Acta Technica *62* (2017), No. 5B, 233–244

An new approach to predict the travel time through ETC and analysis of their influence factors

Shuli Wang[1,3], Meifang Wu[1,4], Chihchia Hsu[2], Fan Wu[2]

Abstract. Since the era of autonomous vehicles is coming, the precise prediction of travel time in a route is more important. The prediction results can help in decision making, such as route selection and perception of accidents in the route. Furthermore, since of the introduction of IOT sensors, like ETC, the data collection for the vehicles in the routes is easier and faster than ever. This paper proposed a method, which used several training model to predict the travel time for routes. Compared to the previous studies, the paper proposed Double-Factor approach, which separates the two features, current traffic status and weather condition, into another training model. The extra training model is used to tune the predicting results derived from the traditional models. A lot of experiments are performed to show its effectiveness. In addition, the paper further analyzed the influence of each feature with precision measures, and identified which features and when are helpful to the prediction.

Key words. Travel time prediction, ETC, IOT, autonomous vehicle.

1. Introduction

Recently travel time prediction has been an important issue in intelligent transportation system, especially in handling congestion problems. According to 2015 urban mobility scorecard [1], in 2014 the congestions in America had caused $160 billion congestion cost and wasted Americans 6.9 billion hours since of the congestions. To reduce the congestion cost and the waste of time, travel time prediction is needed. With that, travelers can make smart decisions about when to travel and on

[1] Workshop 1 - Department of Industry Engineering and Systems Management, Feng-Chia University, Taichung, Taiwan

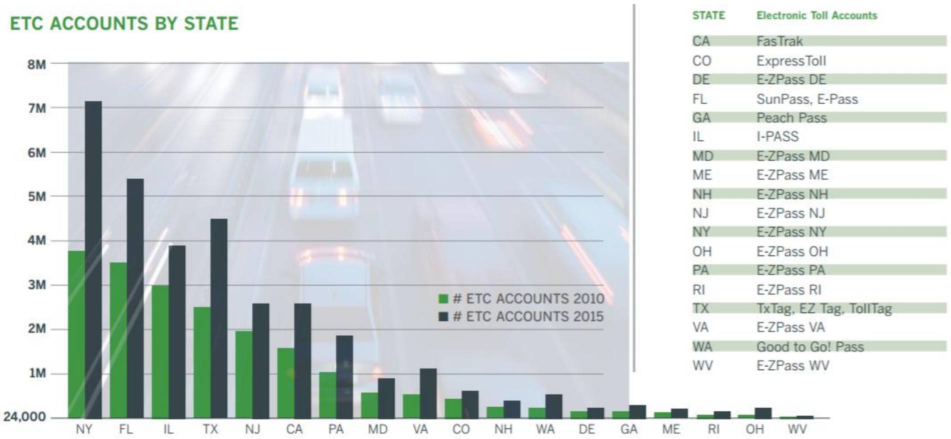
[2] Workshop 2 - Department of Management of Information System, National Chung-Cheng University, Chia-Yi, Taiwan

[3] Workshop 3 - Department of Dental Technology and Materials Science, Central Taiwan University of Science and Technology, Taichung, Taiwan

[4] Corresponding author: Meifang Wu; e-mail: `slwang20170101@gmail.com`

what routes to travel if the prediction is accurate.

In the past, it is hard to make travel time prediction since the traffic data is hard to collect for prediction. Fortunately, the advance of Internet of Things (IoT) makes it easier to collect traffic data in recent years. For example, the electronic toll collection (ETC) system is set up commonly in highway traffic system in many countries, and its popularity is still growing. ETC can collect numerous traffic data for the tolls and traffic control. From 2010 to 2015 in US, the electronic transponders on America's roads increased 19.3M (million), while the ETC accounts increased from 19.9M to 32.7M [2]. Figure 1 shows the increase of toll accounts and transponders in United States, compared to the statistical records in the two years in 2010 and 2015.



