Transportation Research Part C 85 (2017) 476–493

Contents lists available at ScienceDirect

Transportation Research Part C

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

Probe data-driven travel time forecasting for urban expressways by matching similar spatiotemporal traffic patterns

Zhihao Zhang[a], Yunpeng Wang[a], Peng Chen[a,*], Zhengbing He[b], Guizhen Yu[a]

[a] School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Infrastructure System and Safety Control, Beihang University, Beijing 100191, China
[b] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Keywords:
Travel time forecast
Probe data
Pattern-matching
Spatiotemporal traffic patterns
Urban expressway

ABSTRACT

Travel time is an effective measure of roadway traffic conditions. The provision of accurate travel time information enables travelers to make smart decisions about departure time, route choice and congestion avoidance. Based on a vast amount of probe vehicle data, this study proposes a simple but efficient pattern-matching method for travel time forecasting. Unlike previous approaches that directly employ travel time as the input variable, the proposed approach resorts to matching large-scale spatiotemporal traffic patterns for multi-step travel time forecasting. Specifically, the Gray-Level Co-occurrence Matrix (GLCM) is first employed to extract spatiotemporal traffic features. The Normalized Squared Differences (NSD) between the GLCMs of current and historical datasets serve as a basis for distance measurements of similar traffic patterns. Then, a screening process with a time constraint window is implemented for the selection of the best-matched candidates. Finally, future travel times are forecasted as a negative exponential weighted combination of each candidate's experienced travel time for a given departure. The proposed approach is tested on Ring 2, which is a 32km urban expressway in Beijing, China. The intermediate procedures of the methodology are visualized by providing an in-depth quantitative analysis on the speed pattern matching and examples of matched speed contour plots. The prediction results confirm the desirable performance of the proposed approach and its robustness and effectiveness in various traffic conditions.

1. Introduction

Traffic congestion has become a critical problem. Advanced Traveler Information Systems (ATISs) and Advanced Traffic Management Systems (ATMSs) provide a practicable solution for both traffic managers and travelers, which can improve the efficiency and service standard of existing transportation systems and mitigate road congestion problems. Numerous studies (Tu et al., 2008; Liu, 2008; Khosravi et al., 2011; Cai et al., 2016; Chen et al., 2017a, 2017b) have shown that travel time, as an important measure of roadway performance in ATIS and ATMS, can be easily perceived and is applicable to the perspectives of both road operators and users. Consequently, accurate travel time information is urgently needed for better travel planning, route choice, and congestion alleviation.

Due to the highly dynamic and nonlinear nature of traffic states over time and space, travel time forecasting remains a difficult yet important challenge. Thanks to advanced traffic sensing technologies (Zhu et al., 2013), various travel time-related information can

* Corresponding author.
E-mail addresses: zhhhzhang@buaa.edu.cn (Z. Zhang), ypwang@buaa.edu.cn (Y. Wang), cpeng@buaa.edu.cn (P. Chen), he.zb@hotmail.com (Z. He), yugz@buaa.edu.cn (G. Yu).

http://dx.doi.org/10.1016/j.trc.2017.10.010
Received 12 April 2017; Received in revised form 15 August 2017; Accepted 8 October 2017
Available online 24 October 2017
0968-090X/

be conveniently collected, which offers a new solution to congestion by monitoring and disseminating traffic information to road users. These technologies include direct measurement and estimation (Yeon et al., 2008). Direct measurement can be obtained by point-to-point travel time collection, e.g., license plate recognition systems, automatic vehicle identification (AVI), and Bluetooth technologies. Estimation is derived from station-based traffic state measuring devices, e.g., loop detector, global positioning system (GPS) devices and probe vehicle techniques. Because direct measurement usually cannot guarantee sufficient statistics for low sampling rates, estimation approaches are pursued. Currently, a surge of studies has focused on the development of data-driven travel time estimation/prediction approaches. For an extensive review, one can refer to the study by Karlaftis and Vlahogianni (2011), Oh et al. (2015), Mori et al. (2015). The existing data-driven methods can be classified into two major categories, i.e., parametric approaches and non-parametric approaches.

