Multi-Zone Taxi Order Demand Prediction

Pengshun Li
lps19@mails.tsinghua.edu.cn

Chenxi Cheng
ccx19@mails.tsinghua.edu.cn

Chenyu Tian
tiancy19@mails.tsinghua.edu.cn

ABSTRACT

Travel order demand prediction is of great importance for the improvement of intelligent transportation system to city-scale 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 project, 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. For the prediction step, six methods, LWLR, GA-BP Neural Network, SVR, average fusion-based method, weighted fusion-based method and kNN fusion based method are compared. These models were performed and validated on a real-world taxi order demand data from 3-month consecutive collection in Shenzhen, China.

1 INTRODUCTION

Traffic has always been an important factor in the city management and operation which impacts the daily life of millions of dwellers. Nowadays, with the demand for upgrading the intelligent transportation system to provide more smart and personalized services, travel demand is needed to be calculated in real-time for city-scale system operation because most of the traffic inefficiency are caused by imbalanced travel demand and supply. The better we can predict travel order demand, the better can help the city pre-allocate resources and facilitate city-scale travel operation management in megacity. 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. To conclude, predicting travel demand can not only help individuals to save time and get more benefits but also helps to improve the efficiency of the urban traffic system.

Based on that, an accurate short-term multi-zone travel demand prediction model is needed. In our project, a two-step methodology was developed including order zone division and multi-zone order prediction. In the first step, the input to our algorithm is the geographic coordinates of taxi orders. We then use the K-means++ to output clustered zones. In the second step, the input to our algorithm is the historical taxi order demand data. We then use multiple methods (LWLR, GA-BP Neural Network, SVR, average fusion-based method, weighted fusion-based method, kNN fusion based method) to predict the taxi demand in these zones.

2 RELATED WORK

Actually, there is no exact same but only similar tasks in the literatures. There have been many prediction studies 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[7]. 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. Habtemichael and Cetin[5] proposed an enhanced k-Nearest Neighbours algorithm for short-term traffic forecasting and it indicated that the proposed method can provide promising results. Nikravesh[8] 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. But individual prediction methods do not provide good enough predictive performance.

In order to further reduce prediction error, in recent years, some researchers research methods to combine prediction models to improve prediction accuracy[3]. Qiu proposed an integrated precipitation-correction model to use fusion method with four prediction models to predict freeway traffic flow[9]. Vlahogianni[11] 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 prediction accuracy.

According to the literature, it can be found that many studies 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 the spatial effect on predictive performance. For example, some researches do research on travel order demand prediction for hot spot analysis or grids. Li[7] analyzed the spatial-temporal variation of passengers in a hot spot. Ke[6] used a state-of-the-art 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 the 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, which is the biggest uniqueness of our project.

3 DATASET AND FEATURES

3.1 Data Description

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 as shown in *Table 1*. 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

Learning from Data Project Report, December 31, 2019, Shenzhen, China
.

Table 1: Data Description

Attribute	Description	Example
number_id	car license id	$**0115H$
on_time	boarding time	2015-08-03T04:47:52.000Z
on_GPS	boarding GPS time	2015-08-03T04:47:56.000Z
on_difference	time difference	4
on_longitude	boarding longitude	113.79980944
on_latitude	boarding latitude	22.69250807
off_time	drop-off time	2015-08-03T05:05:40.000Z
off_GPS	drop-off GPS time	2015-08-03T05:05:36.000Z

the city. The latitude and long 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.

This project 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. With 60 minutes as time interval, a day can be equally divided into 24 periods.

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. And cross-validation is used to select parameters.

3.2 feature Selection

Do the correlation analysis of the factors that influence travel demand to the airport by taxi to better select input features as shown in *Figure 1*. Here define $D_n(t)$ is the taxi travel demand to the airport during period $[t, t+1)$ of day n. Based on the correlation matrix, on a certain working day, the demand of the current period $D_n(t)$ has a higher correlation with $D_n(t-1), D_n(t-2)$, which are selected as features. And for the same time period in different days, the correlations of the same time in different working days are quite different. Therefore, when predicting order demand, we can choose $D_{n-1}(t), D_{n-4}(t), D_{n-5}(t), D_{n-6}(t), D_{n-7}(t)$ as input variables.

4 METHODS

4.1 Zone Division

We have learnt the Kmeans algorithm in the class. KMeans++ algorithm is an algorithm that optimizes the initialization centroids based on K-Means algorithm. But for this algorithm, the number of clusters need to be prefined. It may have some problems when have no knowledge about the number. Thus, we refered to BWP (Between-Within Proportion) index to select optimal number of clusters. The basic idea is to find small intra-class distance as well as big inter-class distance. The larger the average BWP 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. Using this index to calculate the value of k clusters in the range of $[k_{min}, k_{max}]$, the optimal number of clusters can be obtained.

