Hybrid Traffic Speed Modeling and Prediction Using Real-world Data

Rong Zhang*, Yuanchao Shu*, Zequ Yang*, Peng Cheng* and Jiming Chen*
*State Key Laboratory of Industrial Control Technology
Zhejiang University, Hangzhou, China
zhangrong@iipc.zju.edu.cn, ycshu@zju.edu.cn, zqyangzju@gmail.com, {pcheng, jmchen}@iipc.zju.edu.cn

***Abstract*—Traffic speed modeling and prediction is of great importance for both individuals and government authorities due to the increasing number of traffic congestion and corresponding social and economic impacts. Various approaches have been proposed to predict traffic speed. However, the long-term prediction accuracy is still unsatisfactory especially when occasional events such as extreme weather conditions occur. Based on over 880,000 traces of taxicab GPS collected in Hangzhou, China, as well as the dataset of weather conditions and holidays, we demonstrate a multi-time-scale correlation of traffic speed and the effects of various related events. We propose a hybrid traffic speed modeling and prediction framework which takes multi-time-scale historical traffic speed data as well as related events as inputs. For all segments of major roads in Hangzhou, we establish corresponding traffic speed models through a recursive model identification algorithm. We validate the effectiveness of our approach under various conditions through extensive trace-driven simulations.**

I. Introduction

Traffic congestion is becoming more serious along with the social and economic development. Nowadays, GPS is widely equipped by taxis and buses which provides information of both location and speed. Recently many research studies have been done to analyze the traffic conditions and even predict the future traffic speed by utilizing GPS data generated by a large number of vehicles over a period of time, which is of great importance for many applications such as easing traffic congestion, route guidance, reducing pollutant emissions and so on [1] [2].

Auto-Regressive Integrated Moving Average (ARIMA) [3] and Exponential Smoothing (ES) [4] have been widely used for short-term traffic speed prediction (usually less than 1-hour interval). With the development of artificial intelligence and data mining technology, Artificial Neural Network (ANN) and Support Vector Regression (SVR) have also been proposed for traffic prediction [5]. Although different approaches have been proposed to predict the traffic speed, the long-term prediction accuracy is still unsatisfactory especially when some closely related external events happen, for example, extreme weather conditions, road regulation by the goverment authorities, etc.

Motivated by these observations, we devise a novel system identification method and propose a hybrid traffic speed model using both a city-scale taxicab GPS dataset and related external events datasets.

With the System Identification (SI) process, we are able to obtain the model order and parameters which reveal the hidden dynamic correlation among inputs (short-term and long-term historical traffic speed, weather, special day and so on) and outputs (the predicted traffic speed). The benefits of employing hybrid traffic prediction method are threefold. First, we are able to consider both the correlation of the predicted traffic speed and the historical traffic speed in terms of both short-term and long-term time scales. Second, by taking into account important external factors, the established model reveals the underlying dynamic correlation between the road traffic and the occasional events, such as heavy rain, road construction and road restriction. Last but not least, different from many machine learning methods, parameters in such identified model have more clear physical meanings which help us better understand the traffic dynamics.

Intellectual contributions of this paper are summarized as follows:

1) We collect taxicab GPS data and other events data, including weather condition, special day, etc., for over 4 months and construct a dataset including more than 880,000 traces in Hangzhou, China. Through time series analysis, we show the multi-time-scale correlations of traffic speed, and important effects caused by various external events.
2) Based on System Identification (SI), we propose a hybrid traffic speed modeling framework which takes multi-time-scale historical traffic speed data as well as external events as inputs. We further design a systematic recursive model identification algorithm to derive the model order and parameters. To the best of our knowledge, this is the first attempt to model traffic speed using system identification method.
3) We build hybrid traffic prediction models for major road segments in Hangzhou respectively, and evaluate the traffic prediction performance based on our collected dataset under various settings. Our hybrid modeling approach achieves satisfactory traffic prediction accuracy for even 10-hour prediction. The 1-hour prediction accuracy under extreme weather conditions can be improved by more than 40%. Meanwhile, the performance improvement compared with the benchmark SVR method is more than 13.5%.

The rest of the paper is organized as follows. In Section II, we briefly describe the dataset used in our work and introduce the system architecture. In Section III, we analyze the correlation of traffic data. Then we introduce how to use SI to model the traffic speed in Section IV. We evaluate the proposal method in Section V. Related works are discussed in Section VI, and Section VII concludes the paper.

II. Preliminary

In this section, we first introduce the real-world traffic dataset collected from taxi-equipped GPS, and describe the dataset of external factors including weather and holidays.

A. Dataset Description

Hangzhou is one of the most renowned and prosperous city in China with area of 16000 square kilometers and a population of 8.8 million. In addition to fare meters, taxicabs in Hangzhou are equipped with GPS. Thus, the physical status of taxicab can be monitored for regulation purposes and enhances the quality of service [6]. Meanwhile, it also provides opportunities for the researchers to address the traffic issues. To this end, we make use of traffic dataset of the taxicabs in Hangzhou. Additionally traffic conditon is also severely affected by some other external factors, such as weather, and public events. Therefore, in this paper, we take the weather condition dataset and special day dataset into consideration along with the taxicab data. The detailed data description can be found in Table I.

