Short Term Prediction of Freeway Exiting Volume Based on SVM and KNN

Xiang Wang*
Ph.D. Candidate, Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University 4800 Cao'an Road, Shanghai, 201804, P.R. of China
Tel)+86-188-0196-3176 Fax)+86-2165962897
wangxiang_tjjt@hotmail.com

Kang An
Ph.D. Candidate, Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University 4800 Cao’an Road, Shanghai, 201804, P.R. of China
Tel)+86-136-21816789 Fax)+86-2165962897
augustuskangan@gmail.com

Liang Tang
Graduate Research Assistant, Department of Civil and Environmental Engineering, Transportation Systems Research Laboratory, 3109 Jeong H. Kim Engineering Building, University of Maryland, College Park, MD 20740,
Tel) 301-405-2926 Fax)+301-314-5139
liang@umd.edu

Xiaohong Chen
Professor, Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University 4800 Cao’an Road, Shanghai, 201804, P.R. of China
Tel)+86-216596270 Fax)+86-2165962897
chenxh@tongji.edu.cn

ABSTRACT
In order to better predict the traffic states on freeways and make management decisions, a hybrid model of support vector machine (SVM) and K-nearest neighbor (KNN) is proposed for short-
term freeway exiting volume prediction.

First, a historical data set is built by using the freeway toll data. The abnormal toll records, such as records that have same entry and exit station, illogical time record and abnormal travel speed, are excluded by data quality control. Based on the historical dataset, it is found that the exiting volume has periodical variation over time which provides the basis of the short-term prediction. Then, the historical data set is cross-classified into twelves groups based on the day of week and time of day. The prediction has been done for each group. Finally, the prediction is accomplished by the hybrid-model of SVM and KNN. The exiting volumes of previous time periods are used as the feature vector for KNN and SVM. Besides, a dynamic weight is adopted for the prediction of current time period based on the latest prediction accuracy of KNN and SVM.

The model results indicate that the proposed algorithm is feasible and accurate. The Mean Absolute Percentage Error is under 10%. When comparing with the results of single KNN or SVM method, the results show that the combination of KNN and SVM can improve the reliability of the prediction significantly. The proposed method can be implemented in the on-line application of exiting volume prediction, which is able to consider different vehicle types.

Keywords: Freeway Exiting Flow, Short-term Prediction, KNN, SVM, Hybrid Model

INTRODUCTION
Travel demand management (TDM) is defined as providing travelers with effective choices to improve travel reliability. (FHWA, 2006) TDM includes four categories of strategies, i.e. operational, infrastructure, financial and institutional. Short-term prediction of freeway exiting volume can help to better recognize future traffic state and make proactive operational management decisions. Through this, transport system efficiency and reliability can be improved.

Toll data on freeways contain information about when and where a car enters and leaves the freeway. This data can provide information about the traffic state on the freeway. Using the whole OD matrix analysis, the overall state of the traffic network, patterns, and characteristics can be obtained.

The key problem of traffic flow forecasting is how to effectively use real-time traffic data to predict the traffic state of the next time period (usually 5-15 minutes), such as traffic volume, speed and so on. Researchers have used many methods to predict traffic flow, such as time-series model, the Kalman filter model, nonparametric regression model and neural network models.

In order to help the freeway administrators arrange the work at toll plaza ahead of time, a short-term prediction algorithm is proposed to predict the exiting volume of freeway toll stations based on the Support Vector Machine (SVM) and K Nearest Neighbor (KNN) methods. The exiting volumes of previous time periods are used as the feature vector for KNN and SVM. Besides, a dynamic weight is adopted for the prediction of current time period based on the latest prediction accuracy of KNN and SVM. The experimental results show that the proposed algorithm is feasible and accurate. The proposed method can be implemented in the on-line application of exiting
volume prediction, which is able to consider different vehicle types.

**LITERATURE REVIEW**

**Support Vector Machine**

Support vector machine (SVM) is a machine learning method based on statistical learning theory, Vapnik-Chervonenkis dimension theory and structural risk minimization principle. It can effectively deal with regression (time series analysis), pattern recognition (classification problem, discriminant analysis) and many other problems. SVM is popular in the field of prediction and comprehensive evaluation [1]. It performs well in solving small sample, nonlinear and high dimensional pattern recognition problems and has many unique advantages, by overcoming the “curse of dimensionality”, “over-learning” and other issues.