STATE	Electronic Toll Accounts
CA	FasTrak
CO	ExpressToll
DE	E-ZPass DE
FL	SunPass, E-Pass
GA	Peach Pass
IL	I-PASS
MD	E-ZPass MD
ME	E-ZPass ME
NH	E-ZPass NH
NJ	E-ZPass NJ
NY	E-ZPass NY
OH	E-ZPass OH
PA	E-ZPass PA
RI	E-ZPass RI
TX	TxTag, EZ Tag, TollTag
VA	E-ZPass VA
WA	Good to Go! Pass
WV	E-ZPass WV

Fig. 1. ETC accounts by state in U.S between 2010 and 2015

Since the traffic data is collected much easier than ever, many studies had addressed travel time prediction in the last decade. Travel time prediction aims at measuring future travel time for the same trip. For example, suppose node A is as a starting point and node B is as a destination. In general, people consider historical data in a period (e.g. a month away from the prediction time) of traffic data of those vehicles who took a trip from nodes A to B. Figure 2 shows the data collection for the vehicles in a specific route from day 1 to day $(m + k)$. Assume we need to predict the traffic time for the route in time slot t as well as in latter time slots in day m. According to rule of thumb, the travel time prediction for a route in time slot m only related to its prior travel time in the same route; the posterior traffic condition, like congestion, and accidents, is unrelated to the prediction. All the literature for the travel time prediction adopts the assumption to train their prediction model. In detail, the traffic date from day 1 to day m are the historical data, and in each of those days the time slots from (t-n) to t are the data used to train the model.

Wu, Ho, and Lee [3] first utilized Support Vector Regression (SVR) model in travel time prediction. Their result shows SVR model outperformed the traditional statistical method. Innamaa [4], Oh and Park [5], and Li and Chen [6] all performed the similar experiment with neural network model to predict travel time, the results

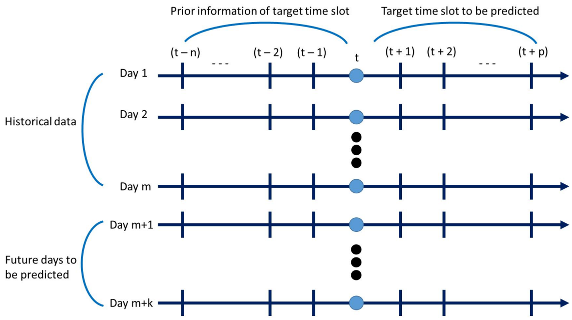

Prior information of target time slot
Target time slot to be predicted
(t – n)
(t – 2)
(t – 1)
t
(t + 1)
(t + 2)
(t + p)
Day 1
Day 2
Day m
Day m+1
Day m+k
Historical data
Future days to be predicted

Fig. 2. The traffic data for the traffic time prediction

being more precise and the execution time shorter. However, the neural network model is regarded as a black box; no influence factors or rules are exposed. Qiao, Haghani, and Hamedi [7] considered weather factors and use K nearest neighbor model to predict the travel time; obviously, it is reasonable to adopt the weather factor in the prediction, in addition to the data collected from ETC system. Recently, Zhang & Haghani [8] and Li & Bai [9] use Gradient Boosting Regression Tree (GBRT) for predicting travel time. Unlike the black-box method, like neural network, GBRT have the explanation ability to show the importance and its effects each feature has.

These literatures discussed how to use different methods to predict the travel time. But the traffic system is complex and the traffic flow may be affected by many factors, like the bad weather, holidays, and accidents. Travel time prediction model should incorporate these factors in order to provide accurate predictions [12]. In this paper, we designed a double-features approach (called DF-approach hereafter), which considers two main features, namely, current traffic and current weather features. Compared to the traditional ones, DF approach adopted four models, namely, SVR, neural-network, GBRT, and K-nearest-neighbor, in a time. Since of the limitation of each model and their precisions being varied for different data distribution, DF approach trains four models with the historical data, and finally selects the best one that has the highest precision. This paper adopted two precision criterion, namely, root mean squared error (RMSE) and mean absolute percentage error (MAPE), which have been widely used for evaluating the precision of the prediction model [6, 9, 10]. To demonstrate the effectiveness of the proposed approach, we conducted many experiments with KDDCup 2017's traffic-flow prediction dataset. The results show that the proposed method outperforms all the other methods in most conditions.

