Neurocomputing 133 (2014) 74–83

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A data mining based approach for travel time prediction in freeway with non-recurrent congestion

Chi-Sen Li, Mu-Chen Chen *

Department of Transportation and Logistics Management, National Chiao Tung University, 4F, No. 118, Section 1, Chung-Hsiao W. Road, Taipei 100, Taiwan, ROC

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form
13 November 2013
Accepted 16 November 2013
Communicated by Manoj Kumar Tiwari
Available online 9 January 2014

Keywords:
Travel time prediction
Non-recurrent congestion
K-means
Classification and regression tree
Neural networks

ABSTRACT

This study integrates three data mining techniques, *K*-means clustering, decision trees, and neural networks, to predict the travel time of freeway with non-recurrent congestion. By creating dummy variables and identifying important variables, not only is the prediction performance increased without increasing investment in equipment, but also important variables are obtained concerning the important locations of equipment in order to effectively assist public transit agencies with system maintenance. The experimental results for a segment of 36.1 km of National Freeway No. 1, Taiwan, with non-recurrent congestion show that, whether or not the data generated by the Electronic Toll Collection (etc) system is used as input variables, the travel time prediction method developed in this study is able to improve the prediction performance. Meanwhile, the proposed approach also reduces the percentage of samples with mean absolute percentage error (MAPE) > 20%. Furthermore, in this study, important variables are extracted from the decision tree in order to predict the travel time. Finally, the prediction models constructed in accordance with six scenarios are highly accurate due to the low MAPE values, which are from 6% to 9%.

1. Introduction

According to a report by the International Energy Agency in 2011 [1], the transport industry made the second largest global contribution to CO_2 emissions, accounting for 23%, following the electricity generation industry, which accounted for 41%. Roads are the main source of CO_2 emissions in the transport industry. Joumard et al. [2] found that CO_2 emissions from vehicles traveling at low speed in urban areas are higher than those from vehicles traveling at high speed. Furthermore, cold engines consume more fuel and generate more pollution than engines that are fully warmed up. Therefore, increasing or maintaining smooth traffic flows for reducing the conditions of stop and go while drivers travel on roads not only reduces the social costs, but also makes an important contribution to reduce greenhouse gas emissions. However, the speed of construction of additional roads generally cannot match the increase in the number of vehicles. Thus, construction of additional roads may be difficult to efficiently ease traffic congestion. In view of this, intelligent transportation systems (ITSs) provide a viable solution that can improve the efficiency and service standard of the existing transportation system and relieve or resolve the road congestion problem. Therefore, in recent decades, ITS has become a mainstream research area. Advanced traveler information systems (ATIS) and advanced traffic management systems (ATMS) are technologies that are often used in ITS to improve the efficiency of the road system. Many studies (e.g., [3,4]) have also pointed out that providing travel time information is an important factor in encouraging the success of ATIS and ATMS. Therefore, in order to relieve traffic congestion or reduce CO_2 emissions, travel time prediction is an important issue. According to the Oak Ridge National Laboratory [5], 55% of the delays encountered by drivers on American freeways are caused by non-recurrent events, of which 72% are freeway accidents [6]. Therefore, in recent years, in order to improve the applicable timing of travel time prediction models, related studies (e.g., [7,8]) have extended the research from exploring general traffic flow conditions to how to improve the prediction performance in the case of non-recurrent congestion.

From the study by Golob et al. [9], there exists a close relationship between traffic flow conditions and traffic accidents (crashes), by type of crash. For example, the congestion flow is apt to result in more serious crashes. In recent years, ATIS and ATMS have been widely utilized to ensure the efficiency of road system and to avoid the congestion flow. From the study by Vanderschuren [10], intelligent transport systems (ITS) can reduce (potential) crashes, and this finding is also demonstrated in Mitretek Systems [11]. Mitretek Systems reported that there are six major objectives/benefits of ITS identified in the literature, which consist of safety, mobility, efficiency, productivity, energy/environment and customer satisfaction.

* Corresponding author.
E-mail address: ittchen@mail.nctu.edu.tw (M.-C. Chen).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.029

Therefore, forecasting travel time, particularly in the freeway with non-recurrent congestion, with a high degree of accuracy, can improve safety and efficiency.

Many studies related to prediction of freeway travel time have shown concrete results. Kwon et al. [12] used a liner model to collect loop detector data in order to predict travel time. Oda [13] and Van Arem et al. [14] used autoregressive integrated moving average (ARIMA) models to obtain appropriate results of travel time prediction in general traffic conditions. Van Arem et al. [14] found that the ARIMA model is applicable to normal traffic congestion, but shows larger deviations in non-recurring congestion. There are also many studies that predict travel time through Kalman filtering [15–20]. Fei et al. [7] developed the Bayesian inference-based dynamic linear model (DLM) and further constructed the adaptive dynamic linear model (ADLM) to predict the travel time in two traffic conditions of recurrent and non-recurrent congestion. Fei et al. [7] divided the data into two data sets, morning and afternoon, and performed travel time prediction with both data sets. The prediction performance of the proposed ADLM is between highly accurate prediction and good prediction.

Moreover, in order to construct the prediction model of actual travel time, previous studies mainly obtain the actual travel time information through probe vehicles [21] and automatic vehicle identification (AVI) [16,22–25]. Petty et al. [21] developed a speed calibrated model through single-loop loop detectors, and then estimated the travel time accordingly. The study by Petty et al. [21] confirmed that the developed stochastic model is applicable in congestion conditions, but not including non-recurrent congestion. Meanwhile, Petty et al. [21] used the travel time data collected by four probe vehicles as the model validation target. Chien and Kuchipudi [16] found that the percentage of probe vehicles in the total traffic flow is a critical factor, which affects the prediction accuracy when using AVI data to predict the travel time. In recent years, neural networks (NNs) [22,26–32] have been utilized to predict traffics with various degrees of success due to their modeling flexibility, predictive ability, and generalization potential [33]. The above-mentioned studies also confirmed that NNs have good prediction capability when applied to the analysis of complex traffic characteristics.

Although previous studies have predicted the short-term travel time with various degrees of success, scientific research with regard to travel time reliability focuses on four fundamental issues as follows:

1. Enhancing the prediction capability with existing equipment.
2. Identifying the important detectors and the critical variables in order to enable the authority to obtain the target detectors and develop the effective imputation methods for missing data.
3. Providing the robust and accurate prediction model which is also applicable in the case of non-recurrent congestion.
4. Classifying different categories of traffic characteristics in order to predict the travel time.