Parametric approaches rely on established theories, including linear regressive models (Zhang and Rice, 2003), time series models such as the Kalman filter (Lint, 2008) and the Auto-Regressive Integrated Moving Average (ARIMA) (Xia and Huang, 2011). Non-parametric approaches consist of neural networks (NN) (Zeng and Zhang, 2013), support vector regression (SVR) (Lam and Toan, 2008), and k-nearest-neighbor (KNN) (Kobrin, 2011), hybrid or ensemble techniques (Zhang and Haghani, 2015). These methods are implemented by direct or indirect procedures to predict travel times using different types of state variables. Direct procedures utilize the travel time in past periods as the state variable to predict future travel time. Indirect procedures are performed using other variables (such as traffic speed, density, flow, and occupancy) as the state variable to predict state condition, and future travel time can be calculated based on the transition.

Parametric approaches, which establish the theoretical framework with straightforward model structures, have been utilized to predict short-term travel time with various degrees of success. For example, time series models were employed to construct the time series relationship of travel time or traffic state, and current and/or past traffic data were employed in the constructed models to predict travel times in the next time interval (Chen et al., 2016). A vast amount of freeway traffic prediction literature addresses these models (Kamarianakis and Prastacos, 2005; Billings and Yang, 2006; Chen et al., 2012; Xiong et al., 2014). Among these types of models, ARIMA is extensively recognized as an accepted framework to construct a freeway traffic prediction model due to its well-defined theoretical foundations and effectiveness in prediction (Karlaftis and Vlahogianni, 2009). Fei et al. (2011) developed the Bayesian dynamic linear model for real-time, short-time travel time prediction in two traffic conditions of recurrent and non-recurrent congestion. Some parametric approaches have been combined with macroscopic traffic models (Celikoglu, 2013a, 2013b) to predict travel time. These estimation models, which are based on flow conservation and propagation principles, enhance the adaptiveness to the actual measurement. Wang et al. (2008) utilized macroscopic traffic flow models and the Kalman filtering technique to predict freeway travel time with synthetic detector and probe vehicle data. However, the accuracy rapidly degrades with an increase in the prediction temporal horizon. Parametric approaches can produce desirable forecasts in traffic conditions that show regular variations. However, they may be computationally extensive and produce larger deviations when highly dynamic and nonlinear traffic states are present over time and space.

Conversely, alternative approaches include building the function of the current and previous traffic states by assuming hidden variables. These approaches are essentially non-parametric and are capable of capturing the underlying structure of data and excavating the intrinsic spatiotemporal traffic features without strong assumptions about its temporal evolution. Support vector regression (SVR) techniques (Lam and Toan, 2008) have been commonly employed to predict travel time. This approach maps the data into a higher dimensional space using a kernel function to find the flattest linear function that relates these modified input vectors with the target variable. Then, the linear function is mapped into the initial space to obtain a final nonlinear function and predict travel time. In addition, neural network (NN) models have gained increasing attention and are frequently applied in traffic state prediction. An NN can be trained using historical data to identify hidden dependencies that can be employed for predicting future states. Lint et al. (2005) proposed a state space neural network (SSNN) method to predicted freeway travel time. Recently, a novel recurrent neural network method that considers temporal-spatial input dynamics for freeway travel time modeling was developed (Zeng and Zhang, 2013). However, NN-based methods employ black-box procedures to predict traffic conditions and lack a suitable interpretation of the dependency relationship. The applications of the methods require long training processes and are usually nontransferable to other sites.