1) Traffic Dataset: The traffic dataset is collected from Oct 1st, 2013 to Jan 31st, 2014. It is composed with 8000 taxicabs' every-2-minute GPS data in Hangzhou, where each record represents corresponding position with latitude and longitude coordinates, in addition to the instantaneous speed and the driving direction. Besides, the road network of Hangzhou is approximately divided into 1325 road segments. The records of our traffic dataset do not contain the information of taxicabs' unique ID, for the sake of privacy protection for taxi drivers. Without loss of generality, we select 500 road segments for thorough analysis. In this paper, the detailed description of this dataset is shown in Table I.

2) Weather Condition: Moreover, the dataset of weather information ranging from Oct 1st, 2013 to Jan 31st, 2014 is obtained from Chinese Weather Report Net [7] including precipitation, temperature, wind speed, etc. The online weather information updates every 60 minutes.

3) Holiday: It is intuitive that the condition of traffic flow varies remarkably on weekday or weekend, owing to different human behavior. Apart from that, notice that there are several public holidays during that period of time, namely National Day (Oct 1st'13 - Oct 7th'13), New Year's Day (Jan 1st'14 - Jan 3rd'14) and Spring Festival (Jan 31st'14), which perhaps contribute to another pattern of traffic flow different from on weekday or weekend and need to be considered.

B. Raw Data Processing

We describe the raw data processing in this part. As a result, some essential notions are presented for further quantitative analysis.

Table I
Dataset Summary

Dataset		Scale	Duration
Traffic data	road segment ID	885600	2013.10.1-2014.1.31
	speed value		
	direction		
	time		
Weather	weather condition	2952	
	time		
Holiday	name	123	
	time		

1) Average Speed: We attempt to estimate the dynamics of traffic flow by leveraging the taxicabs' GPS data. Toward this end, we mainly pay attention to the speed of one road segment, and calculate the average speed of all the GPS records occurring on that segment during a certain period of time.

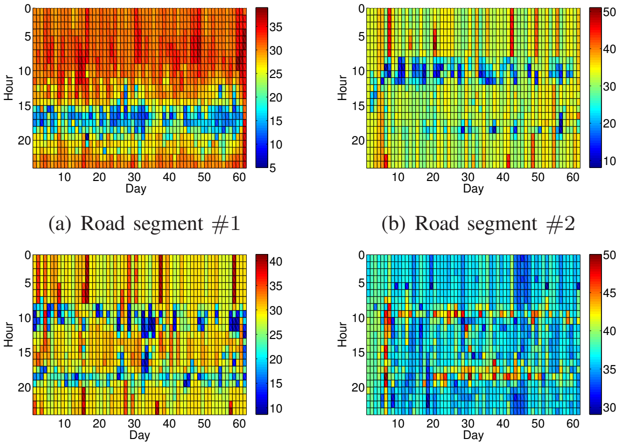
Moreover, given the limited number of taxicabs and large-scale road networks, the GPS records in a certain segment are sparse. Therefore, we divide time into segments and consider GPS data within one hour. It basically assures there exists at least one record in each segment. For example, for road segment No.12, the average speed from 8am to 9am on Oct 1st is the mean value of 20 GPS records' speed value. For simplicity, it denotes the average speed at 8am.

2) Weather Condition: Note that the typical weather conditions include sunny, cloudy, rain, snow, rainstorm, thunderstorm, fog, frost, and so on. Therefore, we categorize all the states into 3 kinds, depending on the precipitation. This is because since the pavement humidity has a direct influence on people's driving behavior and then changes the speed of the traffic flow. Concretely, the first kind corresponds to the basically drying ground, e.g. sunny, cloudy, overcast, flurry spit sprinkle and fog; the second kind features with moderate humidity, e.g. light rain, moderate rain, light snow, frost and sleet; The third kind, with high humidity, contains heavy rain, heavy snow, rainstorm and thunderstorm. Those 3 kinds of states are labeled by w_1, w_2 and w_3, respectively. In our dataset, the occurrence frequency of w_1, w_2, w_3, is 76%, 17%, 7% respectively.

3) Special Day: The traffic flow on different kinds of days are observed with distinct patterns. For example, rush hours happens earlier in morning of weekday, while people are prone to travel around during the public holidays and weekends. Inspired by those special days, we divide days into three catagories, where weekdays, weekends and public holidays are labeled by SD_1, SD_2, SD_3, respectively. The occurrence frequencies of these three kinds of days are 68%, 25%, 7% respectively.

III. Motivation

In this section, we justify the motivation of leveraging the inherent property and external events to predict the traffic speed. Through time series analysis and visualization, we present several insights of traffic data at the multiple time scales, and analyze the correlation between the traffic speed and random factors, specifically the weather condition and special day.