Previous studies (e.g., [16,17,24,25]) mainly focused on the first issue mentioned above, but there are few studies on the third issue. However, for the transportation management unit, there is a trade-off between costs and output benefits. How to improve the prediction performance and applicable timing under the premise of reducing the system implementation cost and operational cost is an important goal of travel time prediction. Therefore, the objective of this study is to use a systematic structure to integrate a variety of data mining methods in order to develop an effective solution to the first three issues mentioned above. For the first issue, there is no doubt that there is a significant positive relationship between improving the explanatory power of traffic characteristics and enhancing the performance of travel time prediction. The previous travel time studies (e.g., [15–17,24,25]) were mainly focused on enhancing the explanatory power of traffic characteristics irrespective of whether the data were derived through model inference, system simulation, parameter calibration, or nonlinear model construction. Therefore, identifying the traffic characteristics at time t could help with improving the accuracy of travel time prediction. Furthermore, collecting more comprehensive traffic data could reflect the traffic characteristics better. For example, stronger explanatory power of traffic characteristics could be obtained when a vehicle detector is set up every kilometer than when one is set up every 10 km. However, more devices will result in higher equipment and maintenance costs. Therefore, increasing devices and enhancing the explanatory power of traffic characteristics have been a trade-off for long time. In order to solve this problem, this study proposes a solution which attempts to avoid the cost of increasing devices and improve the explanatory power of traffic characteristics at the same time. Furthermore, the performance of travel time prediction is expected to be enhanced. In this study, to achieve the above goals, we use *K-means* to categorize the traffic characteristics in every 5 min and create a dummy variable to mark the cluster ID of each 5-min sample.

With regard to the second issue, when compared with collecting data manually, there is no denying that collecting data through a device is better for long-term data collection and makes it easier to control the error. However, it cannot prevent the occurrence of missing data. Therefore, researchers [34–38] have studied the processing of missing data in order to understand the exact traffic characteristics. However, in terms of freeway travel time prediction, relatively few previous studies focused on identifying important device locations and critical variables. For example, if rainfall at detection point A, the spot speed at detection point B, heavy vehicle volume at detection point C, etc. are identified as the critical variables for analyzing traffic characteristics and predicting travel time, managers not only can clearly understand the target detectors but also can develop an imputation method for missing data for a specific variable collected at a particular detection point. In addition, this can also have the benefits of reducing investment in equipment, decreasing maintenance cost and enhancing the applications of model. In this regard, in this study, we use the classification and regression tree (CART) to provide an effective solution to identify important device locations and critical variables for travel time prediction.

For the third issue, in order to construct a robust model which is able to predict freeway travel time in traffic with non-recurrent congestion, a method integrating *K-means*, CART and NN is proposed in this study. Furthermore, in this study, the raw data of the ETC system (vehicle charging time and ID) are used to calculate the actual travel time (ATT), which is used as a target in training the prediction model. This study develops a robust model for predicting the travel time in the case of non-recurrent congestion. With this model, the prediction performance can be improved with existing equipment, and critical variables at important locations can be identified such that the management unit can have a clear objective in operating and maintaining equipment.

The rest of this paper is organized as follows. Section 2 presents the details of the freeway segment in this study and data collection. Section 3 describes the method of constructing the model of travel time prediction. The experimental design and results are presented in Section 4. Finally, conclusions and suggestions of this study are made in Section 5.

2. Data

2.1. The freeway segment in this study

Two freeways, National Freeways No. 1 and No. 3, form the main inter-city transportation corridor between the south and the

north of Taiwan. According to the data reported by Taiwan Area National Freeway Bureau (MOTC) in 2009, the total annual traffic volume of these two freeways was 539,568,273 vehicles, while the annual total traffic volume of National Freeway No. 1 was 329,743,228 vehicles, which accounts for approximately 61.11%. Therefore, National Freeway No. 1 is the main freeway for intercity transportation in Taiwan. In this study, the data were collected at 5-minute intervals between the Yangmei Toll Station and Taishan Toll Station of the freeway in the northward direction from 16 September to 16 October 2009. The studied segment is a one-way three-lane freeway. From Fig. 1, there are eight interchanges including two system interchanges, the Pingzhen System Interchange and the Airport System Interchange in this segment. Its length is 36.1 km accounting for 9.7% of National Freeway No. 1 with a service population of 9,382,332 people accounting for 40.58% of the total population of Taiwan. Furthermore, according to the statistics of September and October 2009 reported by Taiwan Area National Freeway Bureau MOTC, the average daily traffic volume in this freeway segment was 336,895 vehicles, which is the sum of the daily traffic volumes at Yangmei Toll Station and Taishan Toll Station. It accounts for 23.5% of the average daily traffic volume of National Freeway No. 1. With the above information, the segment in this study is the busiest and most complicated segment in National Freeway No. 1.

2.2. Data collection

Considering the methodologies for short-term traffic prediction, NNs are classified as nonparametric statistical methods [39]. Looking for critical variables through the process of establishing the NN-based prediction model is the key to developing an accurate prediction model. Therefore, according to the suggestions of previous studies (cite references), we collected data of critical variables of travel time prediction. Furthermore, the data reliability of automated data collection is dependent on system stability and calibration accuracy. Therefore, in order to enhance the accuracy of travel time prediction in the areas with complicated traffic characteristics, collecting data of critical variables for effectively analyzing traffic characteristics and ensuring data reliability have become the important issues.

According to previous travel time prediction studies [7,40–43], spot speed, rainfall, historical travel time, and the day of the week and time (AM or PM) are important variables for improving the performance of travel time prediction. Therefore, in this study, spot speeds were collected by 11 dual loop vehicle detectors every 5 min. The rainfall data of three areas were collected every 10 min by rainfall detectors, and the data were transformed into those of 5-min intervals using the arithmetic mean method.

According to Petty et al. [21] and Chien and Kuchipudi [16], ATT can be obtained through a probe vehicle or AVI system, and ATT helps with establishing a robust prediction model. However, due to the high system construction cost, the studied segment and samples are limited. In Taiwan, the ETC system was established in 2006 and covers the entire National Freeway No. 1. Up to the end of October 2009, the utilization rate reached 36.48%, and there were a total of 16,247,908 charging records in October 2009, with a charging success rate of 99.9984%. Therefore, through the ETC system, the travel time data can be collected on a long road segment and the sample size can also be increased considerably to ensure sample representativeness. Furthermore, in order to obtain the actual travel time (ATT) and the historical travel time (HTT) respectively as the target and input variable of the prediction model, the vehicle charging time and ID were collected by ETC in this study. ATT and HTT were calculated by the methods presented in Section 3. In addition, Li and Chen [8] pointed out that encoding the days of the week as 1–7 and the time attribute as AM (0:00–12:00) or PM (12:00–24:00) can improve the performance of freeway travel time prediction in the case of non-recurrent congestion. To sum up, in this study, the spot speed, rainfall, day of the week, time (AM or PM), and HTT obtained from ETC were collected and used as input variables of the travel time prediction model.