As another representative non-parametric approach, the pattern-matching method is easy to implement at different sites without data training, as required by the NN and SVR. The basic assumption of the pattern-matching method is that traffic patterns are recurrent in nature and the historical database, which provides a variety of previously experienced traffic patterns, can be employed to predict future “experienced” traffic states, e.g., travel times by matching current data to historical data. Candidates were selected by the Euclidean distance or other data trend measures and utilized to predict travel times in different conditions. The pattern-matching method has greater flexibility than the parametric approach. The predictions are based on huge amounts of data which are presumed to contain a variety of traffic state patterns. Several studies have applied the pattern-matching method to predict traffic flow rates (Zheng and Su, 2014; Habtemichael and Cetin, 2015). Similarly, other studies have applied this method for travel time prediction (Bajwa and Kuwahara, 2003; Bajwa et al., 2005). The main shortcomings of these studies are that only the simplest forms of KNN were employed with the consideration of high-dimensional spatiotemporal traffic states in a limited manner, which prevents the implementation of the method in large-scale traffic networks. Due to the short spacing of merging and diverging traffic flows, the traffic states on urban expressways are considerably more unstable than freeways. Compared with the extensive application of the pattern-matching method for traffic states or travel time prediction on freeways (Zou et al., 2014), less attention has been given to urban expressways.

In principle, the performance of any data-driven approach is dependent on the representativeness and extensiveness of the data. An accurate forecast of dynamic travel time requires the availability of high-resolution data. In the past few years, interval detectors

are more promising than point detectors since interval detectors provide an extensive range of new possibilities for travel time or space mean speed estimation. As the representative of interval detectors, probe vehicles have gained significant attention, and the research on probe data has considerably flourished (Rahmani and Koutsopoulos, 2013). Speed and position data of probe vehicles are periodically collected, e.g., 30 s, then processed to generate traffic information for each road segment at regular time periods. With the growing popularity of probe vehicles and GPS-enabled smartphones, the amount of probe data has rapidly increased. It has been demonstrated to be a reliable and cost-effective method for gathering traffic data over a wide-area road network (Jenelius and Koutsopoulos, 2013). In this study, a vast amount of probe data collected on an urban expressway of Beijing will be employed to develop methodologies for travel time forecasting.

Taken together, existing methods are either insufficient or have limitations for predicting dynamic travel times on urban expressways. Due to the non-stationarity and nonlinearity of traffic states on urban expressways, it remains a challenging yet important issue to characterize the latent evolution of spatiotemporal traffic patterns so as to improve the travel time prediction performance. Hence, considering the increasing demand for traffic predictive tools and the increasing emergence of data collection for large-scale urban networks, this study proposed a simple but efficient pattern-matching method for travel time forecasting on urban expressways. The proposed approach is essentially a data driven method, which can be easily transferred to other sites without the need for a data training process. Instead of one-dimensional time series data, the high resolution of the evolution of similar traffic states in both time and space dimensions will be discovered from historical probe speed data. Note that the traffic states in this study are defined on both spatial and temporal scales. On the spatial scale, traffic states include the entire section of the expressway. On the temporal scale, a sufficient length of time is included to define traffic states evolution. Specifically, the core idea of the proposed pattern-matching method is to use the image process techniques on speed contour plots for feature extraction, and to use data driven techniques to match current traffic patterns with historical spatiotemporal traffic patterns for travel time prediction.

Unlike most of data driven approaches which are considered as “black-box” prediction methods, this study provides an in-depth quantitative analysis on the process of speed pattern matching. Detailed examples of matched speed contour plots will be given to illustrate how the model arrives to predictions. In addition, multi-step-ahead travel time forecasts in complex scenarios with low resolution data will be evaluated to verify the robustness and effectiveness of the proposed approach. The results show that the proposed approach outperforms the other classical methods, i.e., instantaneous travel time, historical average and naïve KNN, in terms of both accuracy and stability of the forecast results.

The remainder of this paper is organized as follows. The data preparation is explained in the next section. Subsequently, the framework and the detail of the methodology are presented. Next, the methodology is implemented in a case study and the forecast results are discussed. The last section provides the conclusions and future research recommendations.

