DP-LRT: An Urban Short-term Traffic Speed Forecasting Method Based on Data Driven

Yu-long HU$^{1,2,*}$ and Lun LUO$^2$

$^1$School of Earth and Space Sciences, Peking University, Beijing, China  
$^2$China Transport Telecommunications & Information Center, Beijing, China  

*Corresponding author

Keywords: Short-term traffic forecasting, Urban traffic, Likelihood ratio test.

Abstract. Until recently, most short-term traffic speed forecasting algorithms have been applied on freeway, arterial or corridor. Short-term traffic speed forecasting on urban road network formed a more complex problem than freeway predictions due to constraints such as signalization. This paper proposes a Date Pattern and Likelihood Ratio Test statistics based (DP-LRT) method for urban short-term traffic speed forecasting from large scale taxi GPS data. Experiment showed that the proposed method reduced the mean absolute percent error and improved the prediction success rate compared with K-NN method, which was a representative data-driven method.

Introduction

Urban short-term traffic forecasting is a hot topic in the studies of transportation and plays a vital element in the fields of smart city and urban computing, which aims at predicting the evolution of urban traffic over time horizons ranging from few seconds to few hours.

The approaches used in short-term traffic forecast can be broadly classified into four categories: naive, parametric, non-parametric, and hybrid. Naive approaches refer to models that provide simple estimate of traffic in the future, e.g., historic averages. Parametric approaches refer to models-based techniques which require a set of fixed parameter values as part of the mathematical or statistical equations they utilize, e.g., analytical models, macroscopic or microscopic models [1], time series analysis models [2,3]. Most of these methods are plagued by their assumption of considering parametric models and have been shown to perform poorly under unstable traffic conditions, complex road settings, as well as when faced with extensive datasets with both structured and unstructured data [4,5]. Non-parametric methods are primarily data-driven and use empirical algorithms to provide predictions, e.g., approaches based on data analysis [6,7] and neural network techniques [8,9].

Given the fact that there are a variety of models developed for short-term traffic forecast, it is difficult to determine which method is most suitable for a particular situation. Various researchers have concluded that the performance of data-driven model is better compared to parametric model when faced with complex traffic conditions, because they are more suitable for learning more complex data [4,6]. However, in spite of the advantages, existing data-driven methods such as K-NN only concentrates on a single section of a highway and lacks the observation on the whole area of a city [4].

Methodology

The urban short-term traffic speed forecast method proposed in this paper is a data-driven approach that can provides predications on traffic information of a whole city based on the historical traffic patterns of sub-region. Initially, the urban space will be divided into grids based on Quadtree method and a comprehensive analysis of spatial and temporal characteristics of urban traffic speed data will be performed. The periodic pattern of traffic speed will be mined based on a Date Pattern method. Finally, a statistically significant range of traffic speed will be predicted based on Likelihood ratio test (LRT).
**Date Pattern Method.** In this research, the expected traffic speed refers to the mean historical traffic speed based on a certain date pattern. A date pattern represents a series of historical dates. Since the daily traffic information in the city will be slightly different, the historical data of different dates will have different traffic speed expectations, which will exert impact on the forecasting results. To reduce forecasting error and improving forecasting success rate, this research classifies the date pattern into four types: ADs (all days), WHDs (workdays or holidays), SWDs (same day of a week) and PTDs (previous two days). The four abbreviations separately represent selecting data on different historical dates.

**Forecasting Based on LRT.** Considering the recent works of which used LRT statistics to describe traffic patterns and find the anomalous region[10], we use the likelihood ratio test method to predict a traffic speed range with statistical significance for each grid of urban after calculating the expected traffic speed based on the selected date pattern:

Given a data set $X$, the model distribution $f(X, \theta)$, a null hypothesis $H_0$: $\theta \in \Theta_0$ and an alternate hypothesis $H$, LRT is the statistic:

$$
\lambda = \frac{\sup\{L(\theta|X); \theta \in \Theta_0\}}{\sup\{L(\theta|X); \theta \in \Theta\}}
$$