In addition, to ensure data reliability, the raw data were collected from the database of Taiwan Area National Freeway

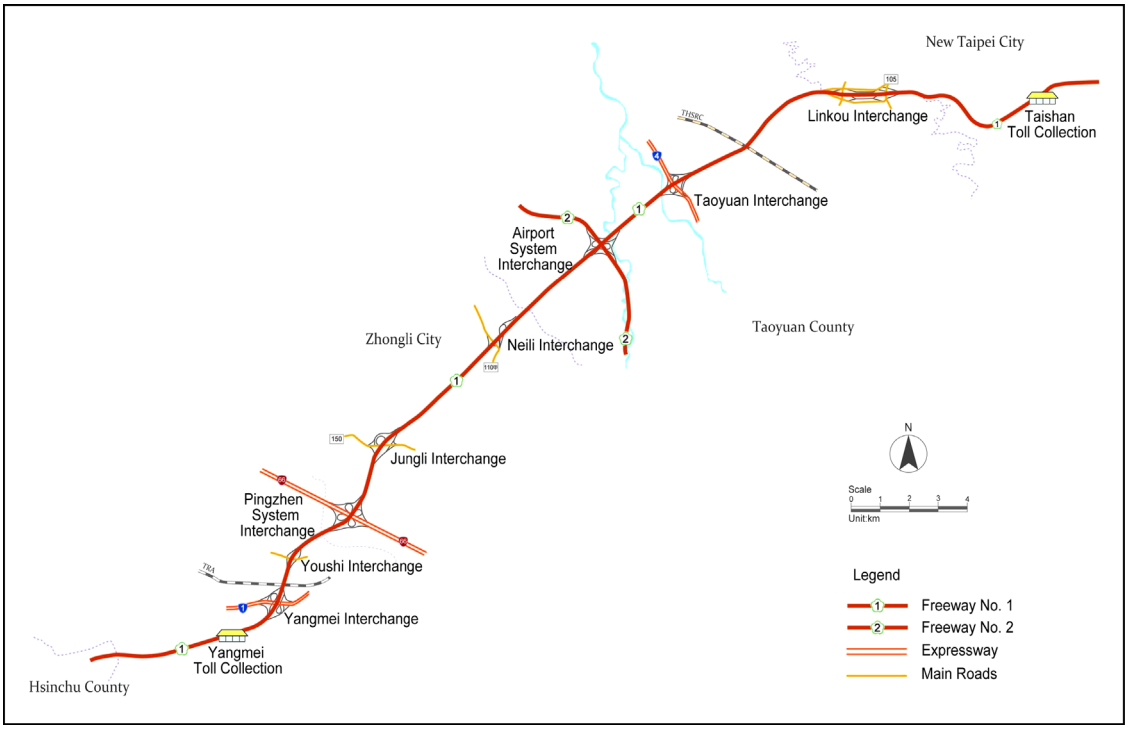

Fig. 1. The freeway segment in this study.

Bureau, MOTC, the database of the Central Weather Bureau, and the accident database of the National Highway Police Bureau. The rainfall data were collected from three rainfall detectors. The above-mentioned three databases are established by Taiwan's governmental agencies to permanently collect the most complete real-time data for information dissemination, management, and research use. Because ATT was used as the target for the model training and test, and HTT was used as the input variable, in order to avoid inconsistency between the result of model training and the actual condition as a result of data imputation error, the sample at time t with missing values of ATT and HTT were removed. Finally, 7908 samples were collected in this study.

2.3. Characteristics of non-recurrent congestion

It is undeniable that accidents and rainfall are the main causes of non-recurrent congestion, and are difficult to predict accurately. There were a total of 76 accidents with 176 vehicles damaged and six people injured in the time span of this study (see Fig. 2). In addition, 94.7% of accidents involved only vehicle damage without injuries to people. It is noteworthy that although fewer accidents occurred on the freeway section between Taoyuan interchange and Linkou interchange than on the freeway section between Jungli interchange and Neili interchange, more vehicles were damaged in accidents occurring in the former section than in the latter. Thus, the impact of accidents on traffic flow is greater in the freeway section between Taoyuan interchange and Linkou interchange.

The traffic flow of Tuesday can be taken as an example (see Fig. 3) to explain the impact of accident and rainfall on traffic flow. Observing the distribution of travel time of 13 October shown in Fig. 3(a), the morning peak hour of the freeway segment in this study occurred at about 7:00–8:20 AM, when there were no accidents or rainfall, whereas the peak time was prolonged, when there was rainfall. This can be approved in the cases of 29 September and 6 October. Due to the intermittent shower occurring during the period 5:25–6:40 AM on 29 September (see Fig. 3(b)), the morning peak hour of this day occurred at 7:20–8:50 AM. Furthermore, the intermittent rainfall during the period 7:15–8:50 AM on 6 October resulted in a morning peak hour of 7:00–9:30 AM.

According to Fig. 3, in the freeway segment in this study, the afternoon peak on Tuesdays was not obvious if there was no accident. However, when there were accidents, the travel time of this freeway segment increased significantly. Taking 29 September as an example, accidents occurred consecutively between 15:10 and 18:45 (see Fig. 3(c)), and accordingly there was a peak hour from 16:30 to 18:45 in this freeway segment (see Fig. 3(a)). Furthermore, there was an accident involving three vehicles between 10:40 and 11:27 AM in the morning of 22 September, which resulted in a morning peak hour lasting from 7:15 to 11:25 AM. Hence, the accident and rainfall are critical factors resulting in non-recurrent congestion. The occurrence of accident has randomness as well as the important variables such as accident occurrence time, number of closed lanes and accident removal time used to estimate the impact of accident on traffic flow cannot be acquired in real time due to the limitation of notification scheme. Therefore, this study attempts to develop a robust model of travel time prediction in the case of non-recurrent congestion without the real time accident related information.

3. The proposed procedure of travel time prediction

In order to establish a robust travel time prediction model for the freeway with non-recurrent congestion, and to obtain the critical variables of important detector locations for the management unit with existing equipment, this study tries to achieve the research objectives based on the procedure as shown in Fig. 4. Each step is described as follows.

3.1. Step 1: input data

In this study, raw data such as spot speed, rainfall, day of the week, time (AM or PM), and vehicle charging time and ID of ETC are collected from the databases established by Taiwan's governmental agencies. Furthermore, ATT, HTT, dummy variables, important detectors, and critical variables are obtained through the following steps.

3.2. Step 2: compute ATT and HTT

The Southwest Research Institute [44] developed two algorithms, TransGuide and TranStar, by using the AVI system to calculate the actual travel time. Both these algorithms use the concept of rolling average to automatically calculate the travel time which meets the threshold-based criterion. Furthermore, TransGuide and TranStar both set 0.2 (20%) as the threshold value. That is, if the travel time of vehicle i is 20% more or less than the previous average travel time $\mathrm{Btt}_{\mathrm{ABt}}$, it is regarded as an abnormal trip, and this sample is removed. For the further details, readers are referred to Dion and Rakha [24]. In addition, differing from the rolling average algorithm, the Transmit algorithm constantly makes the calculation with the travel times from those samples within 15-min intervals. The travel time is the time difference $(t_{Bi}-t_{Ai})$ between downstream point B, t_{Bi}, and the upstream point A, t_{Ai}. Furthermore, the Transmit algorithm collects the travel time samples of two AVI readers, N, in each constant time interval t, with an upper limit of 200 samples, to calculate the average travel time ρ_{ABt} within the time interval by using the following equation [45]:

$$\rho_{ABt}=\frac{\sum_{i=1}^{N}(t_{Bi}-t_{Ai})}{N} \tag{1}$$

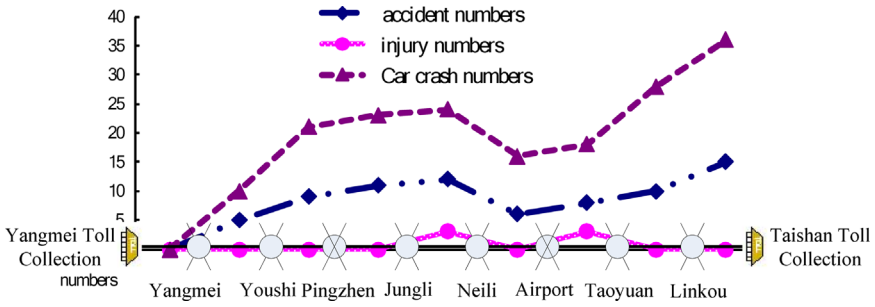

Fig. 2. Number of accidents, injuries, and car crashes in the studied segment.