2. Data preparation

2.1. Data description

The probe vehicle data collected on the urban expressways of Beijing, China during 45 days in January and February 2015 were utilized. Every one minute interval, taxis that were equipped with GPS devices uploaded a set of instantaneous information, such as location, direction, speed, and being occupied or not occupied by passengers. One can refer to Fig. 1 for basic information about the region and data. Fig. 1(a) displays the region on an online map, and Fig. 1(b) presents the instantaneous positions of all taxis that are carrying passengers, where the major network emerges basically. Note that only the taxis that were occupied by passengers were

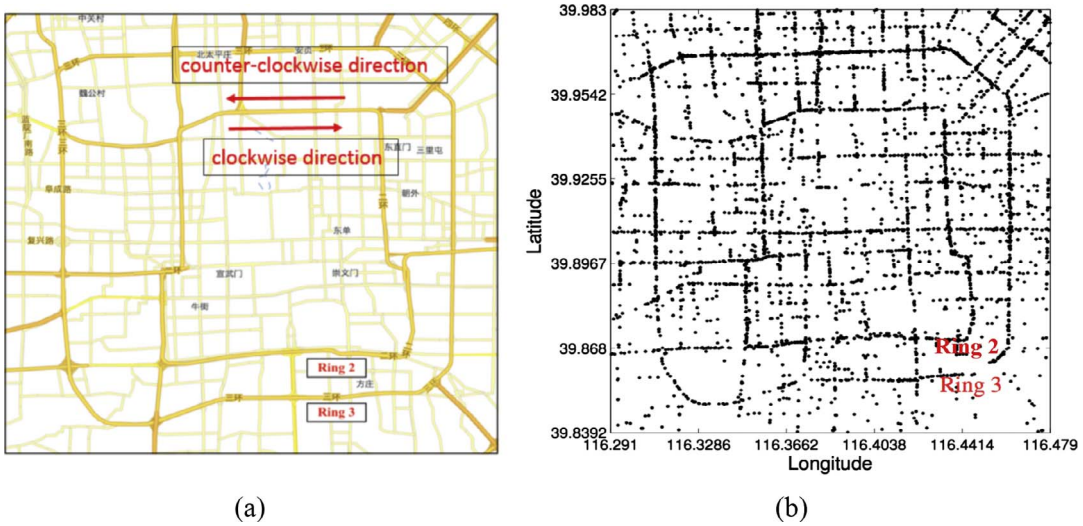

Fig. 1. Schematic of the Beijing urban area and corresponding data: (a) the network that contains Ring 2; (b) Positions of all GPS data points at 8 AM on Jan. 6, 2015.

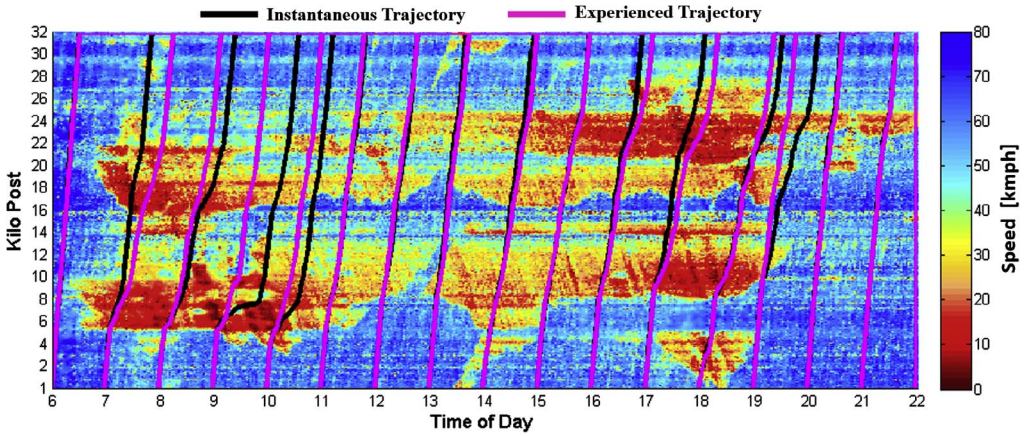

Fig. 2. Spatiotemporal speed diagram of Ring 2 in the clockwise direction on Jan. 6, 2015.

considered here, because the behaviors of these taxis are close to those of regular vehicles.