Where $L(\cdot)$ is the likelihood function. $\theta$ is a set of parameters coming from complete parameter space $\Theta$ and null parameter space $\Theta_0$. This statistic is computed by first computing a maximum likelihood estimate (MLE) under both parameter spaces $\Theta_0$ and $\Theta$, and then computing the ratio of the likelihood obtained via the two MLEs [11]. Under mild regularity conditions, the asymptotic distribution of $\Lambda(X) = -2\log\lambda$ follows a $\chi^2$ distribution with $k$ degrees of freedom, where $k$ is the number of free parameters. For a given confidence level $\alpha$ and a given degree of freedom $k$, e.g., $\alpha=5\%$, $k=1$, the value of the $\chi^2$ is 3.84. The probability of $\Lambda(X)<3.84$ is 95%:

$$
P(\Lambda(X) < 3.84) = 95\%
$$

The traffic speed $n$ obeying Poisson distribution [12]:

$$
P(X = n) = \frac{e^{-m}m^n}{n!}
$$

$m$ is the parameter of the distribution. So, the likelihood of any given grid $G$ is:

$$
L(n|G) = \frac{e^{-m}m^n}{n!}
$$

The MLE of $m$ is expected value under null parameter space $\Theta_0$, and it is real value under complete parameter space $\Theta$. So, the LRT statistic of grid $G$ is:

$$
\lambda_G = \frac{L(n_0|G)}{L(n|G)} = \frac{e^{-(n_0)}(n_0)^n/n!}{e^{-(n)}(n)^n/n!} = \left(\frac{n_0}{n}\right)^n e^{(n-n_0)}
$$

$n_0$ is the expected value given by Date Pattern method. The $\Lambda$ of grid $G$ is given by:

$$
\Lambda_G = -2\ln \lambda_G = 2n\ln \left(\frac{n}{n_0}\right) + 2(n_0 - n)
$$

With the following formula:

$$
2n\ln \left(\frac{n}{n_0}\right) + 2(n_0 - n) = 3.84
$$

We can get the confidence interval of $n$, $[n_{min}, n_{max}]$, which representing the probability of the traffic speed falling within the interval is 95%:

$$
P(n_{min} \leq n \leq n_{max}) = 95\%
$$

When the actual value of traffic speed is placed among the confidence interval, we consider it as a successful prediction. Therefore, the predicted result in this paper is not a single value, but a
statistically significant range. For example, when \( k_0 \) is 10, 50, 100, respectively, the confidence intervals of the \( k \) are showed in Table 1:

| \( k_0 \) (km/h) | Confidence Intervals (km/h) |
|------------------|----------------------------|
| 10               | 5-17                       |
| 50               | 37-65                      |
| 100              | 82-121                     |

**Measuring Performance.** This paper used MAPE (Mean Absolute Percentage Error), SR (Successful Rate) and HFSR (Historical Forecasting Successful Rate) to Measure the performance of the proposed urban traffic forecast method. MAPE provides the forecast error in terms of percentage difference between the observed and predicted traffic speed. SR represents for a certain timeslot, the proportion of the number of successfully predicted grids to the total number of predicted grids. HFSR represents for a certain grid, the proportion of the number of successfully predicted timeslots to the total number of predicted timeslots. The forecast method which provides the least MAPE and the highest SR, HFSR can be identified as an optimal method.

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{F_i - O_i}{O_i} \right| \times 100\%. \tag{9}
\]

\[
SR = \frac{S_g}{N_g} \times 100\%. \tag{10}
\]

\[
HFSR = \frac{S_t}{N_t} \times 100\%. \tag{11}
\]

Where \( F_i \) is the forecast traffic speed of \( i \)th timeslot grids. \( O_i \) is the observed traffic speed of \( i \)th timeslot grids. \( n \) is the total number of grids. \( N_g \) is the total number of grids. \( S_g \) is the number of grids which are forecasted successful. \( N_t \) is the total number of timeslots. \( S_t \) is the number of timeslots which are forecasted successful.