Currently, SVM has been applied to traffic prediction with good performance. Based on nonlinear phenomenon of traffic, complexity and uncertainty, Yang Zhaosheng[2], etc proposed a short-term traffic flow forecasting method using support vector machine model, which performed better than BP neural network model in terms of accuracy, convergence time, generalization capabilities [2]. Road traffic detector data was used in the study. Using license plate recognition historical data, Zhang Juan and Sun Jian built a link travel time short-term prediction model based on non-linear support vector machine regression method, which has better accuracy than traditional exponential smoothing method, multivariate regression, autoregressive integration moving average (ARIMA) models [3]. Tan Man-chun and others proposed a method combining support vector machine and ARIMA to forecast short-term road traffic volume, and found that the combination model had better prediction accuracy than the single model [4].

Freeway exiting flow does not have a simple linear relationship with time. The volume may vary in different time of day. It is unreasonable to predict the short term exiting flow with simple linear regression and similar methods, while SVM is suitable for short-term traffic flow prediction.

**KNN Algorithm**

K-nearest neighbors (KNN) is a non-parametric prediction algorithm. It searches the K most similar feature vectors within the historical database to predict future values. The model has simple structure and high computation efficiency.

KNN has been used for traffic state prediction. Gong and Tang built a road traffic flow short-term forecast model. The traffic flow and speed coming from microwave remote sensor data in previous time interval were used as feature vectors [8]. Zhang proposed a short-term traffic flow forecasting method based on road detection data [9]; Yu used taxi GPS data, taking into account both the time dimension and spatial dimension built feature vectors to achieve a short time speed prediction [10]. Sun built a short-term traffic flow forecasting method, where travel time was used as feature vectors [11]. Bustillos and Chiu proposed single- and multi-neighbor travel time prediction method by combining the N-Curve method and KNN method. The cumulative number of vehicles was used as the feature vector [12]. Myung proposed a short-term travel time prediction algorithm with automated freeway toll data and loop
detection. Travel time and density were used as feature vectors respectively [13].

Freeway toll data contain rich historical information and can be used as input of KNN model. Freeway exiting volume is nonlinearity in time series and is also suitable for the KNN algorithm.

**Data Description and Quality Control**

**Freeway Toll Data Description**

In China, there are two types of freeway toll systems, Manual Toll Collection (MTC) and Electronic Toll Collection (ETC). For MTC, vehicles need to stop at each entry and exit toll station to take toll card and pay toll respectively no matter how many stations in the network it has passed through. For ETC, the vehicles can pass the toll station through ETC line without stopping and the toll collection is automatic. MTC and ETC toll data have same data structure and are centralized storage which makes them convenient to obtain and analyze.

The toll data used in this study come from Jiangsu freeway network during July 1st-28th 2012. There are 18,165,923 MTC data and 4,223,827 ETC data records before data quality control. Each data record contains information like the entry/exit time, entry/exit station, vehicle type, et al. The time-dependent exit volume of toll stations can be calculated by aggregating the toll record in predefined analyzing time interval. However, data quality control is necessary before traffic flow prediction.

**Data Quality Control**

The freeway toll data need to be cleaned using quality control technologies. The abnormal data may include data when a car enters and leaves at the same toll station, data with illogical time records, and data with an unreasonable speed.

1. **The Same Entering and Exiting Toll Station**
   U-turn in service area and toll card exchange to reduce fee can result in data that a car enters and leaves at the same toll station. During the process of extracting data, data with the same toll station were recognized and deleted.

2. **Data with Illogical Time Records**
   If the system time in different toll stations failed to synchronize, the entering time may be later than the leaving time. This kind of error can be found by simple logic sieve.

3. **Data with Abnormal Travel Speed**
   Data with abnormal travel speed is a main part of abnormal toll data. Long time stop at the service area, vehicle breaking down and aggressive driving behavior may result in abnormal travel speed. In this paper, the two-sided test method based on the standard deviation is applied to filter abnormal travel speed. When the gap between the observed data and the average travel speed is greater than $X$ times of the standard deviation, the record is considered as an abnormal one. The following are the detailed steps:
   - Step 1: Calculate the average travel speed between toll station $i$ and toll station $j$ in each 15min time interval.
Step 2: Calculate the standard deviation of travel speed in each time interval.  
Step 3: Calculate the gap between each travel speed and the average travel speed, and then compare the gap with the X times of the standard deviation (X is predefined). If the gap of one data record is larger than the criterion, it is regarded as abnormal data and will be deleted from the database.  
Step 4: Recalculate the average travel speed and standard deviation for each time interval after data filtering.  
Step 5: Repeat from Step 1 to Step 4 until there is no more abnormal data occurs.  