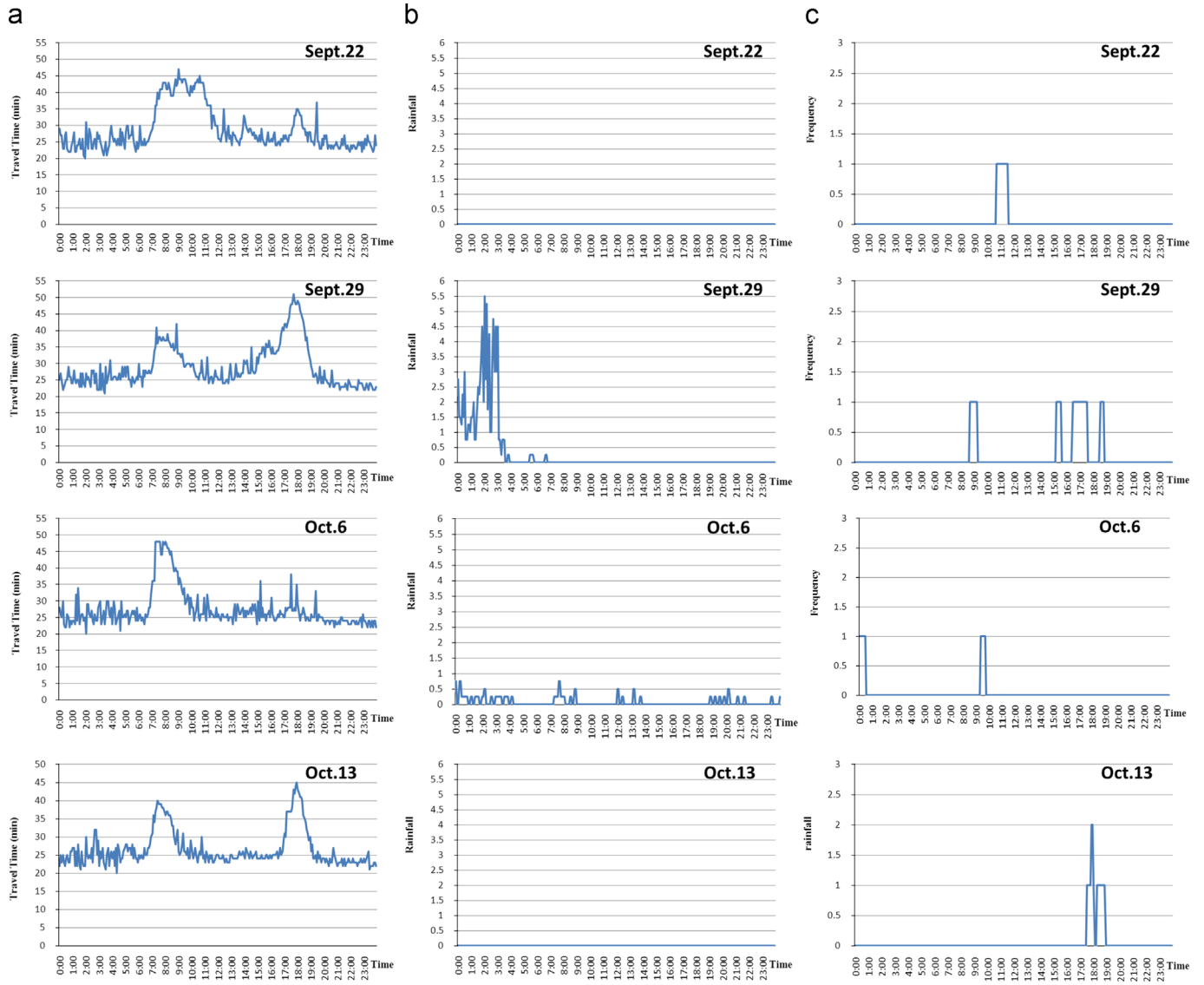

Fig. 3. Analysis of travel time, accident and rainfall on Tuesday. (a) Travel time, (b) rainfall and (c) distribution of number of vehicles involved in accidents and duration of accidents.

In this study, the charging times and ID of northbound vehicles passing through Yangmei Toll Station and Linkou Toll Station were collected through the ETC system. A total of 1,679,868 data points were collected, and the Transmit algorithm was employed to compute HTT and ATT at 5-min intervals without the restriction of 200 samples. The computation of historical travel time (HTT_{ABt}) is expressed by the following equation:

$$HTT_{ABt} = \{t_{Bi} - t_{Ai} | t - t_r \le t_{Bi} \le t \ \text{ and } \ Btt_{ABt}(1-0.4) \le t_{Bi} - t_{Ai} \le Btt_{ABt}(1+0.4)\} \quad (2)$$

In Eq. (2), with the time of vehicle i passing through point B, t_{Bi}, as the judgment basis, data of vehicles passing through point B during an interval of five minutes are collected, and the completed trips between upstream point A and downstream point B are utilized as the samples to compute the average travel time as the historical travel time (HTT_{ABt}). Although HTT does not represent the ATT (ATT_{ABt}) of vehicle i traveling in the freeway section AB from upstream point A, it may imply the historical traffic characteristics of the freeway section AB. According to Fei et al. [7], HTT can be used to effectively adjust the real time prediction model. Thus, in this study, HTT_{ABt} is used as an input variable of the NN-based prediction model.

The computation of actual travel time (ATT_{ABt}) is expressed by the following equation:

$$ATT_{ABt} = \{t_{Bi} - t_{Ai} | t - t_r \le t_{Ai} \le t \ \text{ and } \ Btt_{ABt}(1-0.4) \le t_{Bi} - t_{Ai} \le Btt_{ABt}(1+0.4)\} \quad (3)$$

In Eq. (3), with the time of vehicle i passing through point A, t_{Ai}, as the judgment basis, data of vehicles passing through point A during an interval of 5 min are collected, and the average travel time of vehicles completing the freeway section AB is computed. This travel time represents the ATT taken by vehicle i after passing through point A to enter the freeway section AB. The actual travel time (ATT_{ABt}) is taken as the target of prediction model. The threshold value of checking whether it is a continuous trip is set to 0.4. For example, if the average travel time at time t–1 is 30 min and the travel time of sample i at time t is more than or equal to 18 min or less than or equal to 42 min, it is regarded as a continuous trip. This sample can be used to compute the travel time. Furthermore, both HTT_{ABt} and ATT_{ABt} apply the same threshold value, 0.4. For different areas, the threshold value may be