**Experiments and Results**

Data-driven urban traffic speed forecast method demands a comprehensive analysis of data and exploration of temporal and spatial characteristics of urban traffic. The GPS data in the present research are collected from more than 12000 taxi in Beijing metropolitan area on the days of November 2012, located at 139.703°-40.210° N, 16.082°-116.726° E. The collected data contains more than 972 million GPS records with more than 30 million per day.

**Choosing a Suitable Date Pattern.** Since the daily traffic speed in the city will be slightly different, selecting the historical data of different dates will get different traffic speed expectations, which will exert impact on the forecasting results. It is of necessity to make comparison among testing results in different date patterns to select one with minimum forecasting error and maximum forecasting successful rate.

Figure 1 shows the comparison of forecast performance with different date patterns. It can be observed that the WHDs pattern has the lowest forecast error and the highest forecast success rate. Therefore, the WHDs date pattern is used throughout the remainder of this paper.
Comparing the Results with Enhanced K-NN. To evaluate the performance of the proposed method, the results of DP-LRT proposed in this paper are compared with those from the enhanced K-NN (set key parameter $K=5$) [4]. Figure 2 shows the forecast error and success rate from the proposed approach and K-NN by time of the day. As shown in the figure, the proposed approach is able to provide traffic speed forecasting with lower prediction errors. Figure 3 shows the number of grids of different HFSR of the two methods and proves that the method proposed in this paper shows more accuracy in forecasting more urban areas.

Summary

The urban short-term traffic speed forecast method proposed in this paper is a data-driven approach which can provide predications on traffic information of a whole city based on the historical traffic patterns. To reduce forecasting error and improving forecasting success rate, this research classifies the date pattern into four types: ADs (all days), WHDs (workdays or holidays), SWDs (same day of a
week) and PTDs (previous two days). A statistically significant range of traffic speed was predicted based on Likelihood ratio test (LRT) in one data pattern. Different from previous short-time traffic forecasting studies, it can not only make predictions on traffic speed in different urban areas but also identify the level of intensity of traffic periodicity in different areas, by which urban design and administration can be facilitated with clear direction. For instance, more sensors could be installed in areas with weak periodicity to intensify monitoring and traffic anomaly could be easily traced in areas with strong periodicity, which bring convenience to further studies on traffic anomaly detection.

Acknowledgement
This research was financially supported by the National Natural Science Foundation of China (No.91646207) and National Key R&D Program of China (No. 2017YFC0822003).

References
[1] Fowe A J, Chan Y. A microstate spatial-inference model for network-traffic estimation[J]. Transportation Research Part C Emerging Technologies, 2013, 36(11):245-260.
[2] Guo J, Huang W, Williams B M. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification[J]. Transportation Research Part C, 2014, 43:50-64.
[3] Wang X, Fan T, Li W, et al. Speed variation during peak and off-peak hours on urban arterials in Shanghai[J]. Transportation Research Part C: Emerging Technologies, 2016, 67:84-94.
[4] Habtemichael F G, Cetin M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns[J]. Transportation Research Part C, 2016, 66:61-78.
[5] Vlahogianni E I, Karlaftis M G, Golias J C. Short-term traffic forecasting: Where we are and where were going[J]. Transportation Research Part C Emerging Technologies, 2014, 43(1).
[6] Pascale A, Deflorio F, Nicoli M, et al. Motorway speed pattern identification from floating vehicle data for freight applications[J]. Transportation Research Part C: Emerging Technologies, 2015, 51:104-119.
[7] Shi Q, Abdel-Aty M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways[J]. Transportation Research Part C: Emerging Technologies, 2015:S0968090X15000777.
[8] Li C S, Chen M C. Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks[J]. Neural Computing and Applications, 2013, 23(6):1611-1629.
[9] Wang J, Shi Q. Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet Analysis-Support Vector Machine theory[J]. Transportation Research Part C Emerging Technologies, 2013, 27(2):219-232.
[10] Pang L, Chawla S, Liu W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data & Knowledge Engineering, 2013, 87(9):357-373.
[11] Wu M, Song X, Jermaine C, et al. A LRT framework for fast spatial anomaly detection[C]//Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. DBLP, 2009.
[12] Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics - Theory and Methods 26(6), 1481-1496.