(a) Road segment #1

(b) Road segment #2

(c) Road segment #3

(d) Road segment #4

Figure 1. Different traffic states of different roads

A. Multi-Time-Scale Traffic Speed Correlation

As a sequence of continuous samples, we consider traffic data as a time series data [8], and analyze the internal relations of traffic data using methodology in time series analysis. Time series analysis is a statistical method for dynamic data processing. It has been used to study the statistical regularity of random data sequence with time continuity. We can use it to extract meaningful patterns and characteristics of the data sequence.

Note that the data of hourly average speed for each road segment is obtained. We choose the 4 road segments' data from Dec 1st, 2013 to Jan 31st, 2014 to observe the varies patterns, where Figure 1(a), 1(b), 1(c) and 1(d) indicate the scenario of evening peak, morning peak, morning and evening peak, without morning or evening peak, respectively. It demonstrates the spatial dynamics of traffic flow, as the different road segments have different traffic patterns. Besides, it is observed that each segment owns the temporal dynamics and shows a similar periodic pattern. Specifically, the morning peak or evening peak appears on weekdays while another traffic phenomenon occurs on Jan 31st, Spring Festival in China.

We attempt to investigate whether the current traffic speed correlates with the historical data. Such analysis is important as it is not obvious whether the traffic speed has multi-time-scale correlations. If the answer is positive, our traffic speed model has to take such correlations into consideration in order to achieve more accurate prediction. In this paper we adopt the correlation coefficient method to analyze the correlation of the traffic speed at any adjacent hours.

To assess the correlation between the current traffic speed and the historical data, we introduce the coefficient of correlation, which is shown as follows,

$$R = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 \times \sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}, |R| \le 1, \qquad (1)$$

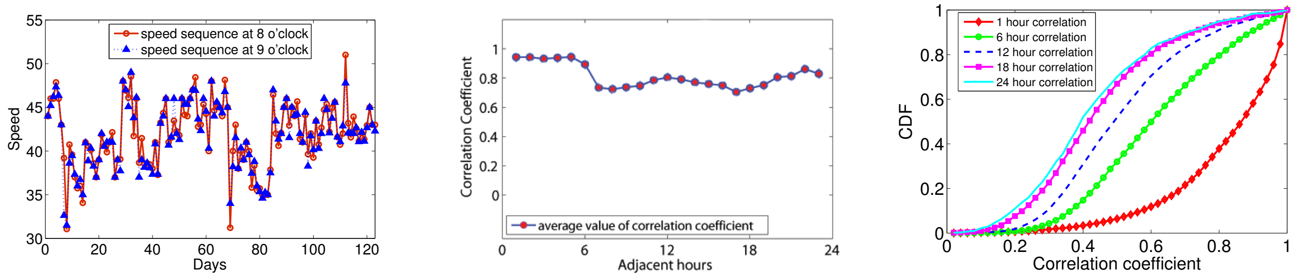
where the sequence of x and y denote one certain road segment's current traffic speed and the past traffic speed respectively. For instance, when evaluating the correlation between traffic speed at 9:00 and its 1-hour-previous historical traffic speed, the sequence of x and y denote the traffic speed at 9:00 and 8:00 respectively, on every single day. The correlation is positive if R is greater than 0, otherwise it is negative. In addition, the correlation gets stronger when R is closer to 1, and vice versa.

Taking the traffic speed at 9:00 and 8:00 as an example (shown as Figure 2(a)), the correlation coefficient $R = 0.9591$, which means the connection between these two sets of traffic speed is strong. Furthermore, we repeat the same analysis on all 500 road segments. Figure 2(b) illustrates the 500-road-segment's average correlation coefficient between every moment's traffic speed and its 1-hour-interval historical speed. It plots that all of the average correlation coefficient are greater than 0.5, which indicates a strong correlation. We further analyze the correlation between traffic speed and its 6,12,18,24-hour-previous historical speed in Figure 2(c). It is obvious that the correlation becomes stronger as the hourly interval is shortened. Apart from that, the correlation is weakened with increasing time intervals. However, the correlation coefficient distribution becomes closer and maintains a certain level, illustrated by curves with the 18-hour and 24-hour intervals. Therefore, we consider the historical speed in multi-time-scale detailed in the hourly-interval and daily-interval historical traffic data to predict the traffic speed.

From Figure 2(c), when the time interval becomes larger, such as 24-hour interval, the correlation becomes less obvious. This may be understandable in the intuitive sense, however, it should also be noted that when the time interval becomes larger, it becomes more often that the correlation coefficient is calculated between weekday traffic speed (i.e., Monday) and weekend traffic speed (i.e., Sunday). Such increased chances in fact pollute the calculated correlation coefficient less positive as we will show in later part that the traffic speed patterns during weekdays and weekends are fairly different. Therefore, in order to better utilize the multi-time-scale traffic speed correlation, we should take the external factors such as special days account.

