Effective Traffic Flow Forecasting Using Taxi and Weather Data

Xiujuan Xu[1,2], Benzhe Su[1,2], Xiaowei Zhao[1,2], Zhenzhen Xu[1,2(✉)], and Quan Z. Sheng[3]

[1] School of Software, Dalian University of Technology, Dalian 116620, China
benzhe.su.123@foxmail.com, {xjxu,xiaowei.zhao,xzz}@dlut.edu.cn
[2] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China
[3] School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
michael.sheng@adelaide.edu.au

Abstract. Short-term traffic flow forecasting is an important component of intelligent transportation systems. The forecasting results can be used to support intelligent transportation systems to plan operation and manage revenue. In this paper, we aim to predict the daily floating population by presenting a novel model using taxi trajectory data and weather information. We study the problem of floating traffic flow prediction with weather-affected New York City, and a new methodology called *WTFPredict* is proposed to solve this problem. In particular, we target the busiest part of the city (i.e., the airports), and identify its boundary to compute the traffic flow around the area. The experimental results based on large scale, real-life taxi and weather data (12 million records) indicate that the proposed method performs well in forecasting the short-term traffic flows. Our study will provide some valuable insights to transport management, urban planning, and location-based services (LBS).

Keywords: Weather data · Prediction model · Big data · Intelligent transportation systems

1 Introduction

In a transportation network, the purpose of Intelligent Transportation Systems (ITS) is to provide dynamic traffic control and management real-time by forecasting traffic flows in the short-term future [6]. With the widespread use of GPS-equipped taxis, a huge amount of data about taxis' trajectories with location information is being generated by an increasing speed. Traffic flow prognostication [4] as an interesting problem draws an intensive attention from researchers for decades.

Traffic flow forecasting (*TFF*) has become important as the number of vehicles in big cities and freeway is continually increasing [2]. TFF could improve

J. Li et al. (Eds.): ADMA 2016, LNAI 10086, pp. 507–519, 2016.
DOI: 10.1007/978-3-319-49586-6_35

the traffic and help people to do wiser decision, especially in big cities. Traffic flow is affected by many factors, for example, human behavior, vehicle trajectory, and transport lights and so on. In our work, we are interested to predict weather-affected traffic flow, and select the New York City as an example. As one of the biggest cities in the world, New York City has been the active target for scientists on transportation study (e.g., visualization of its urban taxi data [5,14]). Meanwhile, New York City Taxi data set was designed to build a wide range of spatio-temporal queries [5]. However, to our best knowledge, there are no research about the weather-affected traffic flow prediction. Generally speaking, New York City has a humid continental climate. It also experiences warm summers with long and cold winters. In the majority of winter seasons, a temperature of –25 °C or lower can be expected in the northern highlands (Northern Plateau) and 5 °C or colder in the southwestern and east-central highlands of the Southern Tier. For example, in January 2016, the winter storm resulted in a travel ban across the city and on Long Island, the shutdown of MTA buses, and the closure of above-ground subway lines through New York.

Undoubtedly, traffic could be affected significantly by the weather conditions. Traffic volume strongly depends on current and post volume of that location and its neighbours. However, to conduct an effective traffic flow prediction study, some important questions need to be answered, namely (i) which are the busiest area (e.g., an airport) of the New York City? (ii) how does the weather affect the traffic flow and the population of reaching the area?, and (iii) how should we design a model to predict the floating traffic flow?

There exist many challenges in the study. The first challenge lies on how we should identify the size of the area. The second challenge is how to deal with data sparseness and coverage. Finally, it is critical to develop consider information loss (e.g., taxi id) due to privacy issues. In this paper, we present our work for effective traffic flow forecasting, using New York City as an example. In a nutshell, the major contributions of our work are listed as the following:

- We study the problem of weather-affected traffic flow prediction in a traffic networks and we introduce a formal definition of short-term traffic flow prediction problem affected by weather.
- We develop an algorithm to extract bondaries of interested areas and also propose a novel approach called *WTFPredict* for traffic flow prediction.
- The results on the real taxis data demonstrate that by considering the weather and traffic flow in the traffic network collectively, the traffic flow prediction accuracy can be significantly improved.