2. Material and Method

2.1. Collected traffic data set and its preprocessing

Since of the advance of internet of things, the data collection is easier than ever. In most countries, they incorporate ETC to reduce the workload of tollgates. The incorporation of ETC can also provide the information about the route of a vehicle along with its entry to the highway, the intersection and its exit from the highway. Figure 3 shows the road network topology in a target area. Figure 3(a) is a bird-eyes view for an area, where we can see the in-flow and out-flow of the vehicles from their entries from intersections, their travelling routes across tollgates, and their exit from intersections. In fact, the ETC is tracing the vehicle in links (or segments), each of which is around 100 meters. Figure 3(b) is the route of a vehicle, which is composed by several links along the route from intersection A to tollgate 1.

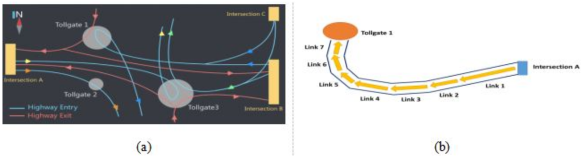
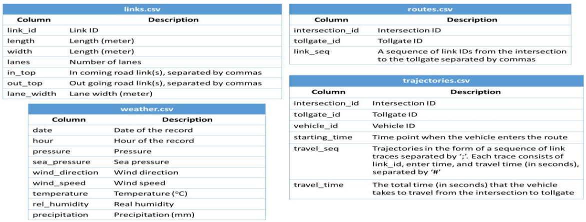


Fig. 3. The road network topology in an area. The bird-eyes view for the area is shown in Figure 3(a); the links of a route traced by ETC is shown in Figure 3(b).

Since ETC sensors are installed in a fixed location along the high way, the collected data is from the dataset of the sensors without route consideration. The route connection will not show is the collection file from ETC. The paper used KDDCup 2017's traffic-flow prediction dataset (https://tianchi.aliyun.com/competition/ information.htm?raceId=231597) in an area in China, containing the information of the route, passing time, and weather along the highway. The dataset are real data, consisted of several files, only four of which are useful for travel time prediction. The other files are used for traffic volume prediction, which is beyond the scope of this paper. Figure 4 shows the description of these four files; in detail, the first file, links.csv, describes the profile of each link in the target area, including its ID, link's width, length, and lanes, and its in-coming links and out-going links. Note that a link is one-directional; there may be more than one links (i.e., in-coming links) flow into that link, and there may be more than one links (i.e., out-going links) flow away that link. The second file, routes.csv, describes a route from an intersection to a tollgate, and all the links along the route. Since a link is around 100 meters away from each other, a route is normally consisted of several links, and their order is arranged according to their connected order in the route. The third file, weather.csv, contains a travel time for some specific vehicle in a route from an intersection to a tollgate. Each record of the file includes the start and end points of a route, the links (or segments) of a rout, the vehicle ID traveling the route, the starting time for the travelling and total time spent on the route for the vehicle. Note that the vehicle ID is masked but represented as another symbols for privacy consideration.