Take the travel speed of passenger car between Huaqiao toll station and Nanjing toll station on July 20th 2012 as an example. Set X=1.2 and after 5 iterations there were no more abnormal data. The final filtered results are shown in Fig.1. There was an serious accident happened around 14:30 and then all the vehicles suffered from a reduced capacity until 16:00. During this influenced time period, all the vehicles traveled with a lower speed. And the method introduced above can determine the available range according to the current traffic condition.

![Figure 1. Results of Abnormal Speed Data Filter Based on Standard Deviation.](image)

After data quality, 86.7% of toll data are available. Within the 13.3% abnormal data, 95.5% are abnormal travel speed, 2.8% are illogical time records, and 1.7% are same entering and exiting toll station.

**Temporal Distribution of Exiting Volume**

Freeway traffic flow changes periodically in Fig 2. The traffic volume increased from Monday to Thursday and reached a maximum on Friday, and then it began to decrease on Saturday and Sunday.
Traffic flow has similar characteristics on different day of week. According to the observed data (Fig 3 shows the data from July 1st to July 14th. Data from July 15th to July 28th is not shown here because of space limit), PM peak is larger than AM peak on Friday and Sunday while it is opposite on the other days. AM peak on Sunday is larger than Friday. In summary, days from Monday to Thursday have similar traffic patterns, while Friday, Saturday, and Sunday are different.

Figure 2. Daily Variation of Freeway Traffic.

Figure 3. Hourly Variation of Freeway Traffic
METHODOLOGY

SVM Forecast

Feature Vector Selection
Freeway traffic has a periodic variation. Traffic flow of the current period is related to the previous periods. Therefore, SVM model can be used to predict the traffic flow of future period. The exiting volume of freeway toll station \( m \) in current time period \( t \) is represented as \( V_{m(t)} \), and it is the output of SVM model. The exiting volume of the latest three previous time periods, \( V_{m(t-1)} \), \( V_{m(t-2)} \), \( V_{m(t-3)} \), are selected as the feature vector and they are the input of SVM model.

Historical Data Set
Freeway toll records are aggregated in 15min time interval for each toll station. The data from July 1st 2012 to July 21st 2012 consists the training data set, and the data from July 22nd to 28th 2012 are used as test data set. According to the temporal distribution of exiting volume, the training data set is divided into four categories: from Monday to Thursday, Friday, Saturday and Sunday. Considering that there are some differences in the morning and evening rush hour traffic, each category is further divided into three parts: morning peak hours (8:00 AM-10:00AM), evening peak hours (13:00 PM-16:00PM) and off-peak hours. In total, the training data set is divided into twelve groups. SVM model is built for each group.

Kernel Function
In the non-linear support vector machine regression analysis, the choice of kernel function directly affects the generalization ability of the model. Some common kernel functions used in literature include polynomial function, Sigmoid function and Gaussian radial basis function (Gaussian Radial Basis Function, RBF). According to previous studies, the Gaussian radial basis function corresponding to the feature space can make an infinite number of dimensions, all theoretically limited data samples in the feature space is certainly linearly separable [1][2][3] . Among the three functions, RBF function is the most widely-used kernel function, which is also used in this study. The formulation of RBF function is as follows:

\[
K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2}\right)
\]

Where \( \sigma \) is the parameter to be calibrated; \( x_i \) is the influence factors of the \( i \)th learning data; \( x \) is the influence factors of the whole learning data.

Calibration of Parameters
The SVM model with RBF function contains three parameters to be calibrated: \( C \), \( \epsilon \), and \( \sigma \). \( C \) is the penalty factor of SVM model; \( \epsilon \) is the insensitive loss parameter of SVM model; \( \sigma \) is the parameter of RBF function. \( \sigma \) determines the width of the local neighborhood. When \( \sigma \) is larger, the accuracy of SVM is higher, but there are more
support vectors, and the computational complexity is higher. \( C \) controls the degree of punishment beyond the allowable error range of samples. Larger value of \( C \) can improve the prediction accuracy of the results. But if \( C \) is too large, it will cause the value of \( C \) to be “over-fitting” state, which will reduce the accuracy of the test data set. These parameters directly affect the prediction accuracy of SVM nonlinear regression model. In this paper we use Kernel Alignment (referred to as the KA) [6] method to determine \( \sigma \) and then use cross-validation (Cross Validation, referred to as CV) method to determine optimal value of \( \sigma \) and \( C \).