The rest of this paper is organized as follows. Section 2 gives a brief overview on the related literature. Section 3 proposes some preliminaries about basic definitions and problem definitions. In Sect. 4, we first introduce the framework of our work and the algorithms, and then present detailed information about the data source used in our study. Section 5 reports the experimental results. Finally, we draw our conclusions in Sect. 6.

2 Related Work

In this section, we review the related efforts from two aspects: *short-term prediction techniques* and *application domains*.

2.1 Prediction Techniques

In recent years, traffic flow prediction (TFP) has received significant attention and many efficient algorithms have been proposed. The typical medium-long-term traffic flow forecasting approach includes four-stage prediction method (traffic generation, traffic distribution, traffic mode, and traffic assignment). However, this four-stage planning tool does not cover the dynamic properties of flow precisely [11]. These different prototypes can be generally divided into two categories: *model-based methods*, and *data-driven methods*.

Many model-based methods have been used for TFP, including support vector machine (SVM) [4], Kalman Filter [13], neural network and some improved neural network. For example, Ding et al. proposed a new traffic flow time series prognostication by SVM in 2002 [4]. A divide-and-conquer method based on neural network and origin-destination (OD) matrix estimation was developed to forecast the short-term passenger flow in high-speed railway system [15]. In [17], the authors propose a Bayesian combined neural network approach to predict short-term freeway traffic flow. Other researchers studied short-term traffic forecasting by tensor techniques. Based on dynamic tensor completion (DTC), a novel short-term traffic flow prediction approach was designed to use the multimode information to forecast traffic flow with a low-rank constraint [12]. In their model, the traffic data are represented as a dynamic tensor pattern so as to capture more information of traffic flow, namely, temporal variabilities, spatial characteristics, and multimode periodicity. Some researchers also have proposed a data-driven method for local traffic state estimation and prediction. A representative effort in this direction is the work in [1], which exploited available traffic and other information and developed data-driven computational approaches.

2.2 Applications

Traffic flow prediction has been applied in many application areas with a rich literature. The interesting applications can be grouped in the following areas: traffic flow prediction on high-speed railway, public bicycle traffic flow prediction, and population prediction on the beaches. Most of work about traffic flow prediction has been applied to the high-speed railway. For example, the researchers used a hybrid model combining clustering with support vector machine (SVM) to predict the uncertainty of traffic flow [16]. The authors in [9] designed a data fusion algorithm to fuse floating car data with stationary detector data on live traffic in order to eliminate individual source bias and alleviate source-specific limitations. However, there are a little work about traffic flow prediction by considering the effect of weather conditions. In a recent effort, the authors in [10] designed a prediction model to predict the daily floating population based on weather factors, through a multiple liner regression analysis approach.

3 Problem Definition

This section defines the problem of traffic flow prediction with the effect of weather conditions. First, we present some basic definitions in Sect. 3.1. Then, we propose the formal problem definition in Sect. 3.2.

3.1 Basic Definitions

Assume a finite set of distinct taxis in a city, i.e., $Taxi = \{Taxi_1, \ldots, Taxi_n\}$. There are m passagers, i.e., $Passager = \{Passager_1, \ldots, Passager_m\}$. A recording database $D = Trip_1, Trip_2, \ldots, Trip_l$ is a set of trips, where each trip $Trip_r \in D$ is a subset of $Trip^*$ and has an unique identifier r, called $Trip_id$. Every record of *GPS* data is denoted by a *GPS* data point. First, we present the definition of a GPS data point.

Definition 1. *GPS data point.* A piece of GPS data point is about a point in a map including a longitude i and a latitude j.

$$gps_{i,j} = \{longitude_i, latitude_j\} \tag{1}$$

A map consists of all GPS data point from taxis. For example, Fig. 1 shows a map of New York City including all GPS data from taxis in January, 2015 where a taxi is represented as a tiny dot in the figure. The formal definition of a map is shown in Definition 2.