In Beijing, four two-way urban expressways, i.e., Ring 2–5, enclose the urban area. The lengths of Rings 2–5 are 32, 48, 65 and 98 km. The four urban expressways are surrounded by auxiliary roads, and the auxiliary roads and urban expressways are connected by various types of interchanges with paired entrances and exits for merging or diverging traffic (Zhao et al., 2009; Song et al., 2015). A pair of entrances and exits is located every 1 km along the urban expressways, which operate as potential bottlenecks in critical traffic conditions.

The probe vehicle data in a highly discrete manner cannot be readily available for travel time prediction on the network level of an urban expressway. The complete knowledge of traffic dynamics, which continuously propagate in time and space along the expressway, will be necessary. In practice, a spatiotemporal traffic diagram is preferred by traffic engineers, in which the x-axis is time, the y-axis is space, and the internal color inside (or the z-axis) represents speed. It is usually constructed using stationary detector data (Kerner et al., 2004; Yildirimoglu and Geroliminis, 2013) and can assist in the identification and tracking of traffic patterns in different regimes, e.g., uncongested or congested (Treiber and Kesting, 2013). However, few studies shed light on the reconstruction of spatiotemporal traffic diagrams using probe data.

A recent study by the authors (He et al. 2017) proposed a simple mapping-to-cells method to partition the network region into small homogeneous square cells. To unveil traffic dynamics from the probe data, the method partitions a network region into homogeneous square cells, and builds a cell network based on the number of the GPS points falling into possible downstream cells. The ranges of traffic flow directions pertaining to the cells are determined according to the moving directions of the GPS data inside cells, and the data within the range are aggregated to reconstruct the spatiotemporal traffic diagram. A detailed description of the mapping-to-cell method is beyond the scope of this paper and one can refer to the study by He et al. (2017). Fig. 2 presents the spatiotemporal speed diagram of Ring 2 in a clockwise direction on Jan. 6, 2015. From the diagram, we can observe abundant information about the traffic characteristics on Ring 2, including traffic breakdowns, wide moving jams, and a variety of congestion patterns.

The reconstruction of a spatiotemporal diagram with the mapping-to-cells methods is simple but efficient, especially in processing mass probe data. It potentially supports online applications by transforming the continuous network space into discrete and homogeneous cells and then matching GPS points with the cells. As will be demonstrated in this study, the information-rich spatiotemporal diagrams that are based on regularly spaced probe data serve as a powerful tool and enable travel time forecasting of urban expressway networks with a vast number of possible combinations of origins and destinations.

2.2. Data processing

To further motivate this research, the travel times for departures at any time in the study site need to be estimated in advance for model prediction and validation purposes. Based on the spatiotemporal speed diagrams, two approaches can be employed to compute travel time, i.e., instantaneous and experienced. Instantaneous travel time is calculated using the current speed measurements along the expressway at the departure time of a trip. It fails to consider speed evolution across time. Conversely, the experienced travel time is the actual and realized one, which serves as the ground truth that a vehicle would traverse the entire trip. The calculation of experienced travel times can be obtained using the standard computational procedure (Coifman, 2002; Lint and Zijpp, 2003; Lint, 2010; Li et al., 2006; Sun et al., 2008). To compare the difference, instantaneous and experienced trajectories were constructed for the hourly departures on Ring 2 in a clockwise direction on Jan. 6, 2015, as shown in Fig. 2. With a number of active bottlenecks beginning at different times and locations, the difference between the instantaneous trajectories and the experienced trajectories can be evidently identified. It indicates that the instantaneous travel time is not a suitable indicator of the actual travel time for traffic state transient periods. Conversely, experienced travel time accounts for traffic state variation on a temporal and spatial scope, such as speed evolution and congestion formation or dissipation at point bottlenecks.

