**ABSTRACT:**

Travel order demand prediction is of great importance for continuous upgrading of intelligent transportation system to city-scale services and personalized services. An accurate short-term travel demand prediction model in both spatial and temporal relations can help the city pre-allocate resources and facilitate city-scale travel operation management in megacity. To address problems like this, in this study, we propose a multi-zone order demand prediction model to predict short-term order demand in different zones at city-scale. Two-step methodology was developed including order zone division and multi-zone order prediction. For the zone division step, the K-Means++ spatial clustering algorithm was used and the parameter k of K-Means++ is estimated by Between-Within Proportion index. For the prediction step, six methods, LWLR, GA-BP Neural Network, SVR, average fusion-based method, weighted fusion-based method, kNN fusion based method, are used for comparison. To demonstrate the performance, three multi-zone weighted accuracy indictors were proposed to evaluate the order prediction ability at city-scale. These models were performed and validated on a real-world taxi order demand data from 3-month consecutive collection in Shenzhen, China. Experiment on city-scale taxi demand data demonstrates the superior performance of multi-zone order demand prediction model with kNN fusion based method used for prediction by proposed accuracy indicator.

**Key word:** zone division, multi-zone prediction, travel order demand, city-scale

1. **Introduction and related work:**

Traffic has become an important factor which impacts the city management and operation and daily life of millions of dwellers. One of the most fundamental problems for smart city is how to build an efficient transportation system. Nowadays, with the continuous upgrading of intelligent transportation system to city-scale services and personalized services, travel demand data concerned and needed to be accessed are rapidly increasing and directly processing city-scale data could put pressure on data-processing system, as data may be so complex as to pose great challenges for real-time calculation of city-scale system operation management. To address this problem, a critical component is an accurate short-term multi-zone travel demand prediction which could divide the order zone to reduce processing time and improve prediction accuracy simultaneously. The better we can predict travel order demand, the better we can help the city pre-allocate resources and facilitate city-scale travel operation management in megacity (Vlahogianni, Karlaftis et al. 2014). This kind of model can benefit many city-scale operational management scenarios. For example, for sharing service, it can help facilitate the schedule of sharing vehicle fleet in advanced to reduce the costly cruise expense; for taxi operation management, it can reduce the imbalance between taxi supply and demand in some areas.

In literature, there has been many studies in prediction for traffic scenarios, including traffic volume, taxi demand, traffic flow volume and these studies propose some prediction methods. To predict traffic, time series analysis methods are the most popular models. Representatively, autoregressive integrated moving average (ARIMA) are well-known time series forecasting models by its short-term prediction performance (Li, Pan et al. 2012; Moreira-Matias, Gama et al. 2013; Davis, Raina et al. 2018). More recently, machine learning methods have been frequently used to predict future traffic data, which attempt to identify historical data that are similar to the prediction instant, including neural networks (NN), support vector regression (SVR), random forest (RF), k-Nearest Neighbours (kNN) and so on. Nikravesh et al. (2016) compared some machine learning methods in terms of predicting traffic data, and the result showed that SVM performed better in predicting the multidimensionality of network traffic data. Habtemichael and Cetin (2016) proposed an enhanced k-Nearest Neighbours algorithm for short-term traffic forecasting and it indicated that the proposed method can provide promising results.

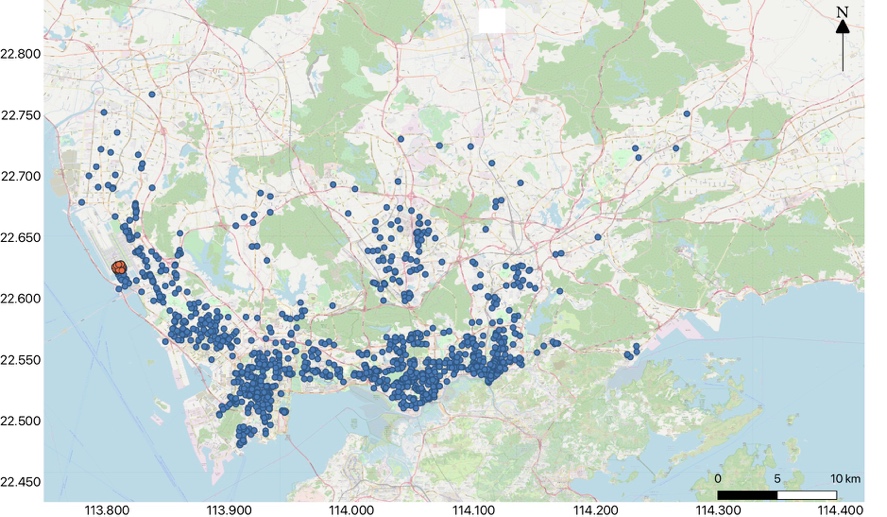
Although individual prediction methods have good predictive performance, in order to further reduce prediction error, in recent years, some researchers research methods to combine prediction models to improve prediction accuracy (Guo, Polak et al. 2010;Guo, Krishnan et al. 2017). Qiu et al. (2016) proposed an integrated precipitation-correction model to use fusion method with four prediction models to predict freeway traffic flow. Vlahogianni (2015) combined three different prediction models to propose a surrogate model for freeway traffic speed prediction. And these studies verified that the fusion based prediction model could improve the prediction accuracy.

According to the literature, it can be found that many researches on travel order demand prediction methods concerned more on temporal change,. In order to have better prediction performance and be adapted to city-scale development, some researches consider spatial effect on predictive performance. For example, some researches do research on travel order demand prediction for hot spot analysis, or grids. Li et al. (2012) analysed the spatial-temporal variation of passengers in a hot spot. Ke et al. (2017) used a novel deep learning approach called the fusion convolutional long short-term memory network to predict passenger demand under on-demand ride services for 77 grids in Hangzhou, China by analyzing spatial-temporal characteristics. However, currently, most researches focus on hot spot or evenly divided grid zones, there is no study about multi-zone prediction analysis considering order spatial distribution and overall performance for zone prediction as well.