Fig. 1. GPS data points of taxis of New York City in January, 2015

Definition 2. *Map.* A map of a city is composed by all GPS data points with k rows and l columns from $GPS_{1,1}$ *to* GPS_{kl}.

$$Map = \begin{pmatrix} GPS_{1,1} \ldots GPS_{1,l} \\ \ldots \quad \ldots \quad \ldots \\ GPS_{k,1} \ldots GPS_{k,l} \end{pmatrix} \tag{2}$$

Based on the definitions of GPS data point and the map, we present the definition of map segment.

Definition 3. *Map Segment.* A map segment *MS* is a boundary sub-map with a list of intermediate GPS points describing the size of the segment using a small block.

A map is divided into $u \times v$ map segments.

$$Map = \begin{pmatrix} MS_{1,1} \ldots MS_{1,v} \\ \ldots \quad \ldots \quad \ldots \\ MS_{u,1} \ldots MS_{u,v} \end{pmatrix} \tag{3}$$

Definition 4. *Trip.* A Trip record $Trip(P_o, taxi_t, time_p - time_q, GPS_{a,b} - GPS_{c,d})$ is about a passager P_o from a point $GPS_{a,b}$ of picking up by a taxi $taxi_t$ at time $time_p$ to another point $GPS_{c,d}$ of dropping off at time $time_q$.

Definition 5. *Traffic State.* A traffic state TS is about the number matrix of picking up in every map segment from time $time_p$ to time $time_q$.

Definition 6. *Traffic Flow.* A traffic flow TF is a sequence of traffic states $TS_1, TS_2, \ldots, TS_T$, i.e., $TF = \{TS_1, TS_2, \ldots, TS_T\}$, where TS_T is the latest matrix in the stream.

As Fig. 2 shows, the traffic flow is considered to be incrementally growing over time. TF_x is the latest matrix in the stream.

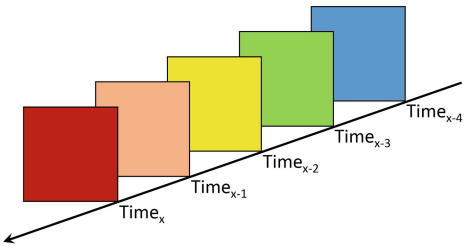


Fig. 2. A 4-order stream: the multimode data are growing incrementally over time

Definition 7. *Weather data.* A Weather data $Weather_u$ is a vector about the average of the weather in an hour, including temperature, humidity, visibility, wind-speed, rain, and so on at time $time_u$, i.e.,

$$Weather_u = \{temperature, humidity, visibility, wind - speed, rainfall\}$$

Similarly, we present the definition of the weather flow.

Definition 8. *Weather Flow.* A weather flow W_u is a sequence of weather data, i.e., $W = \{Weather_1, Weather_2, \ldots, Weather_u\}$.

3.2 Problem Definitions

In general, the floating traffic flow problem of this paper is a time predict problem affected by weather. The definition of traffic flow prediction (TFP) problem is introduced in Definition 9. The formal definition of weather-affected traffic flow prediction ($WTFP$) problem is given in Definition 10.

Definition 9. *Traffic flow prediction (TFP).* For any traffic flow dynamic window, the traffic data in the prediction horizon can be predicted on newly available data, so the prediction problem can be expressed as:

$$TF_{t+1} = f(TF_t) \qquad (t = 1, 2, \ldots, n) \tag{4}$$

where TF_x represents the traffic volume during a time section $time_x$.

Meanwhile, the floating traffic flow problem with the effect of weather is a predict problem, which is to predict the floating traffic flow depending on the weather at a certain time in a certain place.

Given large-scale GPS data and weather data of a city, the goal of this paper is to find the relationship function f between current/pass traffic data and the future traffic volume with the effect of the weather. In the following, the definition of the TFPW is introduced and the related notations is given.

Definition 10. *Weather-effected Traffic flow prediction (WTFP).* For any traffic flow dynamic window, the traffic data in the prediction horizon can be predicted on newly available data, so the prediction problem can be expressed as:

$$TFPW_{t+1} = f(TFPW_t, W_t)(t = 1, 2, \ldots, n) \tag{5}$$

where t represents a time.