In general, the congestion occurs in the same time slot of each day. For example,



links.csv	
Column	Description
link_id	Link ID
length	Length (meter)
width	Length (meter)
lanes	Number of lanes
in_top	In coming road link(s), separated by commas
out_top	Out going road link(s), separated by commas
lane_width	Lane width (meter)

weather.csv	
Column	Description
date	Date of the record
hour	Hour of the record
pressure	Pressure
sea_pressure	Sea pressure
wind_direction	Wind direction
wind_speed	Wind speed
temperature	Temperature (°C)
rel_humidity	Real humidity
precipitation	Precipitation (mm)

routes.csv	
Column	Description
intersection_id	Intersection ID
tollgate_id	Tollgate ID
link_seq	A sequence of link IDs from the intersection to the tollgate separated by commas

trajectories.csv	
Column	Description
intersection_id	Intersection ID
tollgate_id	Tollgate ID
vehicle_id	Vehicle ID
starting_time	Time point when the vehicle enters the route
travel_seq	Trajectories in the form of a sequence of link traces separated by ';'. Each trace consists of link_id, enter time, and travel time (in seconds), separated by '#'
travel_time	The total time (in seconds) that the vehicle takes to travel from the intersection to tollgate

Fig. 4. The data for the travel time prediction from KDDCup 2017.

the congestion often appears in the rush hours, AM 0700-0900 and PM0500-0700, every work day if no accident occurs. Without loss of generality, the time slot is set as one hours in the paper. Before training the model for prediction, we transform the raw data in the above four files into the needed format. The needed data, called travel time features, includes the average travel time, traffic volume, day of week, rainfall, and so on, which are listed in Table 1.

Table 1. The travel time features extracted from the four files from KDDCup 2017.

Feature	Value	Description
Average travel time	Float	Average travel time, measured in seconds
Traffic volume	Integer	Traffic volume
Day of week	0∼6	0: Monday, 1: Tuesday, . . . , 6: Sunday
Workday	0,1	0: not workday, 1: workday
Public holiday	0,1	0: not public holiday, 1: public holiday
Rainfall	Float	Amount of rainfall, measured in millimeter

2.2. The proposed method

This paper used four steps to predict travel time, namely, data collection, data preprocessing, model selection, and DF approach, and model testing. Fig. 5 illustrates these four steps. Except the fourth step, i.e., using DF approach, the other steps are similar to the traditional ones.

In the first step, the paper adopted KDDCup 2017's traffic-flow dataset and then normalized it. In the second step, the paper filtered the missing data and extracted features from the original dataset. In the next step, the paper used the cross validation method to select base model from four models, SVR model, KNN model, NN model, and GBRT model, according to accuracy results of these candidate models. As mentioned before, there are several different models used to predict travel time, each of which has its own edges for different features and in different data distribution. It is hard to artificially determine which model is best to use. Therefore, the paper uses the cross validation method [11] to select base model according to accuracy results of these four candidate models. Fig. 6 illustrates how we train the model and select the appropriate model. First, we split the processed data derived from the second step into two subsets, i.e., the training set and the validation set. Secondly, the paper uses the training set to train four candidate models and then evaluate their individual accuracy against the validation set. Finally, the model with highest accuracy is chosen as the base model.

In the last step, the paper combine DF approach with the base model to enhance the accuracy of travel time prediction. The paper considers the current traffic status and rainfall as dominant factors and processes these two factors in another training model. The reason for the current traffic status is that if there is an accident in front of a vehicle, the delay time is hard to estimate. The traffic system will return to regular status when the accident situation is excluded. Obviously, the current traffic status should not be considered as the training factor in the training process as the previous ones. On the contrary, DF approach used an extra training model, posterior to the selection of the base model, for the current traffic status to tune the prediction travel time. The reason for considering the rainfall as an independent factor is that according to rule of thumb, the amount of rainfall significantly influences the vision of a driver. The driver will slow down the vehicle in instinct to avoid from the car crash, no matter whether there is congestion in front of the driver. However, the rainfall is a nature phenomenon, not a regular condition. DF approach used an extra training model, also posterior to the selection of the base model, for the rainfall condition to tune the prediction travel time. Compared to the traditional ones, they all considered the above two factors as normal features, without considering them separately. The paper considers the two factors as independent factors and expects the prediction of travel time being more accurate than that of others.