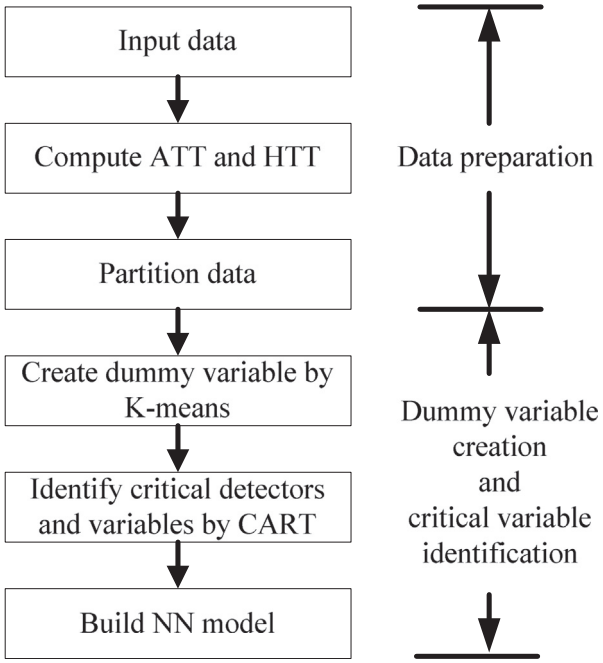

Fig. 4. Procedure of travel time prediction.

different. The threshold value of 0.4 is used by Taiwan Area National Freeway Bureau, MOTC.

3.3. Step 3: partition data

After processing the data with Step 2, the processed data were divided into the training data set and the test data set in the ratio of 7:3 in order to build the travel time prediction model. However, selecting 70% of the data randomly as the training data set and the remaining 30% as the test data set is not applicable because non-recurrent congestion is considered in this study. If the data are directly divided in a random manner, the samples of non-recurrent congestion may be grouped mainly into either the training data set or the test data set. Such a situation may result in over- or under-estimation in the prediction model. Thus, in this study, all the samples were firstly clustered by using *K*-means. Then, each cluster was randomly divided into the training data set and test data set in the ratio of 7:3. The clustering is used to ensure that the samples of various traffic characteristics are randomly selected in the training data set and test data set for avoiding the error in implementing the prediction model.

3.4. Step 4: create dummy variable by K-means

In this study, clustering is used to assign samples which have the same traffic characteristics to the same group, and then generate a dummy variable to the samples. The training data set was clustered through *K*-means, and a dummy variable was added to the input variables to represent the cluster ID in order to build the travel time prediction model. Firstly, the samples of training data set were used to construct the CART-based classification model by utilizing the input variables obtained in the previous steps as the attributes of classification, and the group ID as the class label. Secondly, the cluster ID (e.g., group 1, group 2, etc.) of each sample in the test data set was labeled through the constructed classification model. To summarize, the process of creating the dummy variable is illustrated in Fig. 5. The cluster ID was taken as the dummy variable for building the model of travel time prediction with the training data set, and the built model was used to predict the travel time of the test data set.

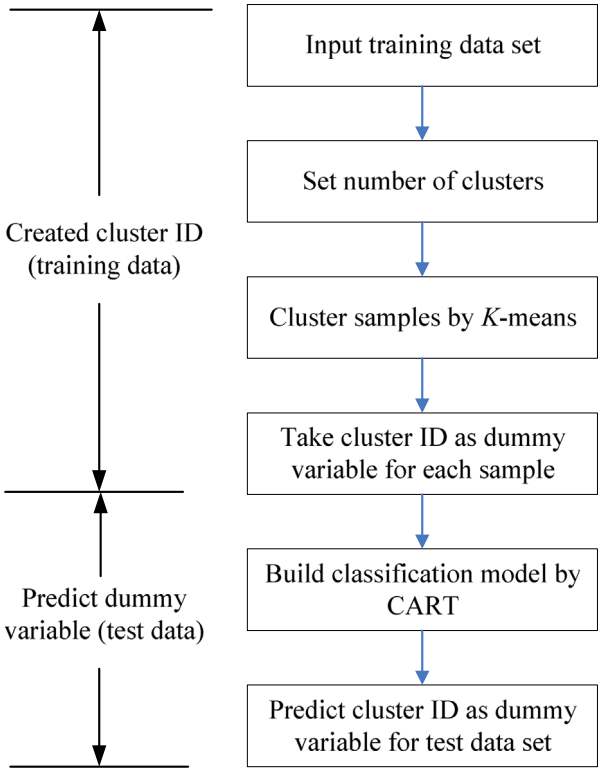

Fig. 5. Process of creating the dummy variable.

3.5. Step 5: identify critical detectors and variables by CART

The CART-based classification model can be used not only to predict the cluster ID of each sample in the test data set, but also to identify the important variables, which appear in the decision tree. The identified important variables can be used to construct the NN-based model of travel time prediction. This prediction model could be used in real-time data forecasting, information publishing, or traffic management. The procedure described above is illustrated in Fig. 6.

3.6. Step 6: build NN-based prediction model

After adding the dummy variable and identifying the important variables, various NN-based models of travel time prediction with different combinations of variables can be constructed to analyze the prediction performance. In order to analyze the impact of HTT, dummy variable and identified critical variables on the prediction performance, six experimental combinations are designed, and the three-layer NN is used to construct the prediction model. Furthermore, each experimental combination uses the training data set to construct the best NN-based prediction model. Then, it is tested with the test data set.

In this study, SAS Enterprise Miner 5.3 is used to construct the back-propagation neural network (BPN) with three layers, input, hidden and output layers, CART, and *K*-means. The detailed algorithm of BPN can be found in [46], and the detailed algorithms of *K*-means and CART can be found in [47].

4. Experimentation

4.1. Experimental design

This study focused on exploring how to establish a robust model to predict the travel time of freeway with non-recurrent