1. **Dataset and Features**

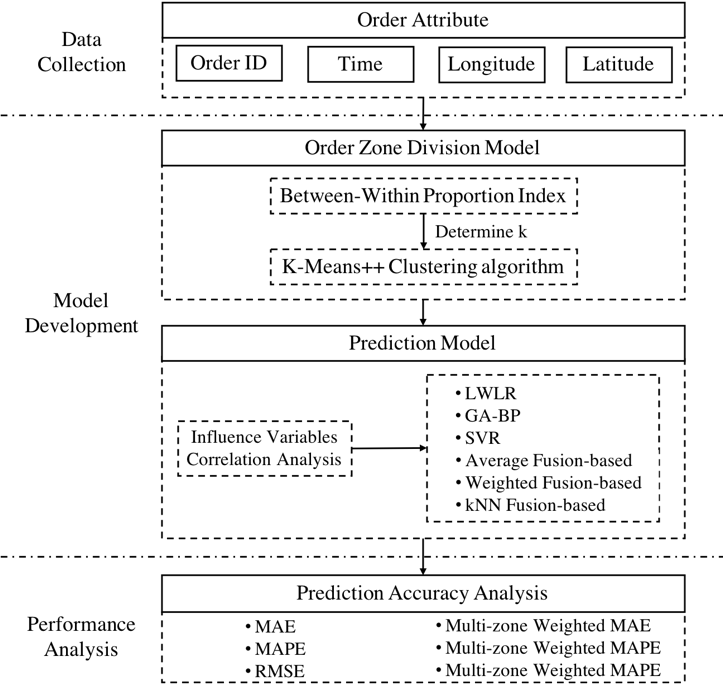
The data in this study come from the taxi order data consecutively collected from August 10 to October 23, 2015 in Shenzhen, China. The order data include information such as order ID, order time, order longitude and latitude.

The content of this study is to take order prediction in the urban area to Shenzhen Airport for example, so it does not consider order demand outside the city. The latitude and coordinate range of Shenzhen is 22˚45′~22˚82′ North latitude, and 113˚71′~114˚37′ East longitude which is the area of the study zone. Then the data in the above range is selected. Furthermore, as we predict travel order demand to the airport, so it is necessary to extract order data which the order destination points are at Shenzhen Airport. The extraction results are visualized as shown in Fig.1. (blue points are order origin points, and the orange points are order destination points at Shenzhen Airport).



1. Visualization of order demand on August 10, 2015 (orders to Shenzhen Airport)
2. **Methods**

The proposed multi-zone order prediction model aims to predict short-term taxi order demand in multiple zones. The framework of this model is shown in Fig.2. For data collection, we collected information including order ID, order time, order longitude and order latitude and this could provide an adequate basis for the model and analysis below. For model development, we proposed the multi-zone order prediction model to predict short-term order demand in multiple zones. This model contained two parts. The first model was order zone division model. In this model, we firstly proposed cluster validity index called Between-Within Proportion index to determine the optimal number k of zones which was an important parameter of K-Means++ clustering algorithm. Then we used K-Means++ clustering algorithm to divide the whole order zone. The second model was prediction model. In this model, firstly we did influence variables correlation analysis to determine the input variables of prediction model, then we used training set and validation set to estimate parameter for prediction model. Following that, we proposed six prediction models, Locally Weighted Linear Regression (LWLR), Genetic Algorithm-Backpropagation Neural Network (GA-BP), Support Vector Regression (SVR), average fusion based method, weighted fusion based method and kNN fusion-based method to do travel order prediction. The three fusion based methods combine the predicted outputs of LWLR, GA-BP, SVR based on some principles to obtain the final prediction results. For performance analysis, we compare prediction performance of these six different prediction models in different zones by using prediction accuracy indicators including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and three multi-zone weighted accuracy indicators including Multi-zone Weighted MAE (MZW-MAE), Multi-zone Weighted MAPE (MZW-MAPE) and Multi-zone Weighted RMSE (MZW-RMSE) are used to compare the overall performance of these prediction models.



1. Research framework
   1. **Order Zone Division Model**
      1. **Spatial Clustering Algorithm**

Currently, the popular spatial clustering algorithms are K-Means, DBSCAN and other spatial algorithms. The K-Means clustering algorithm aims to partition the n observations into k clusters in so as to minimize the within-cluster sum of squares (Mao, Ji et al. 2016). DBSCAN is a typical density clustering algorithm which clusters based on region density, and the region density must exceed the predefined density threshold in the given radius neighbourhood, so it can find clusters of any shape (Birant and Kut 2007). However, DBSCAN algorithm is slightly more complex than K-Means algorithm. It needs to coordinate the neighbourhood sample number threshold and the distance threshold . The final clustering effect will be different due to the combination of different parameters. Besides, the dataset usually can be very large, so when the DBSCAN clustering algorithm is used, its convergence time can be significantly long. Conversely, K-Means algorithm is more suitable for large data sets, and the algorithm has relative scalability and high efficiency. Its time complexity is ,where n is the number of samples, k is the number of clusters, and t is the number of iterations.

The choice of k initial centroids directly affects the accuracy of the final clustering result and the duration of the algorithm operation, so it is necessary to select the appropriate k centroids. K-Means++ algorithm is an algorithm that optimizes the initialization centroids based on K-Means algorithm. So K-Means++ clustering algorithm is chosen to divide the traffic community.