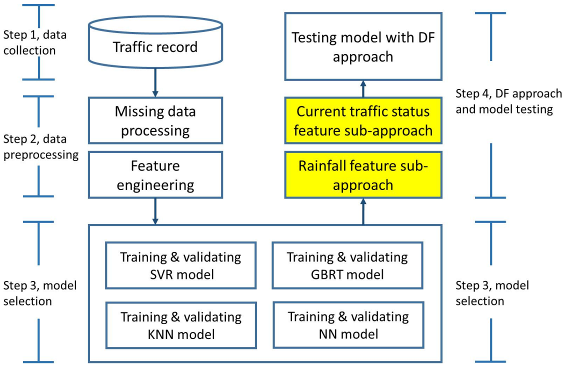

Step 1, data collection
Traffic record
Step 2, data preprocessing
Missing data processing
Feature engineering
Step 3, model selection
Training & validating SVR model
Training & validating GBRT model
Training & validating KNN model
Training & validating NN model
Step 3, model selection
Rainfall feature sub-approach
Current traffic status feature sub-approach
Testing model with DF approach
Step 4, DF approach and model testing

Fig. 5. Four steps to predict the travel time.

3. Experiments and Discussion

To demonstrate the advantages of the proposed method, we perform a lot of experiments. The experiments are executed in a PC-based platform with i7 CPU and 1G Ram. The data comes from KDDCup 2017's traffic-flow prediction dataset, which contains six routes, namely, route A$\sim$2, A$\sim$3, B$\sim$1, B$\sim$3, C$\sim$1, and C$\sim$3, in an area in China, where the naming rule for a route is the concatenation of intersection

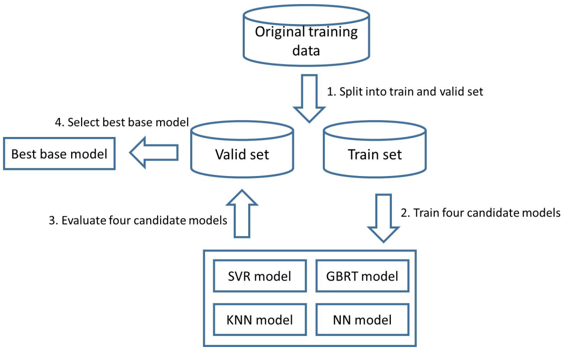
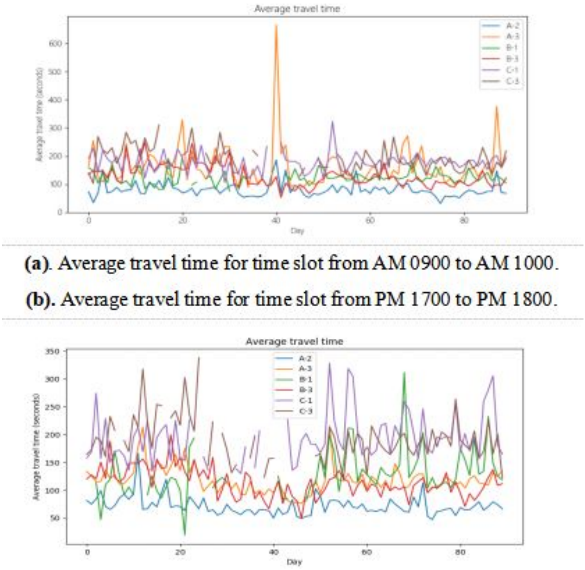

Original training data
1. Split into train and valid set
4. Select best base model
Best base model
Valid set
Train set
3. Evaluate four candidate models
2. Train four candidate models
SVR model
GBRT model
KNN model
NN model

Fig. 6. Model training, testing and selection

ID, "∼", and then tollgate ID. For example, for two routes' notation, A∼2 and A∼3, they mean that the two routes have the same starting point, i.e., intersection A, but have different destination points, i.e., tollgates 2 and 3, respectively. In addition, the dataset also includes the time records of thousands of vehicles travelling on the six routes as well as the weather information in the area from the date 2016-07-19 to the date 2016-10-24.