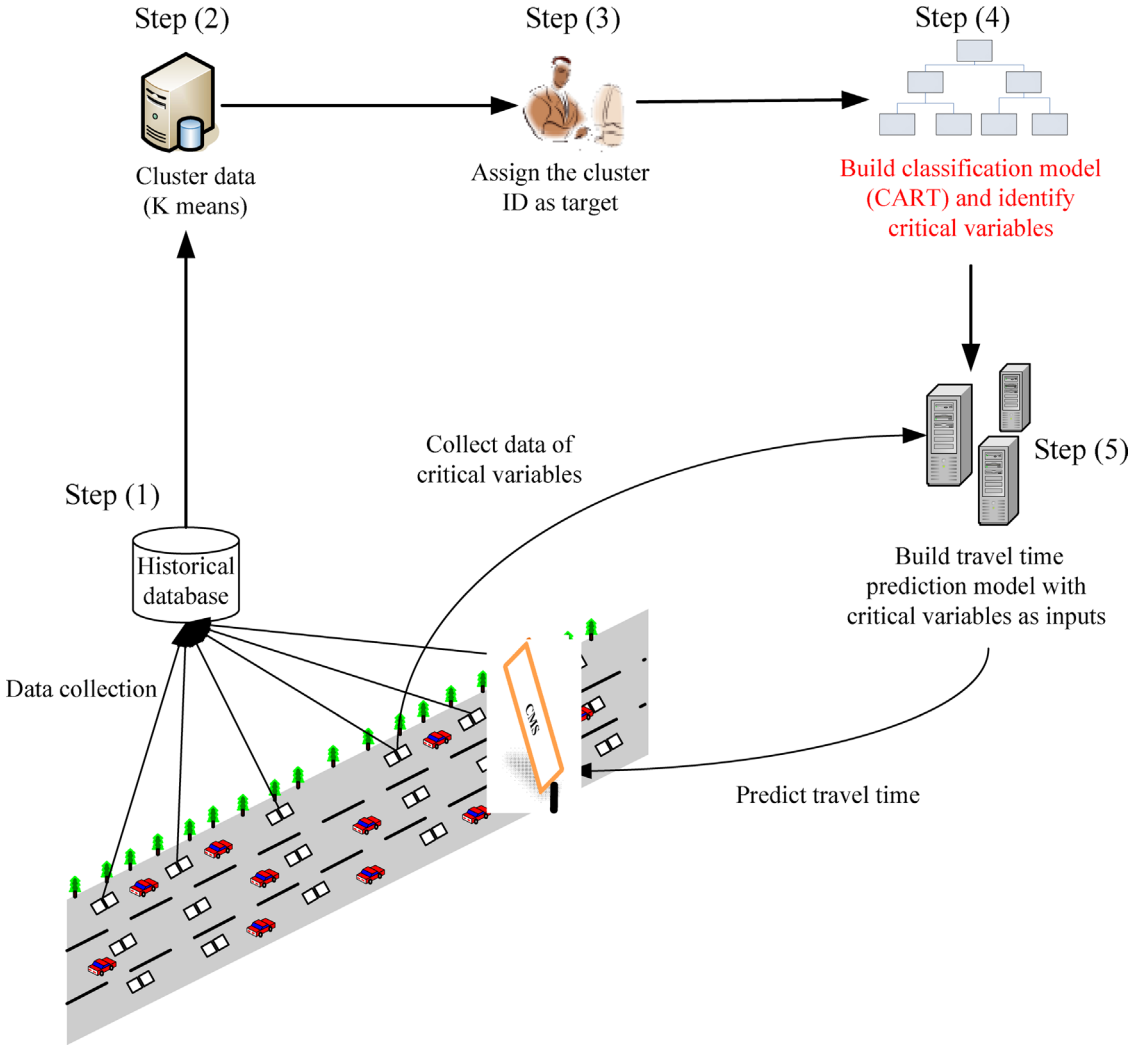

Fig. 6. The procedure of identifying important variables and constructing travel time prediction model.

congestion by using existing or even simplified equipment. Six scenarios were designed for to investigate the performance of proposed prediction approach in this study. Scenario 1 uses spot speeds collected by 11 vehicle detectors, rainfalls collected by detectors in three locations, day of the week, and time (AM or PM) as input variables to build the NN-based model of travel time prediction. Because the number of hidden nodes is an important parameter for the BPN-based prediction model, various numbers of hidden nodes were investigated to select the best model for predicting the travel time. The experimental details of BPN can be found in [8]. In contrast with Scenario 1, Scenario 2 adds HTT calculated from ETC as the input variable. Scenario 3 adds the dummy variable (i.e., cluster ID) generated by *K*-means clustering with the input variables of Scenario 1. The input variables of Scenario 4 include the variables in Scenario 2 and cluster ID. Scenario 5 takes the cluster ID as the class label of the classification in Scenario 3, and the CART is adopted to select the important variables for predicting the travel time with NN. The prediction procedure of Scenario is described in Step 5 in Section 3. The input variables in Scenario 6 include the important variables identified by using CART with the variables in Scenario 4. These six scenarios are summarized in Table 1.

From the above discussion, the important variables in Scenarios 5 and 6 are all selected by using CART. Observing the decision trees in Scenarios 5 and 6, the same important variables are identified,

Table 1
Summary of six scenarios.

Scenario	Description of input variables
1	Spot speeds collected by 11 VDs, rainfalls from three detectors, the day of the week, and time (AM or PM)
2	The input variables of S1, and HTT calculated from ETC
3	The input variables of S1, and the dummy variable (i.e., cluster ID)
4	The input variables of S2, and the dummy variable (i.e., cluster ID)
5	The important variables identified by CART with the variables in Scenario 3
6	The important variables identified by CART with the variables in Scenario 4

which include time (AM or PM), the day of the week, and the spot speed collected by the VD located at 51.6 km, which is denoted as speed 5160.

4.2. Experimental results

The general traffic flow contains the morning peak and afternoon peak, and a noon peak in the areas with the complicated or special traffic flow. Moreover, the traffic management unit usually distinguishes the midnight off-peak traffic characteristic from the

general off-peak condition. Therefore, it is common to divide the traffic characteristics into eight groups in the areas with complicated traffic flow, and to develop corresponding management strategies according to their traffic characteristics. In this study, the data of traffic flow are also divided into eight groups. Furthermore, before clustering and classification, the data are standardized by using $\delta_{ij}=(\varepsilon_j-\varphi_{ij})/\varepsilon_j$, where δ_{ij} represents the standardized value of ith sample's variable j, φ_{ij} represents the original value of ith sample's variable j, and ε_j represents the maximum value of all samples of variable j plus 20%, that is $\varepsilon_j=\max\{\varphi_i\}\times 1.2$. In the CART-based classification, the training data set are randomly the ratio 7:3 divided into the training stage and the validation stage with the ratio of 7:3. From the results of decision trees, the classification accuracy rates of the training stage and validation stage for Scenarios 3 and 4 are above 99.85% for both S3 and S4. Therefore, the constructed CART-based models are able to effectively differentiate the traffic characteristics of the samples collected in this study.

In order to assess the performance of prediction models, the mean absolute percentage error (MAPE) is adopted as the performance metric, and it is expressed by the following equation [48]:

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{ATT_i-PTT_i}{ATT_i}\right| \tag{5}$$

where ATT_i represents the actual travel time of ith sample, PTT_i represents the predicted travel time of ith sample, and n is the number of samples. Therefore, the smaller MAPE is, the higher prediction accuracy is. MAPE was proposed by Lewis [48] and has been widely used as a performance metric of prediction. In the case of MAPE $\leq 10\%$, the model has the "highly accurate prediction capability". In the case of $11\% <$ MAPE $\leq 20\%$, the model has a "good prediction capability". In the case of $21\% <$ MAPE $\leq 50\%$, the model has a "reasonable prediction capability". In the case of MAPE $> 51\%$, the model has an "inaccurate prediction capability". Therefore, if we can construct a model with a "highly accurate prediction" capability and a smaller number of samples with MAPE $> 21\%$, the prediction model not only can help with satisfying the road users' requirements regarding the travel time prediction but also have a positive effect on ITS implementation.

According to the experimental results as shown in Table 2, the MAPE values of the six scenarios are between 6% and 9%. Thus, the prediction models of freeway travel time constructed in this study all have "highly accurate prediction" capability. From Table 2, the MAPE of Scenario 1 (6.68%) is higher than that of Scenario 3 (6.49%) as well as the MAPE of Scenario 2 (6.41%) is close to that of Scenario 4 (6.47%), it is demonstrated that additionally including the cluster ID as the input variable can enhance the model's prediction capability in the environment of having an ETC system or in the environment of having traditional VD detectors.