* + 1. **Determination of Parameter k**

The most critical part of the K-Means++ algorithm is the determination of the value of k. But in actual work, k is difficult to be accurately determined. The indexes used to test the validity of clusters have also been proposed by scholars from various countries, including the Calinski-Harabasz index (CH), the Davies-Boudin index (DB), the Krzanowski-Lai index (KL), Weighted inter-intra index (Wint), In-Group Proportion index (IGP) and so on (Dimitriadou, Dolničar et al. 2002; Dudoit and Fridlyand 2002). Using these indexes to calculate the clustering validity of k clusters in the range of , the optimal number of clusters can be obtained. However, these indexes have defects. When the clustering structure cannot be discriminated, the test results are not ideal enough, so it is difficult to obtain a perfect optimal number of clusters. Zhou S B et al., (2010) proposed a new index according to the geometric structure of the sample, and after a large amount of data verification, the clustering result obtained by this index is better, so for the determination of the value of k, this study choose BWP (Between-Within Proportion) index.

Let be the clustering space, where , n is the number of samples, and c is the number of clusters.

Definition 1: Let the minimum interclass distance of the th sample of the th class be expressed as the minimum of the average distance of the sample to all samples in each of the other classes, given by:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where k and i are class labels, is the th sample of the th class, is the th sample of the th class, is the number of samples in the kth class, is square Euclidean distance.

Definition 2: Let the minimum intra-class distance of the th sample of the th class be expressed as the average distance of the sample to all the other samples in th class, given by:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where is the th sample of the th class, and , is the number of samples in the th class.

Definition 3: Let clustering distance of the th sample of the th class be expressed as the sum of the minimum inter-class distance and intra-class distance of the sample, given by:

(3)

Definition 4: Let clustering subtraction distance of the th sample of the th class be expressed as the difference between the minimum inter-class distance and intra-class distance of the sample, given by:

(4)

Definition 5: Let Between-Within Proportion index of the th sample of the th class be expressed as the ratio of the clustering subtraction distance to the clustering distance of the sample (see equation(5)):

(5)

From the perspective of the intra-class distance, the smaller the value of is, the better the result is. From the perspective of the inter-class, the biggest the value of is, the better the result is. In order to achieve equilibrium in both two cases, linear combination is chosen to consider both requirements. Clustering subtraction distance can be used to evaluate the clustering result. The bigger the value of is, the better the clustering effect is. At the same time, in order to reduce the influence of dimension on clustering, clustering distance is introduced. The index can be made dimensionless through compressing by , thus the range of the value of index is [-1,1]. If , it indicates that the sample is correctly clustered. If , it indicates that the sample is incorrectly clustered.

Index only reflects the clustering of a certain sample, and it does not reflect the clustering of all the samples. However, if the average of the of all the samples in the data set is obtained, the clustering effect of the data set can be reflected. The larger the average value is, the better the clustering effect of the data set is, and among them, the number of clusters corresponding to the maximum value is the optimal clustering number, given by:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |
|  |  | (7) |

where is the average value of BWP obtained when the sample set is clustered into k classes, is the optimal clustering number.

* 1. **Order Prediction Model**
     1. **Locally Weighted Linear Regression**

Locally Weighted Linear Regression is a nonparametric learning algorithm proposed by W. S. Cleveland in 1979 (Cleveland 1979). When predicting new sample values, the training samples need to be trained to update the value of the parameter for each prediction. The values of new samples need to be predicted based on the training data sample set, so the values of the model parameter in each iteration are uncertain. The model allocates weights to each point near the point to be predicted based on the principle that the closer the distance, the greater the weight. Then the ordinary linear regression is performed using the minimization criterion of mean square error to predict the predicted value of the point to be predicted. The wavelength parameter k affects the accuracy of the model, a reasonable value of k is conducive to the improvement of prediction accuracy. In the optimisation process, the trial and error method is used, so the optimal wavelength parameter can be set.

* + 1. **Genetic Algorithm Back Propagation**

BP (Back Propagation) Neural Network is a multi-layer feedforward neural network based on error back propagation (Chauvin and Rumelhart 2013). The basic idea is to use the network mean square error as the objective function. Based on the gradient descent strategy, the parameters are adjusted in the negative gradient direction of the target to minimize the error mean square error between the expected output value and the true value. BP Neural Network model has the advantage of approximating non-linear function with arbitrary precision, but it also has the disadvantage of being easily trapped in local optimum. While GA-BP can use genetic algorithm to make the solution jump out of the local optimum and approximate the global optimum to optimize the weights and thresholds of the BP neural network, so the prediction results can be more accurate. Cross-validation for the selection of parameters is used in this paper. After minimizing the Mean Absolute Percentage Error in the process of optimization, the number of hidden layer could be determined.

* + 1. **Support Vector Regression**

Support Vector Regression (SVR) is a popular machine learning method that emerged at the end of the 20th century. It is mainly used for classification and prediction, and has good generalization ability. It is proposed by the world-renowned scholar Vapnik (Vapnik 2013). The SVR continuously adjusts the parameters by training the samples to derive a model that minimizes the sum of the deviations between the predicted and true values of all training samples. By inputting the predicted input vector into the model, the prediction can be made. In this paper, a Radial Basis Function (RBF) is used as the kernel function because it is shown to be more suitable for traffic prediction under different conditions (Guo, Krishnan et al. 2017). After optimising, the capacity values C of SVR can be determined and the -insensitive loss function is used in this research.

* + 1. **Average fusion based method**

Average fusion-based method: The average of the prediction results of each individual predictor is taken as the final result. Given a set of predictors, , we seek to compute final prediction , given by:

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

where is the prediction result using predictor.

* + 1. **Weighted fusion based method**

In this method, weights of different predictors are not same. Weighted hybrid method is written as:

|  |  |  |
| --- | --- | --- |
|  |  | (10) |
|  |  | (11) |

where is the weight of predictor. Weights are calculated using training dataset. In this paper, the weights are calculated by the inverse of Mean Absolute Percentage Error (MAPE).