The experiments selects some arbitrary time slot in some arbitrary date to predict the traveling time for the time slot of all the six routes from the starting point to the end point. Note that the selected date to predict should be located in the latter half of the dataset to avoid from the insufficient training data. In the prediction, the proposed method used the traveling data in the nearest seven days as the validation data set, while the data in the dates before these seven days is used as the training data set. Moreover, we choose two time slots, namely, AM 0900 to AM 1000 and PM 1700 to PM 1800, as the target time slot to predict. Suppose we choose the date, 2016-10-24, to predict its travelling time in the above two time slots. According to the above discussion, the data from the date 2016-07-19 to the date 2016-10-17 is used as the training data set, while the data from seven days from the date 2016-10-18 to the date 2016-10-24 is used as the validation data set. Figure 7 shows the average travel time of the above six routes in the training data set for the two chosen time slots. From the figure, we can see some line (representing the travel time) being fragmented since there is missing data in that day. For these cases, we skip the data of that date for keeping the consistence of the training data set.

To evaluate the models, the paper implemented the four training models. We use the open-source machine learning library 'Scikit-learn' to implement SVR, KNN, and NN models. For GBRT model, we use a common GBRT framework named lightGBM to implement it. In addition, the paper used two most popular evaluation metrics, namely, root mean squared error (RMSE) and mean absolute percentage error (MAPE), to evaluate the prediction performance, as being shown in Equations 1 and 2, respectively. In these equations, $\hat{y}_t$ represents the ground truth value (i.e., the actual travel time), and y_t represents the predict value (the predict travel time). Without loss of generality, we calculate the average travel time of every 5 minutes to be the ground truth value, which is a regular time range (Myung et al, 2011; and



(a). Average travel time for time slot from AM 0900 to AM 1000.

(b). Average travel time for time slot from PM 1700 to PM 1800.

Fig. 7. The average travel time from the date 2016-07-19 to the date 2016-10-17, where 7(a) and 7(b) are for the average travel time for time slots AM 0900 to AM 1000, and PM 1700 to PM 1800, respectively

Fei, Lu, & Liu, 2011).

In subsection 2.1, the paper processed and extracted the raw data, and formed into a preprocessed data consisted of six features, namely, average travel time, traffic volume, day of week, workday, public holiday, rainfall, and minute slot, as being shown in Table 1. However, we are not sure whether and how much these features are important. Hence, the paper performed several experiments to compare the accuracy with or without the adoption of some feature into the training data. Table 2 shows the scenario of feature selections for the training data set. Note that SVR model is chosen as the baseline model since of its highest accuracy after experiments. We then performed another set of experiments to determine which feature set should be adopted in predicting the travel time.

Table 2. The scenario of feature selection in the training data set.

feature selection	Experiment no.	Description
All features	1.1	All features adopted.
∼Average travel time	1.2	All features exclusive average travel time adopted.
∼Traffic volume	1.3	All features exclusive traffic volumes adopted
∼weekday	1.4	All features exclusive weekday adopted.
∼Workday	1.5	All features exclusive workday adopted
∼Public holiday	1.6	All features exclusive public holiday adopted
∼Rainfall	1.7	All features exclusive rainfall adopted
∼Minute slot	1.8	All features exclusive time window adopted

Figures 8 and 9 show the experiment results about the predicting and actual travel time of each route for the two time slots from AM 0900 to AM 1000 and from PM1700 to PM1800 in the duration from the date 2016-10-18 to the date 2016-10-24. In the two figures, the actual travel time is drawn in blue, and the predicting travel time under different feature sets is draw in other colors. We can see that the actual travel time is in zigzag line. We guess that he traffic condition varied since of many unknown status, such as rush hour, holiday, rainfall, etc. We can also observe that if the feature of average travel time is exclusive (i.e., experiment no. is No. 1.2), its predicting travel time is worse than that of all other experiments. It is reasonable that because current travel time is mainly correlated to the historical average travel time rather than other features, like holiday, rainfall, etc. On the contrary, we observe that the results of the experiment exclusive minute slot have the highest accuracy. The minute slot is to split the granularity of time slot from an hour to 5 minutes. We expect the predicting results should have higher precision. The results match our expectation that the smaller the granularly of the time slot, the higher precision the predicting results. Note that the above statement is valid under the condition that the training data set should be sufficient for each smaller granularity of time slots.