4 Time Series Prediction Model

Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. In this section, we solve the following problems by using large-scale data generated by taxi GPS devices around the airport area. First we introduce our data source, including map data, taxi data and weather data. Then, we identify the boundary of an airport and compute the traffic flow of an airport. Finally, we design a novel algorithm to predict the short-term traffic flow in an airport.

4.1 Prediction Framework

Figure 3 depicts overall framework of the proposed prediction model including the analytical components. Algorithm 1 shows the pseudocode of our *WTFPrediction* model. In the following, we discuss different components of our model in detail.

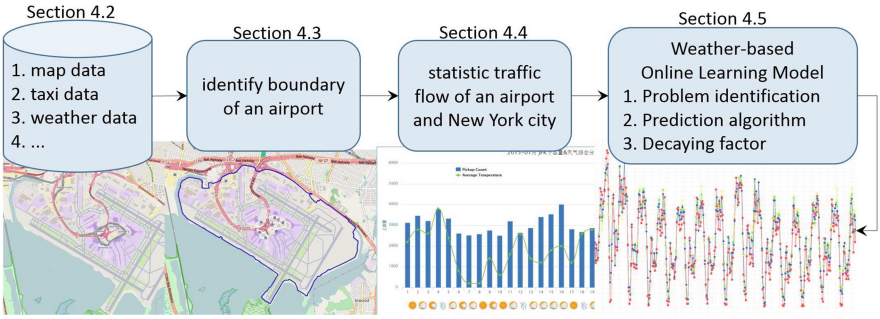


Fig. 3. WTFPrediction algorithm framework

The *WTFPredict* algorithm is detailed in Algorithm 1.

Algorithm 1. *WTFPredict* algorithm

Input: a data source DB, includes map data, weather data, traffic flow
Output: predicted traffic flow

1. Data preprocessing. All data are rescaled to a specific range for a time series forecasting problem.
2. The initial input data are the weather data and traffic flow.
3. Train a model according to Definition 10 to predict the results by a prediction algorithm.
4. Test the model.
5. Get the evaluation results.

4.2 Data Source

In this section, we present our data source about weather and traffic flow.

Taxi Data Source. Our dataset includes trip records from all trips completed in yellow and green taxis in NYC in 2014 and some selected months of 2015 [3]. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP)[1]. In our dataset, we selected the data from 1, Jan, 2015 to 31, Jan, 2015 which involve 12 millions records. The size of our data is 1.84 GB.

Map Data. Map data is used to identify the boundary of an airport. We use OpensStreetMap (OSM) [7] is used to extract the boundary of the airport.

Weather Data. Two types of weather data are used in our study: observation data and forecast data from the website (www.wunderground.com). From January 1, to January 31, 2015, there are 990 records. The weather data was observation data which recorded the basic information about the weather of a day: the highest temperature, the lowest temperature, average temperature, average humidity and so on.

4.3 Algorithm for Extracting Boundary

For detecting the change of traffic flow with weather-effect, we should identify the boundary of the area in order to compute the traffic flow in the area (e.g., the airport). In this section, we introduce a algorithm to extract the boundary, as shown in Algorithm 2. Figure 4 shows the boundary of the *JFK* airport.

Algorithm 2. Algorithm of Extracting Boundary

Input: a data source DB of map.ost.
Output: the boundary of the airport

1. Find root node from the file osm which includes the boundary of the airport;
2. Find a set of all nodes where (Name = the airport);
3. For each item in node
   (a) Create a new node;
   (b) Add an attribute of longitude;
   (c) Add an attribute of latitude;
   (d) Add the node into the file;
4. Save and output the new file.

[1] The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

4.4 Number of Floating Taxis in Three Airports

After we identify the boundary of an airport, we compute the traffic flow by the real taxi data generated by yellow taxis in New York, USA.

Figure 5 shows the number of picking up and dropping off in the *JFK* airport in Jan, 2015. The tendency of the number of picking up marked by the purple line in the JFK airport is a little similar with the tendency of the temperature in the month in Fig. 5.