* + 1. **kNN fusion based method**

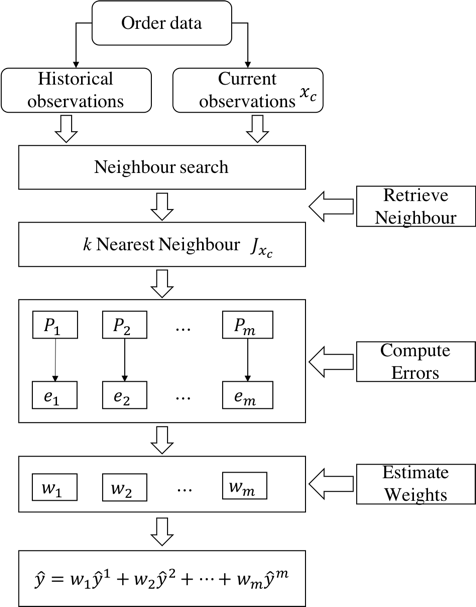
The kNN fusion based method is highly unstructured and does not require any pre-determined model specification. The basic idea of kNN fusion-based method is, under the circumstance of current traffic state, to search the nearest neighbour to this state in the training used historical datasets to compute the prediction errors of the nearest neighbour set, estimate weights of each predictor, and combine the final predicted outputs of each individual predictor based on these weights (Guo, Polak et al. 2018). Fig. 3 depicts the flowchart of the kNN fusion-based method. There are two steps in the kNN fusion-based method.

**Step 1: Neighbourhood searching process:**

The search process finds the nearest neighbours, which are the historical observations that are most similar to the current observation. Euclidean distance is used in this paper to determine the distance between the current input feature vector and historical observations. k is the number of historical observations with the nearest distances to the input feature vector. The set of k nearest neighbours of the input feature vector can be written as and , and n is the dimension of feature space. The trial and error method introduced in Guo et al. (2012) is used for setting the parameters of kNN. k is set to 5 in this paper.

**Step 2: Weighted parameter estimation process:**

This process is used to calculate weights of each predictor. For each vector , the predicted value and error of each predictor can be calculated. Hence, errors can be used to estimate the weights of each predictor at current time, , where MAPE is calculated based on the selected nearest neighbor dataset. The main difference between weighted fusion-based method and kNN fusion-based method is the weighted used in kNN fusion-based method is dynamically updated in every step.



1. Flowchart based on kNN fusion based method
   1. **Prediction Accuracy Indicator**

In order to better compare and analyze actual prediction performance and effect of different prediction models, predictors need to be evaluated and analyzed. In this paper, some prediction performance evaluation indicators are adopted, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), given by:

|  |  |  |
| --- | --- | --- |
|  |  | (12) |
|  |  | (13) |
|  |  | (14) |

where is the total sample size, is the predicted order value , is the observed value.

According to indicators above, the predicted results of LWLR prediction model, GA-BP neural network prediction model, SVR prediction model, average fusion based prediction model, weighted fusion based prediction model and fusion based prediction model are analyzed and evaluated.

In order to better compare the performance of different prediction models in terms of multi-zone prediction, we then propose multi-zone weighted indicators based on MAE, MAPE, RMSE to evaluate the overall prediction performance of these prediction models in all the zones, including multi-zone weighted MAE indicator (MZW-MAE), multi-zone weighted MAPE indictor (MZW-MAE), multi-zone weighted RMSE indictor (MZW-RMSE), given by:

|  |  |  |
| --- | --- | --- |
|  |  | (15) |
|  |  | (16) |
|  |  | (17) |

where O is the number of order demand in all zones, is the number of order demand in zone k, , , are the prediction accuracy indicators for zone k.

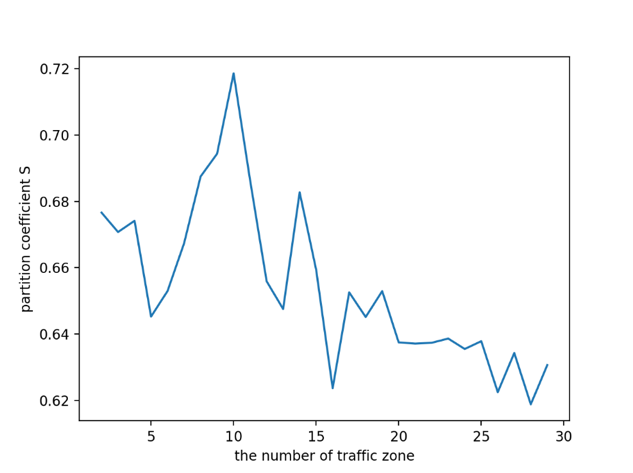
At the same time, a new evaluation rule is defined. When the predicted value of more than half of the predictors in a certain period is less than 7, the time period does not enter the evaluation range, because if the predicted demand is lower than 7, it indicates that during this period few people need travel in taxi, thus the prediction for this period is useless.

1. **Experiment**

**4.1 Zone Division**

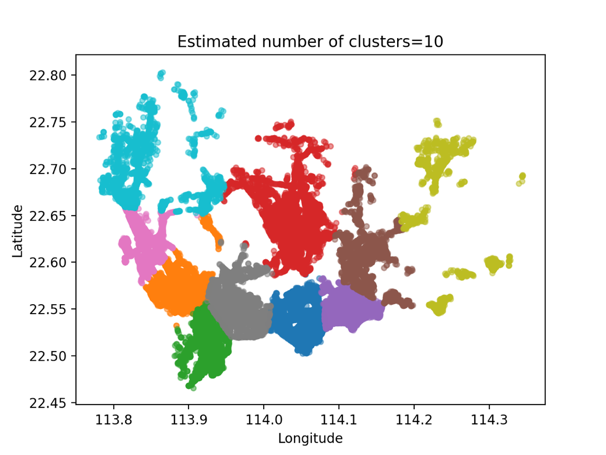
The order demand to Shenzhen Airport differ between working days and non-working days, and passenger travel demands have a strong regularity in working days. Therefore, this section only divides the working days’ order zone as case study.