Tables 3 and 4 show the precision comparison in terms of RMSE and MAPE for the experiments from No. 1.1 to No. 1.8 in the two target time slots. According to the properties of RMSE and MAPE, the smaller the RMSE and MAPE are, the higher precision the models have. In the two tables, all the experiments under the consideration of different feature sets (i.e., experiments from No. 1.2 to No. 1.8) are compared to the experiment (called base experiment) under the consideration of all the feature sets (i.e., experiment No. 1.1). The cells in green mean their precision measures of the corresponding feature sets are worse than that of the base experiment, while the cells in red means their precision measures of the corresponding feature sets are better than that of the base experiment. We can see that when we excluded average travel time (experiment No. 1.2), traffic volume (experiment No. 1.3), rainfall (experiment No. 1.7), and minute slot (experiment No. 1.7), their results get worse. The results show that these four features are helpful in predicting

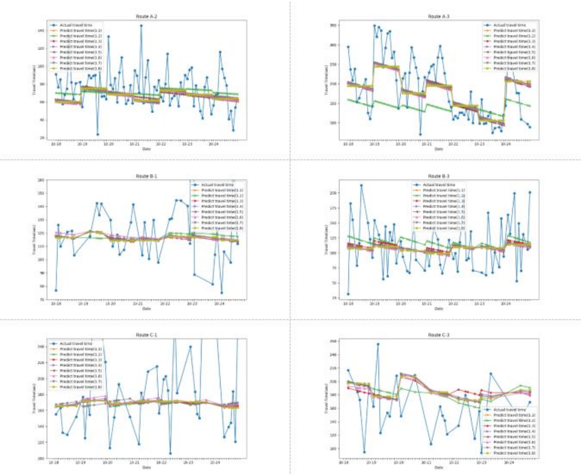


Fig. 8. The predicting and actual travel times for the time slot from AM 0900 to AM 1000 for each route.

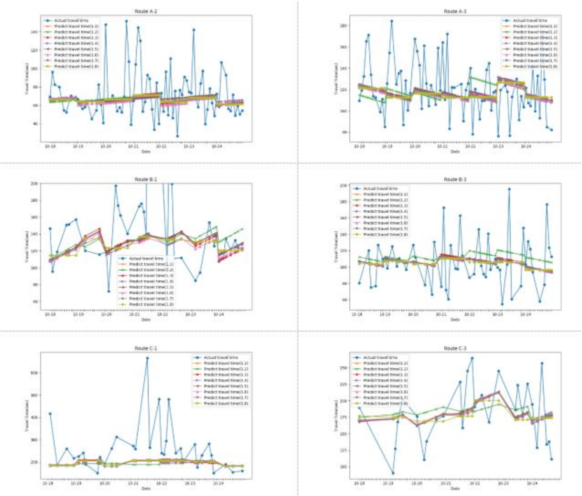


Fig. 9. The predicting and actual travel times for the time slot from PM 1700 to PM 1800 for each route

travel time. In addition, when we excluded workday and public holiday, their MAPE score decrease in both of the target time slots, but their RMSE score increase in the time slot of PM 1700 to PM 1800. The results represent workday and public holiday features are not helpful in predicting travel time for time slot of AM 0900 to AM 1000. That is, the workday and public holiday features has little influence on the travelling time in the time slot of AM 0900 to AM 1000, but has influence in the time slot of AM 0900 to AM 1000. These results can be explained the phenomena that in the work day and public holiday, the frequency that people drive their cars in the routes are similar in the time slot of AM 0900 to AM 1000, regardless of that the day is work day or holiday. But the frequency that people drive their cars in the routes are different significantly in the time slot of PM 1700 to PM 1800 (the rush hour), according to the day being work day or public holiday.