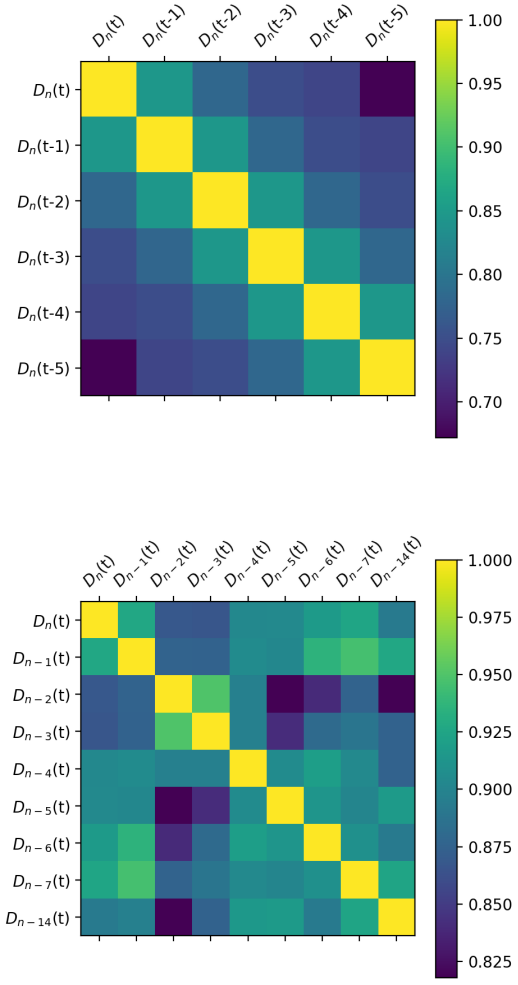


Figure 1: Correlation analysis

4.2 Order Prediction

In this part, six prediction models are trained and tested.

4.2.1 Locally Weighted Linear Regression. Locally Weighted Linear Regression is a nonparametric learning algorithm[2]. 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.

4.2.2 Genetic Algorithm Back Propagation. BP Neural Network is a multi-layer feedforward neural network based on error back propagation[1]. 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 study. After minimizing the Mean Absolute Percentage Error in the process of optimization, the number of hidden layer could be determined.

4.2.3 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[10]. 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 project, 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[3]. After optimising, the capacity values C of SVR can be determined and the insensitive loss function is used in this research.

4.2.4 Average Fusion based Method. Average fusion-based method (or blending) gives the average of the prediction results of each individual predictor. Given a set of predictors $\hat{y}^i$, we seek to compute final prediction , given by

$$\hat{y} = \frac{1}{m}\left(\hat{y}^1 + \hat{y}^2 + \ldots + \hat{y}^m\right) \tag{1}$$

4.2.5 Weighted fusion based method. In this method, or called bagging, weights of different predictors are not same. Weighted hybrid method is written as:

$$\begin{gathered}\hat{y} = \alpha_1\hat{y}^1 + \alpha_2\hat{y}^2 + \ldots + \alpha_m\hat{y}^m = \textstyle\sum_i \alpha_i\hat{y}^i \\ \alpha_1 + \alpha_2 + \ldots + \alpha_m = 1\end{gathered} \tag{2}$$

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

4.2.6 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 neighbours to this state in the training used historical dataset 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[4]. There are two steps in the kNN fusion-based method.

The first step is to find the nearest neighbours, which are the historical observations that are most similar to the current observation. Euclidean distance is used in this study 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 second step is to calculate weights of each predictor. For each vector x, the predicted value $\hat{y} = \sum_i \alpha_i\hat{y}^i$, which has the same form as weighted fusion based method. The main difference between them is the weighted used in kNN fusion-based method is dynamically updated in every step calculated by the inverse of MAPE on the selected nearest neighbor dataset.

4.3 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 work, some prediction performance evaluation indicators are adopted, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE). And 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) which are tailored for this task:

$$\begin{aligned}MZW\text{-}MAE &= \textstyle\sum_k \frac{O_k}{O} MAE_k \\ MZW\text{-}MAPE &= \textstyle\sum_k \frac{O_k}{O} MAPE_k \\ MZW\text{-}RMSE &= \textstyle\sum_k \frac{O_k}{O} RMSE_k\end{aligned} \tag{3}$$

where O is the number of order demand in all zones, O_k is the number of order demand in zone k, MAE_k, $MAPE_k$, $RMSE_k$ 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.