4.5 Prediction Algorithms

In this section, we select three different prediction algorithms to predict traffic flow, including linear regression (LR), multi-layer perception (MP), and SMOreg.

Prediction by Line Regression. Linear regression is a statistical technique for modeling the relationship between variables. We apply multiply linear regression to predict traffic flow.

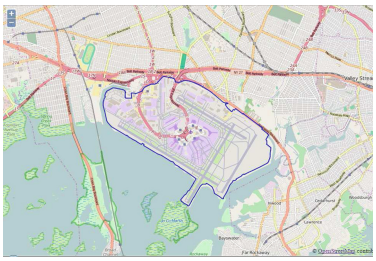


Fig. 4. Boundary of JFK airports in New York City

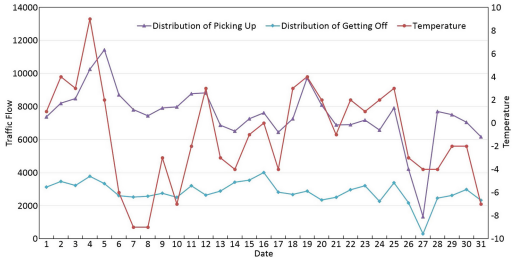
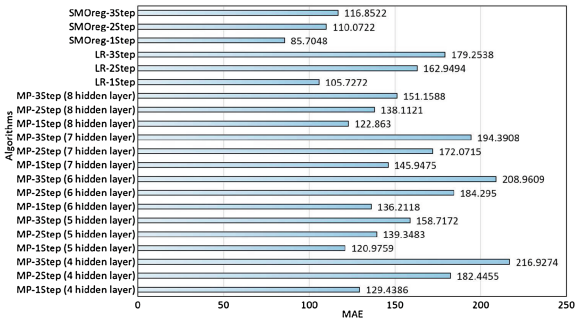


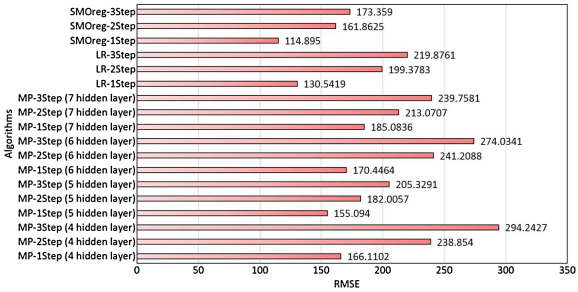
Fig. 5. Distribution of picking up and dropping off in *JFK* airport by day (Jan-2015)

Prediction by Multi-layered Perception. Current weather status is usually affected by previous time. In fact, the weather change could be regarded even as a Markov chain. Therefore, we suppose that the weather change is correlated. Traffic flow is correlated, too. This is the theoretical foundation to adopt artificial neural network (abbreviated as *ANN*) to predict the traffic flow. ANN is a powerful tool to capture and represent complex input/output relationships. Backpropagation (*BP*) feed forward network is a common method of training *ANN* used in conjunction with an optimization method such as gradient descent. Therefore, we decide to use a multi-layered feedforward neural network and adopt the BP learning algorithm to predict the traffic flow.

Prediction by SMOreg Algorithm. The sequential minimal optimization algorithm for training a support vector classifier using RBF kernels (SMOreg), is used in this paper.



(a) MAE



(b) RMSE

Fig. 6. Comparison results of different algorithms

5 Experiment Study

In this section, we show that our proposed method can improve the weather-affected traffic flow accuracy compared with the baseline that only uses traffic flow information. We use weka [8] to test performance of the proposed prediction algorithm.

5.1 Prediction Performance for the *JFK* Airport

In this section, we consider the effect of the different algorithms in prediction of traffic flow. To compare the error of prediction, we select 70 % of the data to train a model and the rest 30 % to test the model. We use mean absolute error (MAE) and root mean squared error ($RMSE$) to measure the performance of our algorithm, which are calculated using:

$$MAE = sum(abs(predicted - actual))/N \tag{6}$$

$$RMSE = sqrt(sum((predicted - actual)^2)/N) \tag{7}$$

The test results of algorithms can be seen from Fig. 6(a) and (b). Smaller values of MAE and $RMSE$ indicate better performance of the algorithms in

prediction. From the figures, we can see that the better performance is obtained by the *SMOreg* algorithm in traffic flow prediction. It should be interesting to note that relatively good performance is also obtained by the *LR* algorithm.