### KNN Forecast

#### Feature Vector Selection

The selected feature vector of KNN is the same as SVM. When predicting the exiting of current time period, \( V_{m(t)} \), the exiting volume of the latest three previous time periods, \( V_{m(t-1)}, V_{m(t-2)}, \) and \( V_{m(t-3)} \), are selected as the feature vector.

#### Historical Data Set

The historical data set is the same as SVM model. The training data set is divided into twelve groups, and KNN model is established for each group.

#### Calibration of K value

Integer \( K \) is the only parameter needs to estimate in KNN algorithm. It affects the model prediction accuracy directly. Cross-validation method is applied to determine the optimal value of \( K \) for each training sample.

First, set the minimum and maximum \( K \) as \( K_{\min} \) and \( K_{\max} \), respectively. Usually, \( K_{\min} = 1, K_{\max} = 50 \). Then, each of the training data set is randomly divided into \( V \) parts, \( D_1, D_2, D_V \). And in turn the respective parts of the data set \( D_j \) (\( j = 1,2,\ldots, V \)) as a new test data set, the other \( V-1 \) parts of new data set as the training data set. Then when , calculate the mean absolute percentage error of the predicted data sets \( D_j \), and then calculate the average mean absolute percentage error of the \( V \) datasets. When obtains the minimum, the optimal value of \( K_i \) should be the optimal \( K \) value.

\[
M_{(K_i, D_j)} = \frac{100\%}{N} \times \sum_{n=1}^{N} \left| \frac{A_n - P_n}{A_n} \right|
\]

\[
\bar{M}_{K_i} = \frac{1}{V} \sum_{j=1}^{V} M_{(K_i, D_j)}
\]

Where \( N \) is the number of periods required for prediction; \( A_n \) and \( P_n \) are the true value and prediction value of the \( n \) th record respectively in test dataset \( D_j \) when \( K = K_i \).

#### Measured Method

We use Euclidean distance to measure the difference between the predicted values and the true values:
Where \( d(V_{P,m(t)}, V_{A,m(t)}) \) means the Euclidean distance between the predicted feature vector \( V_{P,m(t)} \) and the observed feature vector \( V_{A,m(t)} \); \( n \) is the dimension of feature vectors and \( n=3 \) in this study; \( N \) is the amount of historical data; \( x_{P(i)} \) and \( x_{A(i)} \) are the \( i^{th} \) components of feature vector \( V_{P,m(t)} \) and \( V_{A,m(t)} \) respectively.

**Locally Weighted Estimation**

After selecting the \( K \) nearest historical data records, the prediction value is calculated by a weighted estimation. When the Euclidean distance is smaller, the value should have a higher contribute to the prediction. The weighted estimation is shown as follows:

\[
V_{P,m(t)} = \sum_{i=1}^{k} w[V_{P,m(t)}, V_{A,m(i)}] V_{A,m(i)}
\]

(5)

\[
w[V_{P,m(t)}, V_{A,m(i)}] = \frac{\exp\left(d(V_{P,m(t)}, V_{A,m(i)})\right)}{\sum_{i=1}^{k} \exp\left(d(V_{P,m(t)}, V_{A,m(i)})\right)}
\]

(6)

Where: \( V_{P,m(t)} \) is the predicted value of \( m \) toll station at the time period \( t \); \( V_{A,m(i)} \) is the \( K \) values for the \( i \)-th history flow; \( w(V_{P,m(t)}, V_{A,m(i)}) \) is the weight determined by Euclidean distances of \( V_{P,m(t)} \) and \( V_{A,m(i)} \).

**Hybrid Model Forecasting**

Many factors affect the freeway exiting flow, and therefore a simple prediction method can only take into account a certain kind of factors from their own perspective. In order to improve the prediction accuracy, different methods can be tested and then choose the best one. However, each prediction method has its own applicable situation, and it is difficult to find an all-powerful prediction method for all traffic pattern. In order to take full advantage of all the useful information to get better prediction, Bates and Granger proposed Hybrid Model Forecasting concept: combine the individual prediction methods according to certain rules to improve the prediction accuracy.