5 EXPERIMENTS

5.1 Zone Divison

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 BWP value curve is obtained

as *Figure 2*. When the value of k is 10, the value of BWP is the

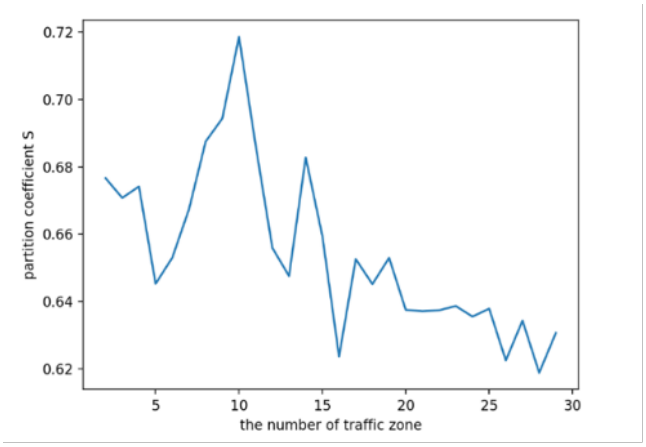


Figure 2: Curve of BWP value with k value

largest. Hence, the number of order area in this task is determined to be 10. *Figure 3* is a clustering result graph, and after visualizing the clustering data. Then, we connect the boundary points and

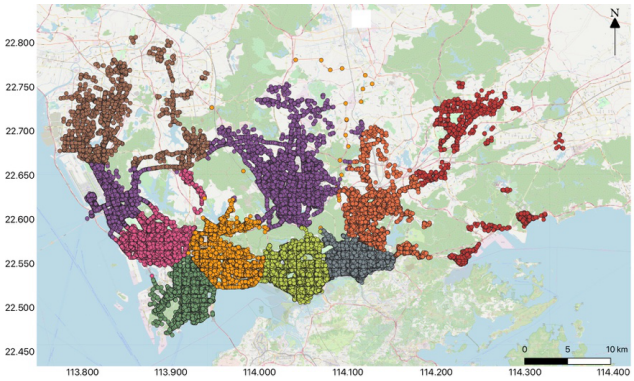


Figure 3: The visualization of clustering result

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.3 is obtained.

5.2 Order Prediction

It can be seen from *Figure5* that the prediction curves of the various prediction models are generally consistent with the overall trend of the actual observations (Instead of showing all the figures of 10 zones, here just show prediction in zone 1). 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

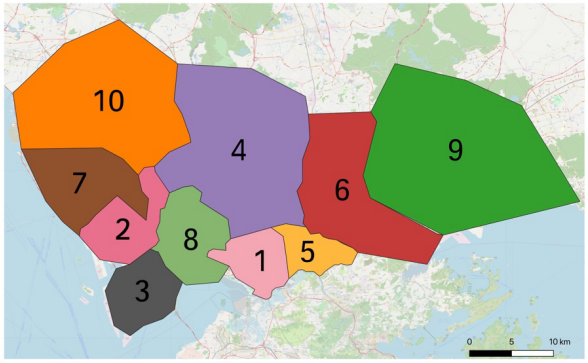


Figure 4: Order zones are adjusted and numbered

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). To better compare

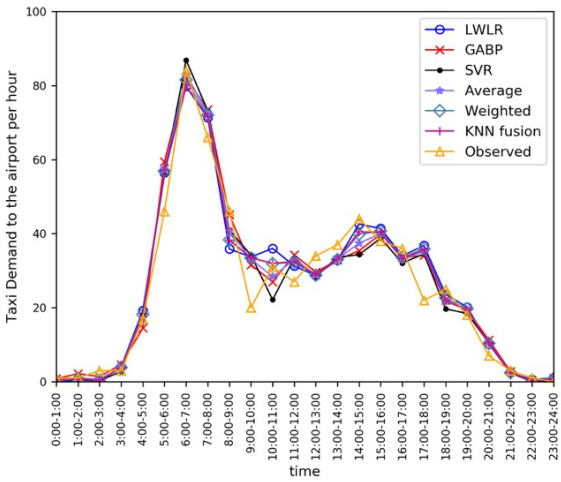
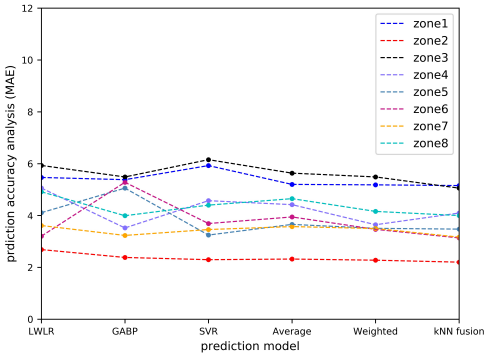
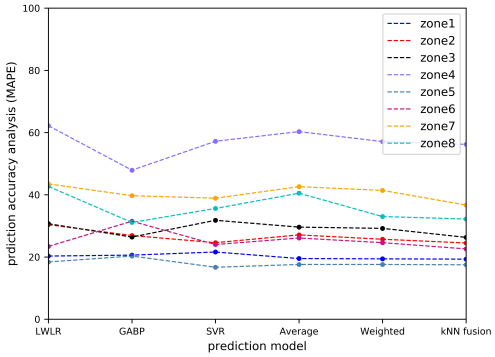