The range of the value of k is limited to [2, 30] by experience and actual conditions. After calling the K-Means++ algorithm, the k-value curve is obtained (shown in Fig.4).



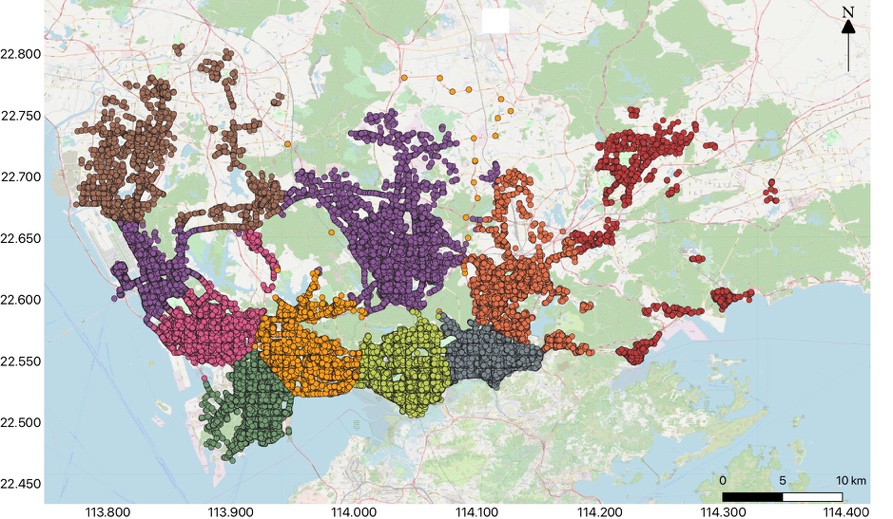
1. Curve of BWP value with k value

As can be seen from Fig.4 when the value of k is 10, the value of BWP is the largest, which is 0.71826. Hence, the number of order area in this paper is determined to be 10.



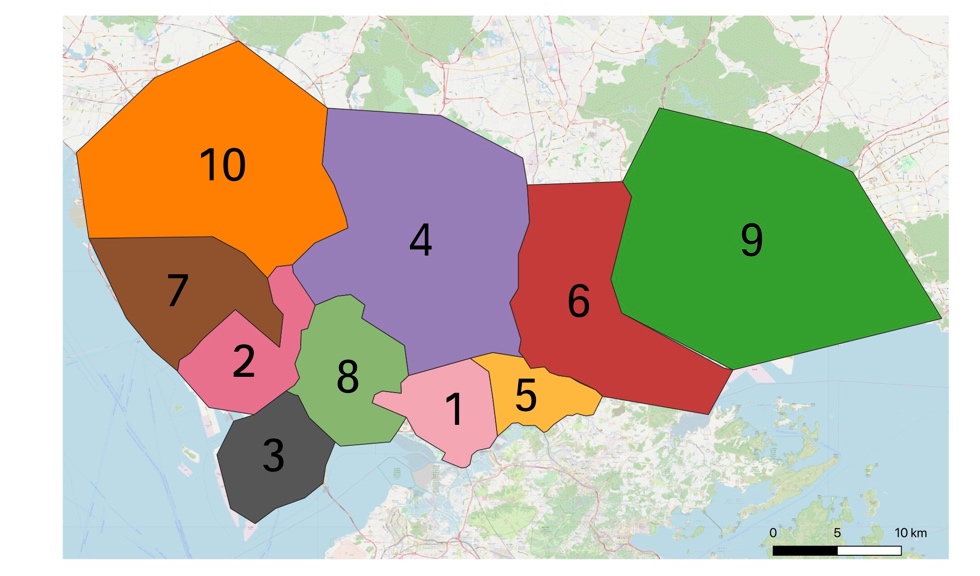
1. Clustering result graph with k=10

Fig.5 is a clustering result graph, and after visualizing the clustering data, Fig.6 can be gotten.



1. The visualization of clustering result

Then, we connect the boundary points and fine-tune them based on the principle that one zone does not cross the main road as soon as possible, because when a driver cruises to find the passengers, they do not need to cross the main road, which could save cruise time. The order zones are fine-tuned and numbered, and Fig.7 is obtained.

**

1. Order zones are adjusted and numbered

**4.2 Influence Variable Correlation Analysis**

This study takes the taxi order demand to Shenzhen Airport on Tuesday as an example to predict the travel order demand to Shenzhen Airport in different zones within 60 minutes on this day. We select the GPS data of all Shenzhen taxi orders from September 1, 2015 to September 31, 2015 and calculate the order demand to the airport within time interval of 60 minutes, and then do the correlation analysis of the factors that influence travel demand to the airport by taxi.

(1) Travel demand correlation analysis of different periods on a certain working day. is the taxi travel demand to the airport during period of day . The correlation analysis is shown in Table 1.

1. Travel demand correlation analysis of different periods on a certain working day

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cc |  |  |  |  |  |  |
|  | 1 | 0.846 | 0.781 | 0.749 | 0.741 | 0.672 |
|  | 0.846 | 1 | 0.846 | 0.781 | 0.749 | 0.741 |
|  | 0.781 | 0.846 | 1 | 0.846 | 0.781 | 0.749 |
|  | 0.749 | 0.780 | 0.846 | 1 | 0.846 | 0.781 |
|  | 0.741 | 0.749 | 0.781 | 0.846 | 1 | 0.846 |
|  | 0.672 | 0.741 | 0.749 | 0.781 | 0.846 | 1 |

With 60 minutes as time interval, a day can be equally divided into 24 periods. denotes taxi order demand during period and denotes taxi order demand during period and so on. According to the correlation coefficient from Table 1, it can be known that the demand of the current period has a higher correlation with the demand of the adjacent period, and the correlation with the demand during period is up to 0.75. Therefore, when predicting the taxi travel demand, we can select and as input variables.