Combination forecasting includes linear combination and nonlinear combinations. Linear combination include: simple average, the simple weighted average, weighted variance countdown, countdown weighted mean square, etc. Nonlinear combination of different models is to use the outcome of different models as input values and then get the final result by nonlinear combination forecasting function.

\[
\hat{y} = \sum_{i=1}^{M} w_i f_i
\]

(7)
\[ \hat{y} = (f_1, f_2, \ldots, f_M) \]  

Where \( \hat{y} \) is the combination of predicted values; \( f_i \) is the result of model \( i \); \( M \) is the number of models; \( w_i \) is the weight of linear combination; \( \phi \) is the combination function.

Linear combination model has simple model structure, high computational efficiency, good adaptability and is widely used in practical research. Nonlinear combined model can express the nonlinear traffic conditions more accurately, but has lower computational efficiency. Considering calculation efficiency, this study chose linear combination of SVM and KNN models. Because SVM and KNN models have different performance of prediction accuracy in different time periods, the weighted combination should be considered. In this study, the weighted hybrid prediction model of SVM and KNN is as follows:

\[ V_{PC(m,t)} = w_{S(m,t)} V_{PS(m,t)} + w_{K(m,t)} V_{PK(m,t)} \]  

Where \( V_{PC(m,t)} \) means the predicted value of toll station \( m \) in period \( t \) with hybrid model; \( V_{PS(m,t)} \) is the predicted value by SVM; \( V_{PK(m,t)} \) is the predicted value by KNN; \( w_{S(m,t)} \) and \( w_{K(m,t)} \) is the weighting factors of SVM and KNN respectively.

Weighting factors of current time period is calculated by mean absolute percentage error (MAPE) of the last three time periods (t-1, t-2, t-3), and the formula is calculated as follows:

\[ w_{S(m,t)} = \frac{\exp(E_{S\text{ MAPE}(m,t)})}{\exp(E_{S\text{ MAPE}(m,t)}) + \exp(E_{S\text{ MAPE}(m,t)})} \]  

\[ w_{K(m,t)} = 1 - w_{S(m,t)} \]  

\[ E_{S\text{ MAPE}(m,t)} = \frac{100\%}{3} \sum_{i=1}^{3} \left| \frac{V_{PS(m,t-i)} - V_{A(m,t-i)}}{V_{A(m,t-i)}} \right| \]  

\[ E_{K\text{ MAPE}(m,t)} = \frac{100\%}{3} \sum_{i=1}^{3} \left| \frac{V_{PK(m,t-i)} - V_{A(m,t-i)}}{V_{A(m,t-i)}} \right| \]  

Where \( E_{S\text{ MAPE}(m,t)} \) means the average percentage error of SVM prediction when comparing with true values in the last three time periods; \( E_{K\text{ MAPE}(m,t)} \) means the average percentage error of KNN prediction in the last three time periods before time period \( t \) in station \( m \); \( V_{PS(m,t-i)} \) is the SVM prediction value of time period \( t-i \) in station \( m \); \( V_{PK(m,t-i)} \) is the KNN prediction value of time period \( t-i \) in station \( m \).
Due to the fundamental principle of SVM and KNN, these two prediction algorithms have different performance. When the similar historical data can be found based on the feature vector of the prediction object, KNN will have a good performance. However, when the abnormal traffic condition exists and the similar historical data cannot be found, SVM can speculate the prediction based on the transferred linear regression and it always has a better performance than KNN. Therefore, the hybrid model about SVM and KNN can ensure the moderate prediction results for both normal and abnormal traffic conditions when using the weighting factors based on the previous prediction error.

RESULT
Scenario Description
Huning Freeway is the busiest freeway in Jiangsu Province. Huaqiao toll station and Nanjing toll station are the two busiest stations on Huning Freeway. Due to the large traffic, vehicle queues often appear in these two stations. Short-term prediction of exiting volume could indicate the administrators to get prepared in advance for possible queues and take corresponding measures, such as increasing number of exit ramps, opening a dedicated truck ramp and other measures to reduce waiting time of exit vehicles. Monday, July 23th 2012 is selected as the test day and then the historical dataset ranging from July 1st to 21th only contains the weekday data. The time interval is 15min and there are 96 intervals for the whole day. According to different time of day, the prediction models are built separately for morning peak hours (8:00 AM-10:00AM), evening peak hours (13:00 PM-16:00PM) and off-peak hours.