Figure 5: Predicted and observed results in zone 1

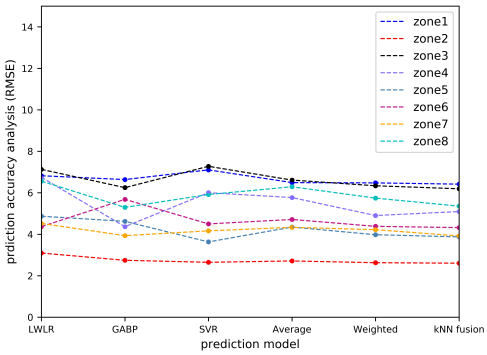
and analyse the prediction performances of different prediction models, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) are adopted. Furthermore, multi-zone weighted MAE indicator (MZW-MAE), multi-zone weighted MAPE indictor (MZW-MAE), multi-zone weighted RMSE indictor (MZW-RMSE) are adopted to evaluate the overall prediction performance of these prediction models in all the zones. 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).



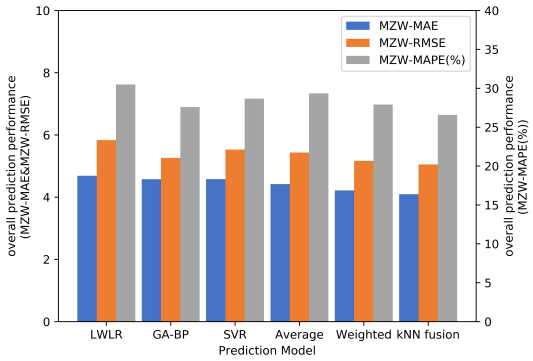
(a) MAE analysis



(b) MAPE analysis



(c) RMSE analysis



(d) Overall prediction performance

Figure 6: Prediction accuracy analysis

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, 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.

6 CONCLUSION

In order to help facilitate city-scale travel operation management in megacity, this project 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.

7 CONTRIBUTIONS

We are three member team, Chenyu worked on the data extraction part, prepared all the features with necessary plots; Pengshun worked on the algorithm implementation with the sklearn framework; Chenxi worked on the overall methodology design on the benchmark on this task and paper write up. All the members joined in the results analysis and future work discussion and write-up collaborations.

REFERENCES

[1] Yves Chauvin and David E Rumelhart. 2013. *Backpropagation: theory, architectures, and applications.* Psychology Press.

[2] William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 74, 368 (1979), 829–836.

[3] Fangce Guo, Rajesh Krishnan, and John 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 (2017), 214–226.

[4] Fangce Guo, John W Polak, and R Krishnamoorthy. 2018. Predictor fusion for short-term traffic forecasting. (2018).

[5] Filmon G Habtemichael and Mecit Cetin. 2016. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation research Part C: emerging technologies* 66 (2016), 61–78.

[6] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen. 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies* 85 (2017), 591–608.

[7] Xiaolong Li, Gang Pan, Zhaohui Wu, Guande Qi, Shijian Li, Daqing Zhang, Wangsheng Zhang, and Zonghui Wang. 2012. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science* 6, 1 (2012), 111–121.

[8] Ali Yadavar Nikravesh, Samuel A Ajila, Chung-Horng Lung, and Wayne Ding. 2016. Mobile network traffic prediction using MLP, MLPWD, and SVM. In *2016 IEEE International Congress on Big Data (BigData Congress).* IEEE, 402–409.

[9] Han Qiu, Ruimin Li, and Hao Liu. 2016. Integrated model for traffic flow forecasting under rainy conditions. *Journal of Advanced Transportation* 50, 8 (2016), 1754–1769.

[10] Vladimir Vapnik. 2013. *The nature of statistical learning theory.* Springer science & business media.

[11] Eleni I Vlahogianni. 2015. Optimization of traffic forecasting: Intelligent surrogate modeling. *Transportation Research Part C: Emerging Technologies* 55 (2015), 14–23.