Furthermore, from Table 2, the percentage of samples with MAPE $> 20\%$ is lower in Scenario 3 (2.62%) compared to that Scenario 1 (3.74%). The percentage of samples with MAPE $> 20\%$ is lower in Scenario 4 (2.74%) compared to Scenario 2 (2.99%). Therefore, additionally including the cluster ID as the input variable can effectively lower the percentage of samples with MAPE $> 20\%$. No matter there is AVI or ETC, the dummy variable (i.e., cluster ID) can improve the performance of travel time prediction. In particular, in the case of collecting the traffic data only with VDs, the dummy variable can more significantly improve the performance of travel time prediction, and notably reduce the percentage of samples with MAPE $> 20\%$.

Table 2
The performance of six scenarios.

Scenario	MAPE (%)	The percentage of samples with MAPE > 20%
1	6.68	3.74
2	6.41	2.99
3	6.49	2.62
4	6.47	2.74
5	8.94	10.22
6	8.94	10.22

In addition, if the prediction models (Scenarios 5 and 6) are constructed only with important variables, the percentage of samples with MAPE $> 20\%$ increases two to four times compared with other scenarios. The prediction model containing more important variables increases the understanding of traffic characteristics and enhances the prediction performance. Therefore, the experimental results of Scenarios 5 and 6 are not unusual. From the results of Scenarios 5 and 6, the day of the week, time (AM or PM) and speed 5160 are extracted as the important variables. Two of these three important variables, the day of the week and time (AM or PM), can be collected by VDs. Therefore, for effective management, speed 5160, i.e., the spot speed collected by the VD at 51.6 km, is a critical element requiring careful maintenance. With the important variables extracted in Scenarios 5 and 6, only one out of 14 detectors (11 VDs and 3 rainfall detectors) is identified as the important detector such that the operational cost and maintenance cost can be significantly reduced. Furthermore, the management unit can maintain and calibrate the data collection system, and develop the imputation method of missing data based on the experimental results. Although the models of travel time prediction constructed in Scenarios 5 and 6 perform worse than the models of other scenarios, the models of Scenarios 5 and 6 have the "highly accurate prediction capability" according to the classification defined by Lewis [46]. From the results of Scenarios 5 and 6, time (AM or PM) is also identified as an important variable by CART, it is similar to the finding in Fei et al. [7] confirming that the traffic characteristics with non-recurrent congestion in the morning and afternoon are indeed different. Therefore, taking time (AM or PM) as the input variable or partitioning the data with respect to time (AM or PM) and analyzing accordingly can improve the performance of travel time prediction in the case of non-recurrent congestion.

5. Conclusions and suggestions

Predicting the travel time of freeway with non-recurrent congestion is essential in the area of traffic and transportation, but it is a challenge to achieve a high degree of prediction accuracy with less data and lower cost. Furthermore, the ability to (1) enhance the model prediction capability with existing equipment and (2) obtain important variables in important locations in order to reduce the equipment maintenance cost and retain the prediction accuracy is an important issue that has been paid much attention by management and research organizations. In this study, about several million data collected by ETC have been used to obtain the actual travel time for predicting the travel time. In addition, following the empirical analysis of National Freeway No. 1 between the Yangmei Toll Station and Taishan Toll Station in the northward direction, we found that a robust travel time prediction model with non-recurrent congestion could be constructed by integrating K-means, decision tree, and neural network.

From the experimental results of this study, the performance of freeway travel time prediction with non-recurrent congestion could be improved by the added dummy variables and the proposed method of extracting important variables. According to the results of this study, increasing the number of dummy

variables and using them as input variables could enhance the prediction capability of model and lower the percentage of the sample whose MAPE > 20% without increasing the amount of equipment needed. This could enhance public acceptance of travel time prediction. For example, in the six-lane two-way freeway, in a specific direction (outbound or inbound), if the travel time information is updated every 5 min, the travel time prediction model runs continuously for 30 days, and there are 600 PCU (the total in three lanes) passing each changeable message sign (CMS) every 5 min and two passengers in each vehicle, then there will be 10,368,000 passengers in total receiving travel time information from each CMS within a month. Even if we only eliminate 0.1% of the sample whose MAPE > 20%, this could reduce negative perception of the prediction model by 10,368 people every month. If we set up a number of CMSs, n, along the freeway in two bounds at the same time, it will influence $2n$ passengers ($10{,}368 \times 2n$ people). Furthermore, it is confirmed by this study that the important variable extraction method with decision tree can not only maintain high prediction accuracy but also significantly reduce the cost of equipment maintenance and operation to comply with the demand of management organization.

Increasing the speed of shock wave and calculating the distance of queue may improve the accuracy of travel time forecasting. The future work can take them into consideration in forecasting travel time. Reducing the percentage of samples whose MAPE > 20% has a significant impact on the road users' satisfaction. Although the method proposed in this study can reduce the percentage of samples whose MAPE > 20%, reducing the percentage of samples whose MAPE > 20% is an issue which needs to be addressed continuously, and it is worth to be further studied in the future.

Acknowledgments

This work is partially supported by National Science Council, Taiwan, ROC, under Grant NSC 100–2410-H-009–013-MY3.