Results of Prediction
The model performance of SVM, KNN and the proposed hybrid model are compared to validate the proposed model. Based on the methodology discussed above, the parameters of SVM and KNN need to be calibrated first as shown in Tab.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Huaqiao Toll Station</th>
<th>Nanjing Toll Station</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AM</td>
<td>PM</td>
<td>Off Peak</td>
</tr>
<tr>
<td>KNN</td>
<td>K</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>SVM</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>C</td>
<td>1.208</td>
<td>1.103</td>
<td>0.845</td>
</tr>
</tbody>
</table>

The comparisons among SVM, KNN and hybrid model are shown in Fig. 4 and Fig. 5. The scatter diagram about the observations and model results are represented in Fig. 6 and Fig.7.
The model results can represent the time dependent exiting volume. The MAPE of the whole day is applied to represent the prediction accuracy. At Huaqiao toll station, the MAPE of KNN and SVM are 5.4% and 9.2% respectively. The MAPE of hybrid model is 6.1%. At Nanjing toll station, the MAPE of KNN and SVM are 8.1% and 5.2% respectively. The MAPE of hybrid model is 6.3%. SVM and KNN have different performances in these two toll stations. KNN has a better performance at Huaqiao while SVM is more suitable for Nanjing. Hybrid model can guarantee the prediction accuracy for both toll stations.

Figure 4. Prediction Result of Huaqiao Toll Station

Figure 5. Prediction Result of Nanjing Toll Station
Besides, the algorithm efficiency can satisfy the requirements of real-time application. The different average computation times of one analyzing interval are shown in Tab.2. Due to the combination of SVM and KNN, the hybrid model takes the longest computation time. The average computation times of hybrid model are 0.55s and 0.60s for Huaqiao and Nanjing respectively, which means it only takes less than 1s to predict the exiting volume in the next 15 minutes time interval.

Figure 6. Scatterplot of Hybrid Predictions and Observations of Huaqiao Toll Station.

Figure 7. Scatterplot of Hybrid Predictions and Observations of Nanjing Toll Station.
Tab 2. Algorithm Efficiency of Different Prediction Models.

<table>
<thead>
<tr>
<th>Toll Station</th>
<th>SVM (s)</th>
<th>KNN (s)</th>
<th>Hybrid Model (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huaqiao</td>
<td>0.51</td>
<td>0.03</td>
<td>0.55</td>
</tr>
<tr>
<td>Nanjing</td>
<td>0.55</td>
<td>0.03</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**Error Analysis**
The performance of KNN and SVM varies in different stations. KNN shows better results at Huaqiao toll station while SVM shows better performance at Nanjing toll station. This just proves the necessity to use hybrid model rather than a single KNN or SVM model.

Generally, the prediction results of the hybrid model are more stable than KNN and SVM. It is noteworthy that the prediction results of Nanjing toll stations during 13:00-16:00 are more inferior than the other periods. The main reason is the large fluctuation of traffic flow. The proposed method is a kind of machine learning. Although SVM can extend the prediction range based on the regression model, the model accuracy still depends on the consistence with the historical dataset. The feature vector of the proposed method is the exiting volume of the latest three time intervals, and the model accuracy would be poor when the volume of the last three periods fluctuates greatly.

**CONCLUSIONS**
This study proposed a short-term prediction algorithm of freeway exiting volume by the hybrid model of SVM and KNN. The model results can help the freeway administrators get ready in advance for possible queues and take corresponding measures at exiting toll station. Freeway toll data contains the detail trip information of each vehicle, and more than 20 million toll records of July 2012 has been used in this study.

The experimental results show that the proposed algorithm is feasible and accurate. The mean absolute percentage error is under 10% for both toll stations. When comparing with KNN and SVM, the results of hybrid model indicate that it can guarantee a more stable accurate level. The proposed method can apply in the on-line exiting volume prediction.

In the future research, the hybrid model can also be tested in the prediction of travel time and segment traffic volume. Another research direction is that choose the most suitable model (SVM or KNN) first based on the previous traffic pattern, and then only use the chosen model to realize the short-term prediction. This may lead to a higher accuracy, but has to cost more computing power.

**REFERENCES**