B. Impact of Random Factors

As analyzed above, the historical data has strong influence on future traffic. Furthermore, the current traffic speed will also suffer the interference of some environmental factors, such as weather, special days, road restrictions and so on [9]. In this part, we analyze the correlation between these factors, including weather condition and special day, and the traffic



(a) Traffic speed at 9:00 and 8:00 of one road segment

(b) Correlation between every two subsequent hours' traffic speed in 500 road segments

(c) Cumulative distribution of correlation coefficient of 500 road segments in different hour intervals

Figure 2. Large scale correlation analysis for different hourly intervals

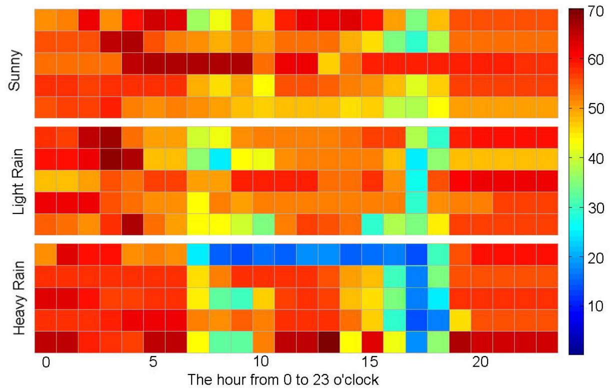


Figure 3. The traffic state in different weather conditions

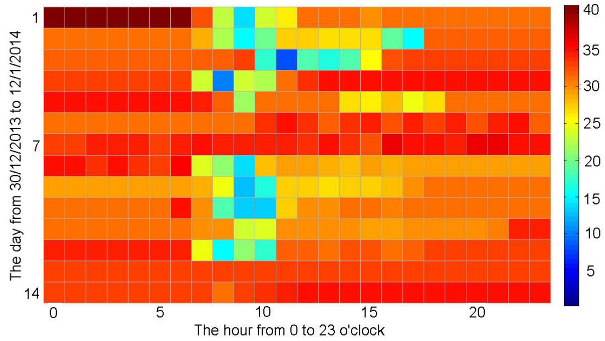


Figure 4. The traffic state of two weeks

speed. Such analysis is beneficial for us to further establish the mathematical model for speed prediction in later sections.

1) Weather Condition: We extract 5 days respectively in sunny, light rain and heavy rain weather conditions. The speed distribution of these days are shown in Figure 3. We find the light rain aggravates the traffic congestion at evening peak while heavy rain is more serious. So the influence of weather on the traffic can not be ignored.

2) Special Day: As metioned in Section II-B3, special days are divided into weekdays, weekends and holidays. We can know the traffic pattern on weekdays, weekends and holidays is different from Figure 1 and Figure 4. It will appear to be crowd at rush hour at weekdays but people don't have this behavior at weekends. And people will be influenced by a large number of tourists and perhaps have another transportation plan on holidays. Therefore, this factor also should be taken into consideration.

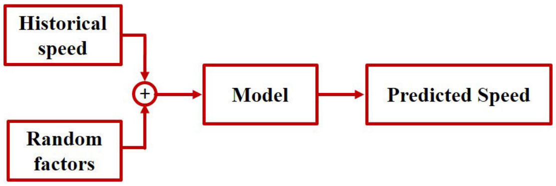


Figure 5. Illustration of System Identification's main idea

C. Problem of Interest

We seek for designing a System-Identification-based model for traffic speed prediction, which updates in real time for each road segment so as to capture the temporal and spatial dynamics. In this part, we discuss about the main idea of System Identification method as shown in Figure 5.

In accordance with the strong correlation between the current traffic speed and multi-time-scale traffic data, as well as the random factors, our input data includes historical traffic data, weather condition and special day. In this paper, we to propose a systematic and efficient approach to establish such a dynamic model through system identification method. The model is expected to characterize the underlying complicated temporal and spatial dynamics of traffic speed of the whole city, and further provides a more accurate traffic speed prediction accuracy at different time scales.

IV. System Identification

Inspired by its concept, we model the traffic flow by leveraging the existing traffic dataset and other external factors. The previous traffic data and external factors correspond to the input while the current traffic speed is the output.

Firstly, we take an overview of analyzing process of System Identification, where two key steps, structure constructing and parameter identification, need to be highlighted. On one hand, in structure constructing, it is determined that the order of system which indicates the current event correlating to the past events happening how long before. Specifically, we utilize a simple but effective method to determine the system's order after trial and error. On the other hand, in the parameter

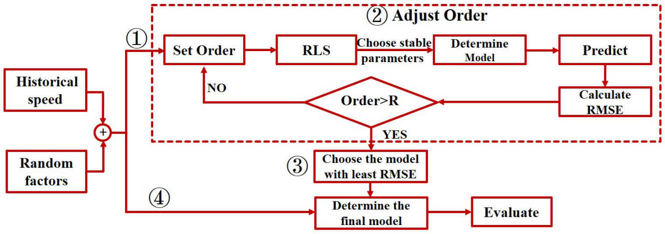


Figure 6. Flow chart of System Identification

identification, the past traffic data as well as random factors, which include weather condition and special day in our case, should be taken into account. The system parameters demonstrate how much the past traffic data and other factors influence the predicted traffic speed.