(2) Travel demand correlation analysis of the same period in different days. We choose Tuesday as an example. denotes travel demand of the previous working day , which is Monday, denotes travel demand 2 days before, which is Sunday and so on. Correlation analysis is shown in Table 2.

1. Travel demand correlation analysis of the same period in different working days

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cc |  |  |  |  |  |  |  |  |  |
|  | 1 | 0.927 | 0.868 | 0.867 | 0.903 | 0.904 | 0.917 | 0.925 | 0.892 |
|  | 0.927 | 1 | 0.876 | 0.875 | 0.906 | 0.902 | 0.936 | 0.946 | 0.927 |
|  | 0.868 | 0.876 | 1 | 0.950 | 0.897 | 0.818 | 0.840 | 0.876 | 0.818 |
|  | 0.867 | 0.875 | 0.950 | 1 | 0.897 | 0.841 | 0.881 | 0.889 | 0.875 |
|  | 0.903 | 0.906 | 0.897 | 0.897 | 1 | 0.906 | 0.920 | 0.904 | 0.875 |
|  | 0.904 | 0.902 | 0.818 | 0.841 | 0.906 | 1 | 0.913 | 0.901 | 0.917 |
|  | 0.917 | 0.936 | 0.840 | 0.881 | 0.920 | 0.913 | 1 | 0.909 | 0.893 |
|  | 0.925 | 0.946 | 0.876 | 0.889 | 0.904 | 0.901 | 0.909 | 1 | 0.924 |
|  | 0.892 | 0.897 | 0.818 | 0.875 | 0.915 | 0.917 | 0.893 | 0.924 | 1 |

According to the correlation coefficient from Table 2, the correlations of the same time in different working days are quite different. The correlation between Tuesday and Saturday is around 0.86, as well as Sunday, but the correlation with adjacent working days is as high as 0.9, and the correlation with last Tuesday is up to 0.925, but it is slightly lower with the Tuesday before last Tuesday. Therefore, when predicting order demand, we can choose , , , , as input variables.

Overall, after correlation analysis, we choose , , , , , , as input variables.

**4.3 Prediction Results**

The data of all working day attributes from August 10, 2015 to October 18, 2015 were selected as a training dataset, and the data of all working day attributes from October 19, 2015 to October 23, 2015 were sorted out as a testing dataset. A day is equally divided into 24 periods, and the whole area is divided into 10 zones. Taxi order demand during each period and each zone on October 20, 2015 can be predicted as an example by six prediction models as mentioned above. Comparisons of predicted and observed taxi demand of each zone are shown in Fig.8-Fig.17.

|  |  |
| --- | --- |
|  |  |
| 1. Predicted and observed results in zone 1 | 1. Predicted and observed results in zone 2 |
|  |  |
| 1. Predicted and observed results in zone 3 | 1. Predicted and observed results in zone 4 |
|  |  |
| 1. Predicted and observed results in zone 5 | 1. Predicted and observed results in zone 6 |
|  |  |
| 1. Predicted and observed results in zone 7 | 1. Predicted and observed results in zone 8 |
|  |  |
| 1. Predicted and observed results in zone 9 | 1. Predicted and observed results in zone 10 |

It can be seen from Fig.8 to Fig.17 that the prediction curves of the various prediction models are generally consistent with the overall trend of the actual observations. During the peak period (5:00-8:00), each individual prediction model including LWLR, GABP and SVR has a good performance with no obvious deviation. Despite this, they have different performances in other time periods, which sometimes have good performance and sometimes have poor performance. Fusion based prediction models including average fusion based method, weighted fusion based method and kNN fusion based method combine the advantages of three individual prediction models and slightly perform better than individual prediction models in other periods (i.e. 10:11:00 in zone 1, 9:00-10:00 in zone 2). At the same time, as we can see from Fig.16 and Fig.17, taxi demand to the airport in zone 9 and zone 10 is very low. According to Fig. 7, zone 9 and zone 10 are located far from Shenzhen Airport, and they are not densely populated areas and commercial developed areas, so the demand to the airport by taxi can be very low.

**4.4 Prediction Accuracy Analysis**

To better compare and analyse the prediction performances of different prediction models, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) are adopted. Furthermore, multi-zone weighted MAE indicator (MZW-MAE), multi-zone weighted MAPE indictor (MZW-MAE), multi-zone weighted RMSE indictor (MZW-RMSE) are adopted to evaluate the overall prediction performance of these prediction models in all the zones (shown in Table 3-Table 5, Fig.9-Fig 12). Meanwhile, according to the newly defined evaluation rule (see Sec 2.3), the demand in zone 9 and zone 10 during each period is less than 7 people, so the analysis of these two zones is worthless, and thus not included in Table 3, Table 4 and Table 5.