5.2 Prediction Results for the *JFK* Airport

The prediction results of *LR* algorithm is shown in Fig. 7 and Fig. 8 shows the results of *SMOreg* algorithm. Compared Fig. 8 with Fig. 7, we can observe that the error is smaller in Fig. 8. Figure 9 presents the results of MP algorithm with 6 hidden layers. We can see that the SMOreg algorithm for regression performs the best in all algorithms. It should be particular mention that there was no taxi data on 27 January because of the weather condition. All algorithms do not consider that situation which leads to some error on that day.

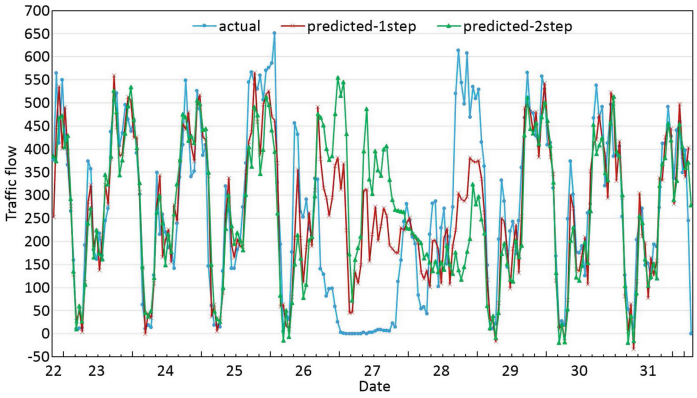


Fig. 7. Predict traffic flow (LR algorithm) for getting off in *JKF* airport by hour

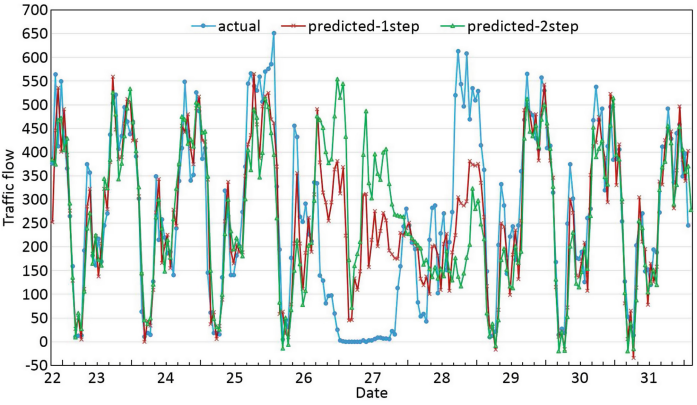


Fig. 8. Predict traffic flow (SMOreg algorithm) for getting off in *JKF* airport by hour

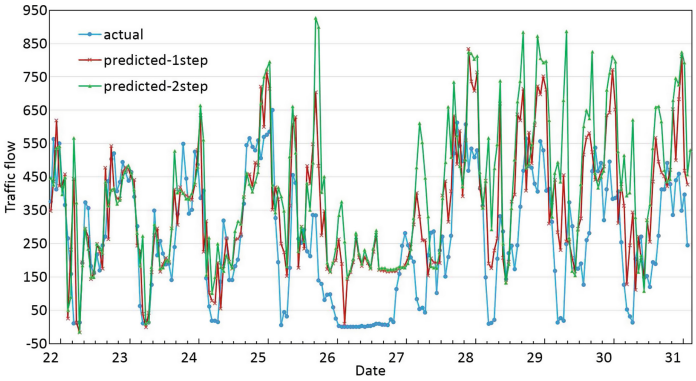


Fig. 9. Predict traffic flow (MP algorithm) for getting off in *JKF* airport by hour

6 Conclusion

Traffic flow prediction is an important problem with the ever stretching of cities nowadays. Weather is one important factor affecting traffic flow, which unfortunately has not be considered in most existing approaches. In this paper, we study the problem of predicting traffic flow affected by weather conditions, using airports of New York City as an example. In particular, we propose the *WTFPredict* model which first defines weather-affected traffic flow, then builds three different algorithms to predict weather-effected traffic flow, and compares those algorithms. These results offer an overall view on the current status of the research and development on intelligent transportation systems (ITS). We hope that our work can help researchers better understand the research status of ITS and gain valuable insight on the future technical trends of the area.