A. The Physical Meaning of System Structure

Hereby, we will introduce the basic structure of system model in the process of SI. The system's input and output at time k are denoted by $u(k)$ and $y(k)$ respectively. The most basic linear equation of the input and output in discrete time shows as follows [10],

$$y(k)+a_1y(k-1)+\cdots+a_ny(k-n_a)=b_1u(k-1)+\cdots+b_mu(k-n_m),$$

which can be naturally transformed into

$$\begin{aligned} y(k) \;\; = \;\; & -a_1y(k-1)-\cdots-a_ny(k-n_a) \\ & +b_1u(k-1)+\cdots+b_mu(k-n_m). \end{aligned} \tag{2}$$

Then we introduce the input vector $\varphi(k)$ and parameter vector θ, where

$$\varphi(k)=[-y(k-1),\cdots,-y(k-n_a),u(k-1),\cdots,u(k-n_m)]^T,$$
$$\theta=[a_1,\cdots,a_n,b_1,\cdots,b_m]^T.$$

In our case, it is realized that the predicted speed is related to not only the hourly-interval and daily-interval history speed but also the current random factors. Therefore in our case, the hourly-interval history speed is denoted by $HHS(k)$, and daily-interval one is denoted by $DHS(k)$. $u(k)$ is the random factors which also affect the predicted speed. Especially, the weather condition, WC, and special day, SD. Consequently, the system structure of predicting the current speed can be described as below,

$$\begin{aligned} PS(k) = & -a_1HHS(k-1)-\cdots-a_{n_a}HHS(k-n_a) \\ & -b_1DHS(k-1)-\cdots-b_{n_b}DHS(k-n_b) \\ & +c_1WC(k)+c_2SD(k). \end{aligned} \tag{3}$$

where $PS(k)$ represents the predicted traffic speed.

B. Online Identification of System Model

In addition to the input data, the process of modeling mainly consists of two parts, while one is identifying the order of model and the other is adjusting the model parameters. The whole idea of determining the order and parameters is to iteratively adjust them until the accuracy of model is maximized, which is illustrated in Figure 6 in detail.

According to the basic structure, the crucial order and parameters of system need to be determined by combining history speed and random factors (shown in Step 1 in Figure 6). Most importantly, we are still unaware of the history speed in hourly intervals and that in daily intervals happening how long before respectively correlate the current speed, i.e. n_a and n_b are unknown. In order to identify the system order, in Step 2 in Figure 6, we begin with the initialization where $n_a = 1$ and $n_b = 1$. Given each setting of n_a and n_b, the parameters of system can be estimated via mining the training data, such that a temporary system is obtained. Then by comparing its prediction result with the test data, we acquire the Root Mean Square Error (RMSE) of the predicting result, which indicates the accuracy of the temporary system. After searching for the minimum RMSE iteratively by increasing n_a and n_b respectively until n_a and n_b both reaching the pre-set thresholds, the optimum of system's order and parameters are determined in Step 3. Finally, the performance evaluation is conducted in Step 4.

In addition, note that Recursive Least Square (RLS) method is adopted in Step 2 for the parameter estimating [11], which helps to reduce the computational complexity, as RLS avoids the repetitive computation once the real-time traffic data is updated. Moreover, the online optimization approach can further improve the performance of parameter estimation [12]. Such a method is also expected to be useful for real-time traffic speed prediction. Therefore, we adopt the recursive least square method to estimate the parameters.

According to the principle of least square method, we have

$$\hat{\theta} = (\varphi^T\varphi)^{-1}\varphi^TY.$$

Then it is turned into the recursive form and shown as

$$\hat{\theta}(k) = \hat{\theta}(k-1) + K(k)[y(k) - \varphi^T(k)\hat{\theta}(k-1)], \tag{4a}$$

$$K(k) = \frac{P(k-1)\varphi(k)}{1+\varphi^T(k)P(k-1)\varphi(k)}, \tag{4b}$$

$$P(k) = [I - K(k)\varphi^T(k)]P(k-1). \tag{4c}$$

Overall, the detailed process of system identification is illustrated in Algorithm.1. With the initialized order of hourly-interval and daily-interval historical data, we use the RLS method to obtain the parameters and choose this set of stable parameters as the system parameters, to obtain a temporary system model. To justify this temporary system model, we train the model with the test data which have not been used, so that RMSE of the predicting result is obtained. Then by gradually increasing n_a and n_b and comparing the corresponding RMSE, those procedures are conducted iteratively until the threshold of n_a and n_b are triggered. As a result, the system's order and parameters with the minimum RMSE is considered as the final determined model.

Algorithm 1 System Identification

Input: Hourly-interval Historical Speed (HHS), Daily-interval Historical Speed (DHS), Weather Condition (WC), Special Day (SD).

Output: Predicted speed.

