) In terms of temporal dimension, data of September 1 in Shenzhen city were selected in this research. Traffic data analysis work is very complicated due to the large amount of data, which requires patient data processing. In the future, if there are sufficient data sources, we can consider processing the data of a week or even a month. More data can make the model better and the forecasting accuracy will be greatly improved.
3) As for the division of TAZ, k-means algorithm was adopted in this paper, which has advantages and disadvantages. The advantage lies in its efficient division, while the disadvantage lies in many problems such as selection of initial center of mass and k value. The improvement methods adopted in this paper cannot completely solve these problems, therefore, other classification algorithms can be considered for later research.
4) For the forecasting method, the multi-layer perceptron model with multiple hidden layers is adopted in this paper, and better forecasting models such as Convolutional Neural Networks can be considered in the future. In addition, the form of input and output of the model can be changed. For example, the distribution of OD matrix can be considered as an image, and the forecasting of OD matrix as an image recognition process, which can be further studied in the future.

REFERENCES

[1] Y. Ji, Y. Fan, A. Ermagun, X. Cao, W. Wang, and K. Das, "Public bicycle as a feeder mode to rail transit in China: The role of gender, age, income, trip purpose, and bicycle theft experience," *Int. J. Sustain. Transp.*, vol. 11, no. 4, pp. 308–317, Apr. 2017.

[2] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2933–2943, Aug. 2019.

[3] Z. Yu, C. Licia, W. Ouri, and Y. Hai, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[4] Y. Guo, Z. Li, Y. Wu, and C. Xu, "Evaluating factors affecting electric bike users' registration of license plate in China using Bayesian approach," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 59, pp. 212–221, Nov. 2018.

[5] Y. Pan, S. Chen, F. Qiao, S. V. Ukkusuri, and K. Tang, "Estimation of real-driving emissions for buses fueled with liquefied natural gas based on gradient boosted regression trees," *Sci. Total Environ.*, vol. 660, pp. 741–750, Apr. 2019.

[6] H. Dong, M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, and X. Zhou, "Traffic zone division based on big data from mobile phone base stations," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 278–291, Sep. 2015.

[7] C. Xu, Y. Wang, P. Liu, W. Wang, and J. Bao, "Quantitative risk assessment of freeway crash casualty using high-resolution traffic data," *Rel. Eng. Syst. Saf.*, vol. 169, pp. 299–311, Jan. 2018.

[8] Q. Cheng, Z. Liu, and W. Y. Szeto, "A cell-based dynamic congestion pricing scheme considering travel distance and time delay," *Transportmetrica B, Transp. Dyn.*, vol. 7, no. 1, pp. 1286–1304, Dec. 2019.

[9] P. Liu, J. Wu, H. Zhou, J. Bao, and Z. Yang, "Estimating queue length for contraflow left-turn lane design at signalized intersections," *J. Transp. Eng. A, Syst.*, vol. 145, no. 6, Jun. 2019, Art. no. 04019020, doi: 10.1061/JTEPBS.0000240.

[10] C. Wang, Z. Ye, E. Chen, M. Xu, and W. Wang, "Diffusion approximation for exploring the correlation between failure rate and bus-stop operation," *Transportmetrica A, Transp. Sci.*, vol. 15, no. 2, pp. 1306–1320, Nov. 2019.

[11] Y. Yuan, M. Yang, J. Wu, S. Rasouli, and D. Lei, "Assessing bus transit service from the perspective of elderly passengers in Harbin, China," *Int. J. Sustain. Transp.*, vol. 13, no. 10, pp. 761–776, Nov. 2019, doi: 10.1080/15568318.2018.1512691.

[12] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," *Image Vis. Comput.*, vol. 21, no. 4, pp. 359–381, Apr. 2003.

[13] C. Wang, C. Xu, J. Xia, Z. Qian, and L. Lu, "A combined use of microscopic traffic simulation and extreme value methods for traffic safety evaluation," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 281–291, May 2018.

[14] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 568–583, Aug. 2010.

[15] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Detecting pattern changes in individual travel behavior: A Bayesian approach," *Transp. Res. B, Methodol.*, vol. 112, pp. 73–88, Jun. 2018.

[16] D. Li, T. Miwa, and T. Morikawa, "Considering en-route choices in utility-based route choice modelling," *Netw. Spatial Econ.*, vol. 14, nos. 3–4, pp. 581–604, Dec. 2014.

[17] D. Li, T. Miwa, and T. Morikawa, "Modeling time-of-day car use behavior: A Bayesian network approach," *Transp. Res. D, Transp. Environ.*, vol. 47, pp. 54–66, Aug. 2016.