Acknowledgment. This work was supported in part by the Natural Science Foundation of China under Grant 61502069, 61300087 by the Natural Science Foundation of Liaoning under Grant 2015020003, by the Fundamental Research Funds for the Central Universities under Grant DUT15QY40.

References

1. Antoniou, C., Koutsopoulos, H.N., Yannis, G.: Dynamic data-driven local traffic state estimation and prediction. Transp. Res. Part C: Emerg. Technol. **34**, 89–107 (2013)
2. Barros, J., Araujo, M., Rossetti, R.J.: Short-term real-time traffic prediction methods: a survey. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 132–139. IEEE (2015)
3. NTL Commission: TLC Trip Record Data. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
4. Ding, A., Zhao, X., Jiao, L.: Traffic flow time series prediction based on statistics learning theory. In: Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, pp. 727–730. IEEE (2002)

5. Ferreira, N., Poco, J., Vo, H.T., Freire, J., Silva, C.T.: Visual exploration of big spatio-temporal urban data: a study of new york city taxi trips. IEEE Trans. Vis. Comput. Graph. **19**(12), 2149–2158 (2013)
6. Ghosh, B., Basu, B., O'Mahony, M.: Bayesian time-series model for short-term traffic flow forecasting. J. Transp. Eng. **133**(3), 180–189 (2007)
7. Haklay, M., Weber, P.: Openstreetmap: user-generated street maps. IEEE Pervasive Comput. **7**(4), 12–18 (2008)
8. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361. IEEE (1994)
9. Houbraken, M., Audenaert, P., Colle, D., Pickavet, M., Scheerlinck, K., Yperman, I., Logghe, S.: Real-time traffic monitoring by fusing floating car data with stationary detector data. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 127–131. IEEE (2015)
10. Lee, K., Hong, B., Lee, J., Jang, Y.: A floating population prediction model in travel spots using weather big data. In: 2015 IEEE Fifth International Conference on Big Data and Cloud Computing (BDCloud), pp. 118–124. IEEE (2015)
11. Saw, K., Katti, B., Joshi, G.: Literature review of traffic assignment: static and dynamic. Int. J. Transp. Eng. **2**(4), 339–347 (2014)
12. Tan, H., Wu, Y., Shen, B., Jin, P.J., Ran, B.: Short-term traffic prediction based on dynamic tensor completion
13. Wang, Y., Papageorgiou, M., Messmer, A.: Real-time freeway traffic state estimation based on extended kalman filter: adaptive capabilities and real data testing. Transp. Res. Part A: Policy Prac. **42**(10), 1340–1358 (2008)
14. Wang, Z., Ye, T., Lu, M., Yuan, X., Qu, H., Yuan, J., Wu, Q.: Visual exploration of sparse traffic trajectory data. IEEE Trans. Vis. Comput. Graph. **20**(12), 1813–1822 (2014)
15. Xie, M.Q., Li, X.M., Zhou, W.L., Fu, Y.B.: Forecasting the short-term passenger flow on high-speed railway with neural networks. Comput. Intell. Neurosci. **2014**, 23 (2014)
16. Xu, H., Ying, J., Wu, H., Lin, F.: Public bicycle traffic flow prediction based on a hybrid model. Appl. Math. Inf. Sci **7**, 667–674 (2013)
17. Zheng, W., Lee, D.H., Shi, Q.: Short-term freeway traffic flow prediction: bayesian combined neural network approach. J. Transp. Eng. **132**(2), 114–121 (2006)