Initialization. θ: the set of parameters, TD: the range of training data, PD: the range of prediction data, n_a: the order for HHS, n_b: the order for DHS.

```
for $i = 1\, to\, n_a$, $j = 1\, to\, n_b$ do
    for $t = 1\, to\, TD$ do
        $\theta \leftarrow RLS(HHS, DHS, WC, SD)$
    end for
    choose a stable $\theta$
    for $p = 1\, to\, PD$ do
        $Speed(i, j) \leftarrow SI(i, j, \theta)$
    end for
    calculate $RMSE(Speed(i, j))$
end for
Get relevant value $(i, j)$ with $\min RMSE(Speed(i, j))$
    and then determine the order $n_a = i, n_b = j$
Determine the model $SI(n_a, n_b, \theta)$
```

V. Evaluation

A. Settings

In our case, the data of four months from Oct 1st, 2013 to Jan 31st, 2014 is chosen to establish and evaluate our identified model. Every single road segment owns altogether 2952 records. Among them, the previous 2400 records are leveraged for training, and the rest 500 ones are for testing. From all the 1325 road segments, we randomly select 300 road segments to model their traffic flow respectively and conduct our evaluation accordingly.

B. Performance Metrics

In order to justify the prediction accuracy, we calculate the Mean Absolute Error (MAE) for s_i at time t_j for the kth prediction horizon $MAE(s_i, t_j, k)$ as follows,

$$MAE(s_i, t_j, k) = \frac{1}{n} \sum_{j=1}^{n} |\hat{z}_k(s_i, t_j) - z(s_i, t_j)|.$$

Additionally, we also adopt the Route Mean Square Error (RMSE) for s_i at time t_j for the kth prediction horizon $RMSE(s_i, t_j, k)$ defined as

$$RMSE(s_i, t_j, k) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(\hat{z}_k(s_i, t_j) - z(s_i, t_j)\right)^2}.$$

C. Baselines: Support Vector Regression

Compared with the existing traffic prediction methods, our System Identification method outperforms in both long-term and short-term prediction by combining multi-time-scale data. Support Vector Regression (SVR), a typical prediction method suitable for short-term prediction, is widely used in prediction due to its high accuracy presented in [5]. In this paper, we adopt the SVR method as the baseline. The input feature vector $x_j \in R^n$ at time t_j for road segment s_i is $x_j = [HS(s_i, t_j - m), HS(s_i, t_j - m + 1), ..., HS(s_i, t_j), WC(s_i, t_j), SD(s_i, t_j)]$. Feature vector x_j contains historical speed, weather, special day. They are denoted by HS, WC and SD respectively. Let $y_j = z(s_i, t_j + 1)$ be the predicted speed at time $t_j + 1$. Our purpose is to find out the internal nonlinear relationships between y_j and x_j, while the training data is $\{x_j, y_j\}_{j=1}^{n}$. And SVR method is adopted to predict the traffic speed according to [5]. Specifically, we utilize Matlab package LIBSVM to realize SVR algorithm [13], where the input data is the same as that of SI method, i.e. historical speed, weather condition and special day.

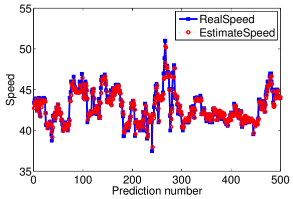


Figure 7. Comparison between one-hour prediction results and real traffic speed for one road segment

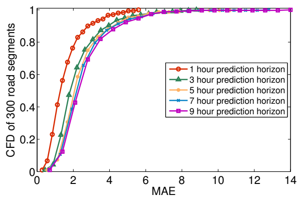
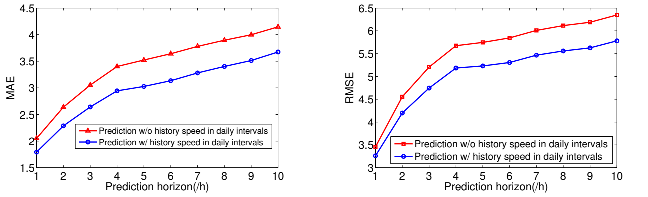


Figure 8. The cumulative distribution of MAE in different prediction horizon

D. Results

1) Performance of Traffic Speed Prediction: In this subpart, we evaluate the effectiveness of System Identification method based on the data of historical speed and random factors. Specially, we utilize the model of a certain road segment trained by 2400 sets of training data to predict the future traffic speed, while the other 500 sets of data are adopted to test the prediction performance. The estimated result is compared with the real one, which is shown in Figure 7. Notice that this result is based on 1 hour prediction horizon, which means we should be aware of the real speed value of past 1 hour as well as the earlier data. Furthermore, for the random selected 300 road segments, we calculate the MAE of each segment for different prediction horizon and draw the cumulative distribution in Figure 8. Especailly, we evaluate the prediction from 1-hour to 9-hour prediction horizons. Figure 8 plots the five different hourly prediction horizon results among them and illustrates that the MAE increases with the prediction horizon increases. Moreover, it is also shown that the cumulative distribution of MAE less than 3 kmph is 0.99, and 0.9867, 0.9767, 0.9500, 0.9167 are for 3-, 5-, 7-, 9-hour intervals respectively.

2) Impact of Multiple Time-scale: In order to show that our prediction model performs better when the long-time-scale historical traffic speed is taken into consideration, we further re-establish the traffic prediction model merely according to hourly-interval history speed. Note that in this subpart, we still do not consider the other factors, such as weather condition and special day. Figure 9 compares the effectiveness of results with/without daily-interval history speed. Obviously, it indeed improves the prediction accuracy, especially for long prediction horizon. For 1-hour to 5-hour prediction horizon, MAE (RMSE) considering daily-interval history speed is 12% (9%) fewer than that without these data, and they tend to be smaller as the prediction horizon increases. As for 5-hour to 10-hour prediction horizon, the difference of these two situations becomes steady.