1. Prediction accuracy of different methods

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Zone1 | | |  | Zone2 | | |  | Zone3 | | |
| MAE | MAPE | RMSE |  | MAE | MAPE | RMSE |  | MAE | MAPE | RMSE |
| LWLR | 5.468 | | 20.3% | 6.821 |  | 2.685 | 30.4% | 3.093 |  | 5.931 | 30.7% | 7.128 |
| GABP | 5.382 | | 20.6% | 6.637 |  | 2.382 | 26.9% | 2.744 |  | 5.487 | 26.4% | 6.248 |
| SVM | 5.926 | | 21.6% | 7.104 |  | 2.296 | 24.6% | 2.648 |  | 6.154 | 31.8% | 7.276 |
| Average | 5.200 | | 19.5% | 6.490 |  | 2.320 | 27.1% | 2.712 |  | 5.634 | 29.6% | 6.611 |
| Weighted | 5.183 | | 19.4% | 6.476 |  | 2.276 | 25.7% | 2.628 |  | 5.488 | 29.2% | 6.338 |
| kNN fusion | **5.155** | | **19.3%** | **6.416** |  | **2.201** | **24.5%** | **2.604** |  | **5.056** | **26.3%** | **6.192** |

1. Prediction accuracy of different methods(continued)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Zone4 | | |  | Zone5 | | |  | Zone6 | | |
| MAE | MAPE | RMSE |  | MAE | MAPE | RMSE |  | MAE | MAPE | RMSE |
| LWLR | 5.059 | | 62.2% | 6.748 |  | 4.109 | 18.4% | 4.870 |  | 3.203 | 23.4% | 4.366 |
| GABP | **3.521** | | **47.9%** | **4.363** |  | 5.056 | 20.3% | 4.622 |  | 5.282 | 31.5% | 5.683 |
| SVM | 4.572 | | 57.2% | 6.005 |  | **3.246** | **16.7%** | **3.630** |  | 3.691 | 24.0% | 4.498 |
| Average | 4.419 | | 60.3% | 5.768 |  | 3.657 | 17.6% | 4.346 |  | 3.944 | 26.1% | 4.713 |
| Weighted | 3.644 | | 57.1% | 4.903 |  | 3.505 | 17.6% | 3.977 |  | 3.466 | 24.6% | 4.381 |
| kNN fusion | 4.096 | | 56.2% | 5.095 |  | 3.475 | 17.5% | 3.874 |  | **3.136** | **22.6%** | **4.321** |

1. Prediction accuracy of different methods(continued)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Zone7 | | |  | Zone8 | | |  | MZW-MAE | | MZW-MAPE | | MZW-RMSE |
| MAE | MAPE | RMSE |  | MAE | MAPE | RMSE |  |
| LWLR | 3.613 | | 43.5% | 4.521 |  | 4.915 | 42.7% | 6.561 |  | 4.687 | 30.491% | | 5.831 | |
| GABP | 3.232 | | 39.7% | 3.933 |  | **3.992** | **31.1%** | **5.296** |  | 4.576 | 27.579% | | 5.260 | |
| SVM | 3.459 | | 38.9% | 4.164 |  | 4.401 | 35.6% | 5.915 |  | 4.578 | 28.664% | | 5.528 | |
| Average | 3.573 | | 42.6% | 4.334 |  | 4.650 | 40.5% | 6.290 |  | 4.417 | 29.339% | | 5.435 | |
| Weighted | 3.498 | | 41.4% | 4.220 |  | 4.158 | 33.0% | 5.742 |  | 4.217 | 27.905% | | 5.165 | |
| kNN fusion | **3.166** | | **36.7%** | **3.919** |  | 3.998 | 32.2% | 5.355 |  | **4.095** | **26.579%** | | **5.050** | |

|  |  |
| --- | --- |
|  |  |
| 1. Prediction accuracy analysis with MAE | 1. Prediction accuracy analysis with MAPE |
|  |  |
| 1. Prediction accuracy analysis with RMSE | 1. Overall prediction performance analysis |

In Zone 1, the values of MAPE using average and weighted fusion based methods are 19.5% and 19.4%, while the value of the MAPE using kNN fusion based method is 19.3%. Meanwhile, compared with average and weighted fusion based method, the value of MAE and RMSE using kNN fusion based method are relatively lower than other prediction models, which are 5.155 and 6.416 respectively. In the same way, kNN fusion based method gives the most accurate results than other prediction models in Zone 2 (e.g. 2.604 orders/h of RMSE), Zone 3 (e.g. 6.192 orders/h of RMSE), Zone 6 (e.g. 4.321 orders/h of RMSE), Zone 7 (e.g. 3.919 orders/h of RMSE).

In Zone 4 and Zone 8, GABP is better. But the performance of the kNN fusion based method is also significantly better than other prediction models in Zone 4. Compared with MAPE of average fusion based method, the improvement of kNN fusion based method is 4%. Besides, the performance of the fusion based method is second only to that of GABP in Zone 8, 3.998 orders/h of MAE, 32.2% of MAPE, 5.355 orders/h of RMSE. By contrast, in Zone 5, the best performing prediction model is SVR, but the second is kNN fusion based method, the performance of kNN fusion based method is only 0.8% difference from that of SVR in terms of MAPE.

In terms of overall prediction performance analysis, the values of MZW-MAE, MZW-MAPE, MZW-RMSE of kNN fusion based method are lowest in these six prediction methods, which are 4.095 orders/h, 26.579% and 5.05 orders/h respectively and this indicates that kNN fusion based method could give a better prediction result in multi-zone prediction.

From the viewpoint of fusion methods, the results in Table 3, Table 4, Table 5 indicate that simple fusion method including average fusion based method and weighted fusion method only provide moderate improvements to the overall prediction accuracy, but kNN fusion based method can have a better prediction performance in this case of predicting short-term order demand. For example, in Zone 7, the MAPE value is reduced to 36.7% from 38.9% which is the best result using the individual prediction model, in contrast with 42.6% in average fusion based method and 41.4% in weighted fusion based method.