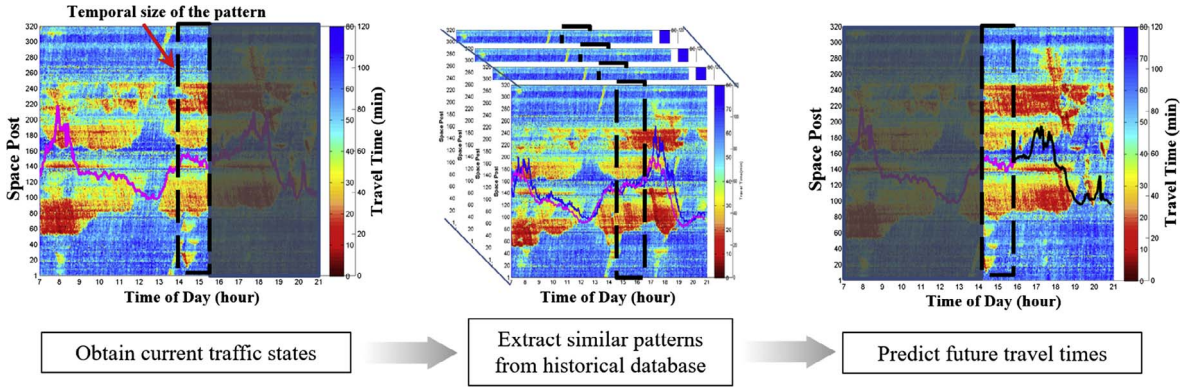

Fig. 3. Basic framework of the proposed methodology for travel time forecast.

2.3. Establishing historical traffic database

Pattern-matching technique is based on the premise that traffic states that are similar to the current state have occurred. Due to this recurrence, historical traffic patterns are crucial to the development of the travel time forecast framework. To manipulate the historical database in an efficient manner, days with similar traffic states are classified such that only similar days from the historical database are searched. Hierarchical clustering has been commonly employed in the transportation field to establish a hierarchy of clusters (Theodoridis and Koutroumbas, 2008; Weijermars and Van Berkum, 2005). Ward's linkage is selected as the similarity measure. It is defined as the incremental sum of squares, i.e., the increase in the total within-cluster sum of squares as a result of joining two clusters.

To prepare for the input to the clustering process, spatiotemporal speed profiles for each day need to be preprocessed. Instead of a matrix, speed profiles at different locations (every 100 m) along the expressway in different time periods (every 2 min from 6:00 to 22:00), i.e., "320 expressway cells" × "480 time periods per day" = 153,600 speed variables, are converted to a time series dataset. The historical dataset can be primarily classified into two clusters, namely, a weekend cluster and a weekday cluster. The weekend cluster includes Saturdays, Sundays and the New Year's Day holiday, i.e., January 1 to 3, whereas the weekday cluster includes Monday to Friday, except for January 4, which falls on a Sunday and involves one day shift of work due to the New Year's Day holiday. It can reduce the magnitude of the search space and improve the computation efficiency of this data-driven approach.

3. Methodology

Based on multiday spatiotemporal diagrams, this study aims to predict expressway travel times by matching real time traffic patterns to historical candidates with similar patterns. As illustrated in Fig. 3, the fundamental framework of the proposed approach comprises the following three steps: (i) obtaining the current traffic states; (ii) extracting similar traffic patterns from the historical records; and (iii) predicting future travel times.