(a) MAE comparison in different prediction horizons

(b) RMSE comparison in different prediction horizons

Figure 9. Comparison between whether considering historical speed in daily intervals

3) Impact of Random Factors: Different from many existing works [14], our model also includes various random factors into consideration. In this subpart, we verify the effectiveness of the model when taking such random factors into consideration. Preliminary, we try to verify the effectiveness of weather. Recalling Figure 3, a record with a serious evening peak when there happens to be a heavy rain. In order to justify the effectiveness, we compare the prediction accuracy of two prediction models where the only difference is that one model takes the weather condition as its input while the other does not. Then we verify the accuracy through testing data. Specifically, we choose the testing data during the day when there was a heavy rain. Figure 10 displays the true traffic speed of a certain road segment, as well as the absolute error of prediction with/without the weather condition. It is obvious that the prediction absolute error with weather information is 45% fewer than that without weather speed. Most importantly, the absolutely error with the weather condition is remarkable lowered by 63% when a heavy rain occurs from 18:00 to 20:00, which means the model with weather condition is able to capture the bursting traffic jam caused by the abrupt change of weather.

Furthermore, we verify the prediction performance when the factor of special day is taken into the model. Figure 11 compares the prediction performance with/without considering Special Day. Specifically, we select one road segment featuring a morning peak at weekdays except weekends. We mainly conduct the comparison at rush hour 8-11 am of weekends as the difference of weekdays and weekends lies in the morning peak. Figure 11 shows the comparison results of three weekends. It is obvious that considering Special Day has a smaller absolute error than no consideration. The prediction accuracy is improved by 5.4% in terms of the mean absolute error.

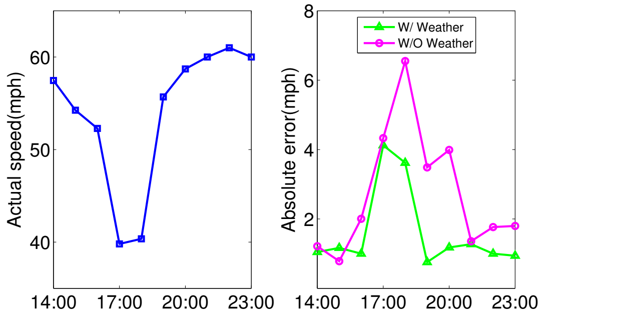


Figure 10. The prediction result comparison between whether considering the weather

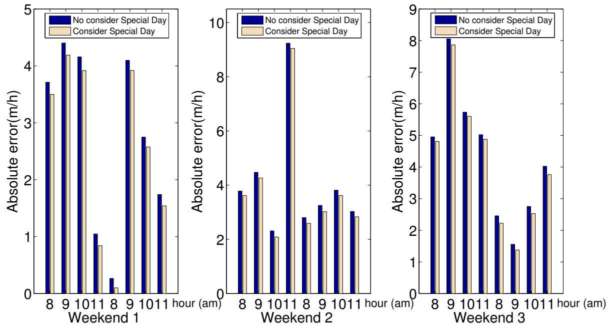
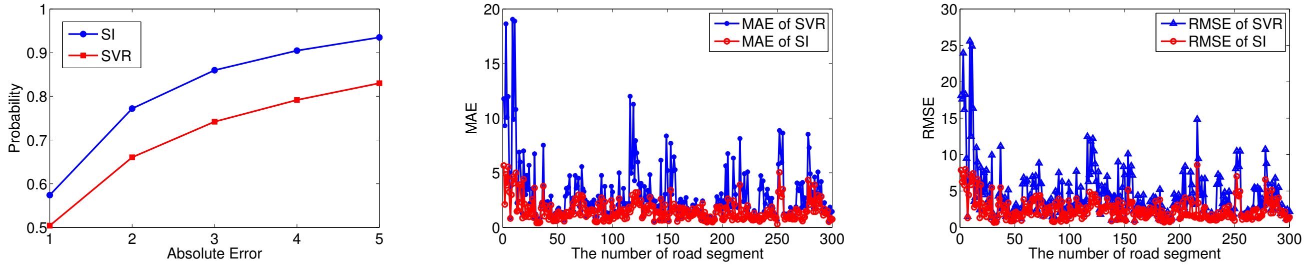


Figure 11. The prediction result comparison between whether considering the special day

4) Comparison with SVR: In this part, we compare our established prediction model with the benchmark SVR method. Both methods use the same training and test datasets. From Figure 12(a), it can be observed that our model fully outperforms SVR in terms of all the prediction accuracy. Specifically, it is observed the probability of MAE less than 1 kmph is 0.5720, as well as 0.7720, 0.8598, 0.9047, 0.9349 for less than 2, 3, 4 and 5 kmph respectively. As for SVR, the probability is 0.5092, 0.6671, 0.7494, 0.7996, 0.8385 respectively for 1, 2, 3, 4 and 5 kmph. Therefore by using our proposed method, the prediction accuracy is overall improved by 13.5% in average. Furthermore, Figure 12(b) compares the results of MAE for these two methods, while Figure 12(c) corresponds to RMSE. Obviously, the prediction error of SI is 58% (64%) fewer than that of SVR with regard to MAE (RMSE). Overall, SI is a dynamic prediction method and more accurate than SVR, an off-the-shelf machine-learning prediction method.