References

[1] International Energy Agency Statistics. CO_2 Emission from Fuel Combustion. IEA Publications, 9, rue de la Federation, 75739 Paris Cedex 15 Printed in Luxembourg by Imprimerie Centrale, October 2011.
[2] R. Joumard, P. Jostb, J. Hickman, D. Hasselb, Hot passenger car emissions modelling as a function of instantaneous speed and acceleration, Sci. Total Environ. 169 (1995) 167–174.
[3] A. Dharia, H. Adeli, Neural network model for rapid forecasting of freeway link travel time, Eng. Appl. Artif. Intell. 16 (7) (2003) 607–613.
[4] W.H.K. Lam, K.S. Chan, M.L. Tam, J.W.Z. Shi, Short-term travel time forecasts for transport information system in Hong Kong, J. Adv. Transp. 39 (3) (2005) 289–306.
[5] Chin S.M., Franzese O., Greene D.L., Hwang H.L., Gibson R.C., Temporary Loss of Highway Capacity and Impacts on Performance: Phase 2. Report No. ORLNL/TM-2004/209, Oak Ridge National Laboratory, 2004.
[6] A. Skabardonis, P. Varaiya, K.F. Petty, Measuring recurrent and nonrecurrent traffic congestion, Transp. Res. Rec. 1856 (2003) 118–124.
[7] X. Fei, C.C. Lu, K. Liu, A Bayesian dynamic linear model approach for real-time short term freeway travel time prediction, Transp. Res. Part C: Emerg. Technol. 19 (6) (2011) 1306–1318.
[8] C.S. Li, M.C. Chen, Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks, Neural Comput. Appl. 23 (6) (2013) 1611–1629.
[9] T.F. Golob, W.W. Recker, V.M. Alvarez, Freeway safety as a function of traffic flow, Accid. Anal. Prev. 36 (6) (2004) 933–946.
[10] M. Vanderschuren, Safety improvements through intelligent transport systems: a South African case study based on microscopic simulation modelling, Accid. Anal. Prev. 40 (2) (2008) 807–817.
[11] Mitretek Systems, Intelligent Transport System Benefits: 2001 Update Under Contract to the Federal Highway Administration. US Department of Transportation, Washington, DC, US, 2001.
[12] J.Y. Kwon, B. Coifman, P. Bickel, Day-to-day travel-time trends and travel-time prediction from loop detector data, Transport. Res. Rec. 1717 (2000) 120–129.
[13] Oda T., An algorithm for prediction of travel time using vehicle sensor data, in: Third International Conference on Road Traffic Control, Institute of Electrical Engineers, 1990, pp. 40–44.
[14] B. Van Arem, M.J.M. Van Der Vlist, M.R. Muste, S.A. Smulders, Travel time estimation in the GERDIEN project, Int. J. Forecast. 13 (1997) 73–85.
[15] M. Chen, S. Chien, Dynamic freeway travel-time prediction with probe vehicle data, link based versus path based, Transport. Res. Rec. 1768 (2001) 157–161.
[16] S. Chien, C.M. Kuchipudi, Dynamic travel time prediction with real-time and historic data, J. Transp. Eng. 129 (6) (2003) 608–616.
[17] A. Stathopoulos, M.G. Karlaftis, A multivariate state space approach for urban traffic flow modeling and prediction, Transp. Res. C 11 (2) (2003) 121–135.
[18] C. Nanthawichit, T. Nakatsuji, H. Suzuki, Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway, Transp. Res. Rec. 2987 (2003) 49–59.
[19] Chu L.Y., Oh J.S., Recker W., Adaptive Kalman filter based freeway travel time estimation. Paper Presented at the 84th TRB Annual Meeting, Washington, DC, January 2005.
[20] Y. Wang, M. Papageorgiou, Real-time freeway traffic state estimation based on extended Kalman filter: a general approach, Transp. Res. B 39 (2) (2005) 141–167.
[21] K.F. Petty, P. Bickel, M. Ostland, J. Rice, F. Schoenberg, J. Jiang, Y. Ritov, Accurate estimation of travel time from Single-Loop detectors, Transp. Res. A 32 (1) (1998) 1–18.
[22] D.J. Park, L.R. Rilett, Forecasting multiple-period freeway link travel times using modular neural networks, Transp. Res. Rec. 1617 (1998) 163–170.
[23] Hoffmann G., Janko J., Travel times as a basic part of the LISB guidance strategy, in: Proceedings of the Third International Conference on Road Traffic Control. Institution of Electrical Engineers, London, England, 1990, pp. 6–10.
[24] F. Dion, H. Rakha, Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates, Transp. Res. Part B 40 (2006) 745–766.
[25] R. Li, G. Rose, Incorporating uncertainty into short-term travel time predictions, Transp. Res. C 19 (2011) 1006–1018.
[26] Cui Y., Huang Q., Character extraction of license plates from video, in: Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1997, pp. 502–507.
[27] H.R. Kirby, S.M. Watson, M.S. Dougherty, Should we use neural networks or statistical models for short-term motorway traffic forecasting? Int. J. Forecast. 13 (1) (1997) 43–50.
[28] D.J. Park, L. Rilett, G. Han, Spectral basis neural networks for realtime travel time forecasting, J. Transp. Eng. 125 (6) (1999) 515–523.
[29] J.W.C. van Lint, S.P. Hoogendoorn, H.J. Van Zuylen, Freeway travel time prediction with state-space neural networks-modeling state-space dynamics with recurrent neural networks, Transp. Res. Rec. 1811 (2002) 30–39.
[30] T. Park, S. Lee, A Bayesian approach for estimating link travel time on urban arterial road network, Lect. Notes. Comput. Sci. 3043 (2004) 1017–1025.
[31] J.W.C. Van Lint, S.P. Hoogendoorn, H.J. Van Zuylen, Accurate travel time prediction with state-space neural networks under missing data, Transp. Res. C 13 (2005) 347–369.
[32] S. Innamaa, Short-term prediction of travel time using neural networks on an interurban highway, Transportation 32 (2005) 649–669.
[33] E. Mazloumi, G. Ros, G. Currie, S. Moridpour, Prediction intervals to account for uncertainties in neural network predictions: methodology and application in bus travel time prediction, Eng. Appl. Artif. Intell. 24 (3) (2011) 534–542.
[34] M. Zhong, P. Lingras, S. Sharma, Estimation of missing traffic counts using factor, genetic, neural, and regression techniques, Transp. Res. C 12 (2004) 139–166.
[35] B. Raj, R.M. Stern, Missing-feature approaches in speech recognition, IEEE Signal Process. Mag. 22 (2005) 101–106.
[36] C. Cerisara, S. Demange, J.P. Haton, On noise masking for automatic missing data speech recognition: a survey and discussion, Comput. Speech Lang. 21 (3) (2007) 443–457.
[37] S. Demange, C. Cerisara, J.P. Haton, Missing data mask estimation with frequency and temporal dependencies, Comput. Speech Lang. 23 (2009) 25–41.
[38] R. Lederman, L. Wynter, Real-time traffic estimation using data expansion, Transport. Res. B 45 (2011) 1062–1079.
[39] van Hinsbergen C.P., van Lint J.W.C., Sanders F.M., Short term traffic prediction models, in: Proceedings of the 14th World Congress on Intelligent Transport System (CD-ROM) Beijing, China, 2007.
[40] F. Yuan, R.L. Cheu, Incident detection using support vector machines, Transp. Res. C 11 (2003) 309–328.
[41] L.Y. Chang, Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network, Saf. Sci. 43 (8) (2005) 541–557.
[42] C.H. Wei, Y. Lee, Sequential forecast of incident duration using artificial neural network models, Accid. Anal. Prev. 39 (2007) 944–954.
[43] J. Yeon, L. Elefteriadou, S. Lawphongpanich, Travel time estimation on a freeway using discrete time Markov chains, Transp. Res. B 42 (2008) 325–338.
[44] SwRI, Automatic vehicle identification model deployment initiative – system design document. Report Prepared for TransGuide, Texas Department of Transportation, Southwest Research Institute, San Antonio, TX, 1998.
[45] Mouskos K.C., Niver E., Pignataro L.J., Lee S., Transmit system evaluation. Final Report, Institute for Transportation, New Jersey Institute of Technology, Newark, NJ, 1998.
[46] S. Kumar, Neural Networks: A Classroom Approach, TATA McGraw-Hill Publishing Company Limited, Boston, 2004.

[47] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education, Inc., Boston, 2006.
[48] C.D. Lewis, Industrial and Business Forecasting Methods, Butterworth-Heinemann, London, 1982.

Chi-Sen Li is a Ph.D. candidate in the Department of Transportation and Logistics Management at National Chiao Tung University, Taipei, Taiwan. He received his M.Sc. degree in the Department of Civil Engineering from National Central University and B.S. degree in Department of Transportation Engineering from Feng Chia University. His research interests include Data Mining, Travel Time Prediction and Logistics Management.

Mu-Chen Chen is a professor of Department of Transportation and Logistics Management in National Chiao Tung University, Taipei, Taiwan. He received his Ph.D. and M.Sc. degrees both in Industrial Engineering and in Management from National Chiao Tung University, and his B.S. degree in Industrial Engineering from Chung Yuan Christian University. His teaching and research interests include Data Mining, Logistics and Supply Chain Management and Meta-heuristics.