[18] D. Li, T. Miwa, C. Xu, and Z. Li, "Non-linear fixed and multi-level random effects of origin–destination specific attributes on route choice behaviour," *IET Intell. Transp. Syst.*, vol. 13, no. 4, pp. 654–660, Apr. 2019, doi: 10.1049/iet-its.2018.5251.

[19] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 301–313, Jan. 2013.

[20] J. A. Alvarez-Garcia, J. A. Ortega, L. Gonzalez-Abril, and F. Velasco, "Trip destination prediction based on past GPS log using a hidden Markov model," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8166–8171, Dec. 2010.

[21] D. Papinski, D. M. Scott, and S. T. Doherty, "Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 12, no. 4, pp. 347–358, Jul. 2009.

[22] E. Castillo, J. M. Menéndez, and P. Jiménez, "Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations," *Transp. Res. B, Methodol.*, vol. 42, no. 5, pp. 455–481, Jun. 2008.

[23] R. M. Assunção, M. C. Neves, G. Câmara, and C. Da Costa Freitas, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees," *Int. J. Geograph. Inf. Sci.*, vol. 20, no. 7, pp. 797–811, Aug. 2006.

[24] T. Shirabe, "A model of contiguity for spatial unit allocation," *Geograph. Anal.*, vol. 37, no. 1, pp. 2–16, Jan. 2005.

[25] M. Fort and J. A. Sellarès, "Solving the k-influence region problem with the GPU," *Inf. Sci.*, vol. 269, pp. 255–269, Jun. 2014.

[26] D. Guo and H. Wang, "Automatic region building for spatial analysis," *Trans. GIS*, vol. 15, pp. 29–45, Jul. 2011.

[27] S. Seyedabrishami and Y. Shafahi, "Expert knowledge-guided travel demand estimation: Neuro-fuzzy approach," *J. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 13–27, Feb. 2011.

[28] J. P. van der Vliet, A. J. Pel, and J. W. C. van Lint, "Enriched travel demand estimation by including zonal and traveler characteristics and their relationships," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2017, pp. 351–355.

[29] Q. Zhou, H.-P. Lu, and W. Xu, "New travel demand models with back-propagation network," in *Proc. 3rd Int. Conf. Natural Comput. (ICNC)*, 2007, pp. 311–317.

[30] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[31] Y. Liu, Z. Liu, and R. Jia, "DeepPF: A deep learning based architecture for metro passenger flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 101, pp. 18–34, Apr. 2019, doi: 10.1016/j.trc.2019.01.027.

[32] D. Li, T. Miwa, T. Morikawa, and P. Liu, "Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets," *Transp. Res. C, Emerg. Technol.*, vol. 67, pp. 31–46, Jun. 2016.

[33] L. Cheng, X. Chen, S. Yang, Z. Cao, J. De Vos, and F. Witlox, "Active travel for active ageing in China: The role of built environment," *J. Transp. Geography*, vol. 76, pp. 142–152, Apr. 2019, doi: 10.1016/j.jtrangeo.2019.03.010.

[34] Y. Zhang, Y. Jiang, W. Rui, and R. G. Thompson, "Analyzing truck fleets' acceptance of alternative fuel freight vehicles in China," *Renew. Energy*, vol. 134, pp. 1148–1155, Apr. 2019.

[35] D. Li, C.-J. Jin, M. Yang, and A. Chen, "Incorporating multi-level taste heterogeneity in route choice modeling: From disaggregated behavior analysis to aggregated network loading," *Travel Behaviour Soc.*, vol. 19, pp. 36–44, Apr. 2020, doi: 10.1016/j.tbs.2019.11.002.

DAWEI LI was born in 1987. He received the B.S. and M.S. degrees from Southeast University and the Ph.D. degree from National Nagoya University.

During his Ph.D. degree, he communicated with Prof. M. Ben-Akiva of Singapore MIT Joint Research Center and has participated in the development of path selection module of large simulation platform SimMobility. He was selected into the Xiangjiang scholars program (in cooperation with Prof. A. Chen of the Hong Kong University of Technology, the research topic is network modeling based on advanced path selection model). He was a special Associate Professor with the Institute of Materials and Systems for Sustainability (Imass, including Nobel laureate Hiroshi Amano), Nagoya University. He has published more than 20 articles in authoritative SCI journals in transportation research series, networks and spatial economics, IET ITS, transportation research record, and other transportation fields. He has presided over and participated in nearly 20 national, provincial, and ministerial projects.

YUXIANG TANG was born in Yancheng, Jiangsu, China, in 1996. He is currently pursuing the M.S. degree with the School of Transportation, Southeast University.

QIONG CHEN received the B.S. degree from Xi'an Jiaotong University and the M.S. degree from Southeast University, where she is currently pursuing the Ph.D. degree with the School of Architecture.