VI. Related Work

The issue of traffic prediction has drawn lots of researchers' attention. Most existing methods focus on predicting traffic via historical traffic data, but ignore the impact resulted from external factors such as weather, special events, point of interests and so on. In [15], B. Pan et al. propose a hybrid forecasting model merely considering the current situation, e.g. rush hour, which combines ARIMA model and Historical Average Model (HAM). However, they do not take more external factors into account, since the process of modeling features high computation complexity given the numerous possible circumstances. Y. Zhang et al. improve the prediction accuracy



(a) The accuracy comparison between SI and SVR in different scope of MAE

(b) MAE comparison between SI and SVR

(c) RMSE comparison between SI and SVR

Figure 12. The results comparison between SI and SVR

and reliability by modeling ARIMA and GJR-GARCH separately [16]. They uncover the underlying nature of periodic and volatility characteristics, though, the model only works in short-term-prediction application scenarios. Moreover, J. Xu et al. create a hybrid predictor aiming at the real-time traffic estimation [17]. Although it adapts to the change of several external situations, it fails to mine the historical data sufficiently. In long-term prediction, its accuracy reduces due to the serious influence of external factors. In short, to the best of our knowledge, this paper is the first to establish the dynamic traffic speed prediction by considering both the short-term and long-term historical data as well as various external factors as the model input. Such model is expected to reveal the inherent traffic dynamics of each road segment, and may be helpful for further analyze of the specific traffic patterns in accordance with the human behavior and the external factors.

VII. Conclusion

In this paper, we fully examine characteristics of traffic speed data and propose a dynamic traffic speed modeling and prediction framework, which combines multi-time-scale historical data and other random factors. The proposed model improves both the short-term and long-term prediction performance. Via extensive trace-driven simulations, our method is validated to outperform benchmarks in terms of the prediction accuracy. In addition, the model also adapt to burst of traffic speed caused by random factors. The results provide insights into traffic speed prediction and may help tackle a series of social problems caused by traffic congestion.

Acknowledgment

The paper was partially supported by the National Basic Research Program (973 Program) under Grant 2015CB352500, and National Program for Special Support of Top Notch Young Professionals.

References

[1] X. Tan, Y. Shu, X. Lu, P. Cheng, and J. Chen, “Characterizing and modeling package dynamics in express shipping service network,” in *Proceedings of the IEEE International Conference on Big Data*, 2014, pp. 144–151.

[2] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, “Inferring gas consumption and pollution emission of vehicles throughout a city,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1027–1036.

[3] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

[4] B. M. Williams, P. K. Durvasula, and D. E. Brown, “Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, no. 1, pp. 132–141, 1998.

[5] M. T. Asif, J. Dauwels, C. Y. Goh, A. Oran, E. Fathi, M. Xu, M. M. Dhanya, N. Mitrovic, and P. Jaillet, “Spatiotemporal patterns in large-scale traffic speed prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 794–804, 2014.

[6] D. Zhang, T. He, Y. Liu, and J. A. Stankovic, “Callcab: A unified recommendation system for carpooling and regular taxicab services,” in *Proceedings of the IEEE International Conference on Big Data*, 2013, pp. 439–447.

[7] T. P. W. S. C. of CMA, “Weather of china,” http://www.weather.com.cn/.

[8] W. Min and L. Wynter, “Real-time road traffic prediction with spatio-temporal correlations,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[9] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, “Discovering spatio-temporal causal interactions in traffic data streams,” in *Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining*, 2011, pp. 1010–1018.

[10] T. B. Schön, A. Wills, and B. Ninness, “System identification of nonlinear state-space models,” *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.

[11] H. Zhou, J. Chen, J. Fan, Y. Du, and S. K. Das, “Consub: incentive-based content subscribing in selfish opportunistic mobile networks,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 669–679, 2013.

[12] X. Cao, P. Cheng, J. Chen, and Y. Sun, “An online optimization approach for control and communication codesign in networked cyber-physical systems,” *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 439–450, 2013.

[13] C. Chang and C. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[14] S. Clark, “Traffic prediction using multivariate nonparametric regression,” *Journal of transportation engineering*, vol. 129, no. 2, pp. 161–168, 2003.

[15] B. Pan, U. Demiryurek, and C. Shahabi, “Utilizing real-world transportation data for accurate traffic prediction.” in *Proceedings of the IEEE International Conference on Data Mining*, 2012, pp. 595–604.

[16] Y. Zhang, Y. Zhang, and A. Haghani, “A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 65–78, 2014.

[17] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. van der Schaar, “Mining the situation: Spatiotemporal traffic prediction with big data,” *IEEE Journal of Selected Topics in Signal Process*, 2015.