1. **Conclusion**

In order to help facilitate city-scale travel operation management in megacity, this paper proposes a multi-zone order demand prediction model to divide order zones through K-Means++ spatial clustering algorithm, and predict order demand in different divided zones based on six different prediction models, including kNN fusion based method, LWLR, GA-BP Neural Network, SVR, average fusion based method and weighted fusion based method. This study takes taxi order demand to Shenzhen International Airport as a case study to divide the order area and predict the order demand in different zones. The result indicates that it is effective to use multi-zone travel demand prediction model to divide the order zones and predict order demands. According to the systematic comparison analysis, GA-BP Neural Network, SVR and kNN fusion based method have relatively good predictive performance in individual zones based on three prediction accuracy indicators (MAE, MAPE, RMSE). However, according to multi zone weighted indicators, MZW-MAE, MZW-MAPE and MZW-RMSE, the overall prediction performance of kNN fusion based method for multi-zone order demand is the best. Overall, we can support that in the case of city-scale order prediction, using multi-zone prediction model with kNN fusion based method can be effective. And it could be suggested, the multi-zone prediction model with kNN fusion based method proposed in this study can be served as a basis of scheduling optimization at city-scale. However, limited by data availability, our case study is special, and in the future work, we would use data from more scenarios for verification, and then get more comprehensive conclusion.

1. **Reference**

Birant, D. and A. Kut (2007). "ST-DBSCAN: An algorithm for clustering spatial–temporal data." Data & Knowledge Engineering 60(1): 208-221.

Chang, H.-w., Y.-c. Tai and J. Y.-j. Hsu (2010). "Context-aware taxi demand hotspots prediction." International Journal of Business Intelligence and Data Mining 5(1): 3.

Chauvin, Y. and D. E. Rumelhart (2013). Backpropagation: theory, architectures, and applications, Psychology Press.

Cleveland, W. S. (1979). "Robust locally weighted regression and smoothing scatterplots." Journal of the American statistical association 74(368): 829-836.

Davis, N., G. Raina and K. Jagannathan (2018). "Taxi demand forecasting: A HEDGE-based tessellation strategy for improved accuracy." IEEE Transactions on Intelligent Transportation Systems 19(11): 3686-3697.

Dimitriadou, E., S. Dolničar and A. Weingessel (2002). "An examination of indexes for determining the number of clusters in binary data sets." Psychometrika 67(1): 137-159.

Ding, M.-j., S.-z. Zhang, H.-d. Zhong, Y.-h. Wu and L.-b. Zhang (2019). "A Prediction Model of the Sum of Container Based on Combined BP Neural Network and SVM." Journal of Information Processing Systems 15(2).

Dudoit, S. and J. Fridlyand (2002). "A prediction-based resampling method for estimating the number of clusters in a dataset." Genome biology 3(7): research0036. 0031.

Feng, X., X. Ling, H. Zheng, Z. Chen and Y. Xu (2018). "Adaptive Multi-Kernel SVM With Spatial–Temporal Correlation for Short-Term Traffic Flow Prediction." IEEE Transactions on Intelligent Transportation Systems 20(6): 2001-2013.

Guo, F., R. Krishnan and J. Polak (2012). "Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbour method."

Guo, F., R. Krishnan and J. Polak (2017). "The influence of alternative data smoothing prediction techniques on the performance of a two-stage short-term urban travel time prediction framework." Journal of Intelligent Transportation Systems 21(3): 214-226.

Guo, F., J. W. Polak and R. Krishnamoorthy (2018). "Predictor fusion for short-term traffic forecasting."

Guo, F., J. W. Polak and R. Krishnan (2010). Comparison of modelling approaches for short term traffic prediction under normal and abnormal conditions. 13th International IEEE Conference on Intelligent Transportation Systems, IEEE.

Habtemichael, F. G. and M. Cetin (2016). "Short-term traffic flow rate forecasting based on identifying similar traffic patterns." Transportation research Part C: emerging technologies 66: 61-78.

Li, X., G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang and Z. Wang (2012). "Prediction of urban human mobility using large-scale taxi traces and its applications." Frontiers of Computer Science 6(1): 111-121.

Ma, X., Z. Dai, Z. He, J. Ma, Y. Wang and Y. Wang (2017). "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction." Sensors 17(4): 818.

Mao, F., M. Ji and T. Liu (2016). "Mining spatiotemporal patterns of urban dwellers from taxi trajectory data." Frontiers of Earth Science 10(2): 205-221.

Moreira-Matias, L., J. Gama, M. Ferreira, J. Mendes-Moreira and L. Damas (2013). "Predicting taxi–passenger demand using streaming data." IEEE Transactions on Intelligent Transportation Systems 14(3): 1393-1402.

Mukai, N. and N. Yoden (2012). Taxi demand forecasting based on taxi probe data by neural network. Intelligent Interactive Multimedia: Systems and Services, Springer: 589-597.

Nikravesh, A. Y., S. A. Ajila, C.-H. Lung and W. Ding (2016). Mobile network traffic prediction using MLP, MLPWD, and SVM. 2016 IEEE International Congress on Big Data (BigData Congress), IEEE.

Qiu, H., R. Li and H. Liu (2016). "Integrated model for traffic flow forecasting under rainy conditions." Journal of Advanced Transportation 50(8): 1754-1769.

Vapnik, V. (2013). The nature of statistical learning theory, Springer science & business media.

Vlahogianni, E. I. (2015). "Optimization of traffic forecasting: Intelligent surrogate modeling." Transportation Research Part C: Emerging Technologies 55: 14-23.

Vlahogianni, E. I., M. G. Karlaftis and J. C. Golias (2014). "Short-term traffic forecasting: Where we are and where we’re going." Transportation Research Part C: Emerging Technologies 43: 3-19.

Yao, H., F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye and Z. Li (2018). Deep multi-view spatial-temporal network for taxi demand prediction. Thirty-Second AAAI Conference on Artificial Intelligence.

Yu, R., Y. Li, C. Shahabi, U. Demiryurek and Y. Liu (2017). Deep learning: A generic approach for extreme condition traffic forecasting. Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM.

Zhou, S.-B., Z.-Y. Xu and X.-Q. Tang (2010). "New method for determining optimal number of clusters in K-means clustering algorithm." Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications) 46(16).