Table 3. Precision evaluation for experiments no. from 1.1 to 1.8 for the time slot AM 0900 to AM 1000

Exp. No Metrics	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
mean RMSE (min)	0.8114	0.8908	0.8121	0.8100	0.8074	0.8078	0.8132	0.818
mean MAPE(%)	25.076	26.538	25.179	25.091	24.755	24.923	25.156	25.252

Table 4. Precision evaluation for experiments no. from 1.1 to 1.8 for the time slot AM 1700 to AM 1800

Exp. No Metrics	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
mean RMSE (min)	0.9031	0.9212	0.9044	0.9040	0.9038	0.9079	0.9101	0.9093
mean MAPE(%)	22.073	23.19	22.218	22.101	22.021	21.878	22.155	22.100

4. Conclusion

Travel time prediction is worthy of study and is beneficial to society, especially in the coming era of autonomous vehicle. Since of the introduction of ETC, the data collection for the vehicles in the routes is easier and faster than ever. The travelling data can be used to provide decision making, such as route selection and perception of accidents in the route. This paper proposed a method, which used several training model to predict the travel time for the routes. Compared to the previous studies, the paper separated the two features, current traffic status and weather condition, into another training model. The extra training model is used to tune the predicting results derived from the traditional models. In addition, the paper further analyzed the influence of each feature with precision measures, and identified which features and when are helpful to the prediction. Since of the coming of big data, the similar analysis combing data from different areas is popular. The future work is to incorporate more data, such as the air pollution data and the stability of electronic power supply, etc., to see their influence of the travelling time.

References

[1] L. Martino, J. Read, V. Elvira, F. Louzada: *Cooperative parallel particle filters for online model selection and applications to urban mobility.* Digital Signal Processing *60* (2017), 172–185.

[2] RS. Bowman, S. Taylor, A. Zupan: *A Zupan. Bridge Spectra of Twisted Torus Knots.* International Mathematics Research Notices *495* (2015), No. 16, 74–80.

[3] Wu, C. H. Ho, J. M, Lee, D. T: *Travel-time prediction with support vector regression.* IEEE transactions on intelligent transportation systems *5* (2004), No. 4, 276–281.

[4] Innamaa. S: *Short-term prediction of travel time using neural networks on an interurban highway.* Transportation *32* (2005), No. 6, 649–669.

[5] Oh. C, Park. S: *Investigating the effects of daily travel time patterns on short-term prediction.* KSCE Journal of Civil Engineering *15* (2011), No. 7, 1263–1272.

[6] Li. C. S, Chen. M. C: *Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks.*Neural Computing and Applications *23* (2013) , No. 6, 1611–1629.

[7] Qiao. W, Haghani. A, Hamedi. M: *Short-term travel time prediction considering the effects of weather.* Transportation Research Record: Journal of the Transportation Research Board *2308* (2012),61–72.

[8] Zhang. Y,Haghani. A: *A gradient boosting method to improve travel time prediction.* Transportation Research Part C: Emerging Technologies *58* (2015),308–324.

[9] Li. X, Bai. R: *Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree.*In IEEE International Conference on Machine Learning and Applications (2016), 1010–1015.

[10] Myung. J, Kim. D.. K., Kho. S. Y, Park. C. H: *Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system.* Transportation Research Record: Journal of the Transportation Research Board *20*, (2011), No. 2256, 51–59.

[11] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel:Journal of Machine Learning Research *12* (2012), No. 10, 2825–2830.

[12] Vlahogianni. E. I, Karlaftis. M. G, Golias, J. C: *UShort-term traffic forecasting: Where we are and where we're going.*Transportation Research Part C: Emerging Technologies *43* (2014), 3–19.

Received November 16, 2017