The fundamental assumption of the proposed pattern-matching approach is that traffic patterns are recurrent in nature. Specifically, the intuitive assumption is that the subsequent traffic states of historical candidates are similar to that for the current traffic state. The proposed methodology makes full use of the relationship between the traffic states and the travel times. Given the current traffic pattern as illustrated in the black dashed frame in Fig. 3, the potential candidates are searched among the historical database based on certain similarity or distance measures. A screening process with a time constraint window is implemented for the selection of the best matched candidates. Future travel times for any departure can be predicted by aggregating the experienced travel times for each candidate with associated weights. Some key issues remain to be explored, e.g., determining the similar traffic patterns, the appropriate temporal size for pattern searching, the appropriate number of potential candidates and the principle of aggregation. Detailed explanations of these aspects are provided in the following sections.

3.1. Extracting spatiotemporal traffic patterns

The premise of pattern matching methods is to extract traffic patterns from constructed spatiotemporal speed diagrams that are similar to that for the target profile. Thus, the design of the feature extraction procedure governs the prediction accuracy. In the field of image analysis and pattern recognition, the Gray-level Co-occurrence Matrix (GLCM) (Nixon and Aguado, 2008), which is constructed from pair-wise pixel information at different distances and relative inclination, provides a more effective way to extract more information and make better representation of the prevailing traffic state. Considering that spatiotemporal diagrams provide a pictorial representation of traffic states in time and space, the traffic state which is regarded as an image can be applied in an analogous manner to establish similar measurements for traffic texture characterization. In the context of this study, different values of observed traffic speeds correspond to the gray levels in image analysis.

The GLCM, as the measurement of spatial relationships, is used to describe the texture and degree of variation of an image by

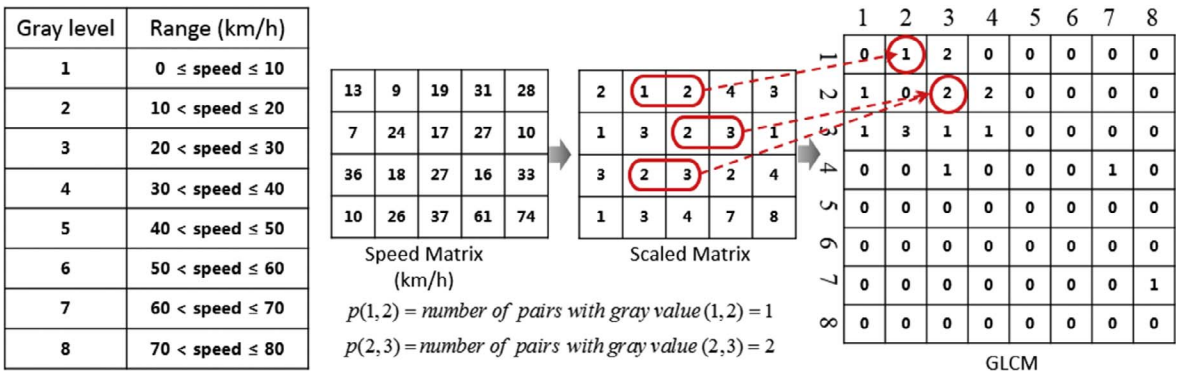

Fig. 4. Calculation of Grey-level Co-occurrence Matrix.

exploring the space relations underlying the pixels gray level distribution, and capture the range, local variation, and the trend of traffic over a short time period (Theodoridis and Koutroumbas, 2008). Thus, the GLCMs derived from traffic states are capable of detecting the high-resolution properties of traffic states evolution and maintaining useful information while suppressing data redundancies. Fig. 4 illustrates the calculation of the GLCM.

Given the speed limit of 80 km/h on Ring 2, the collected speed data are integers ranging from 0 to 80 km/h. As many as 80 possible gray levels would result in excessive computational cost and lower performance. For simplification, gray levels are discretized into eight groups with equally spaced 10 km/h intervals. As an example, the traffic states are defined as an intensity image with a 4 × 5 speed matrix, as shown in Fig. 4. Then, the image is scaled to eight Gray-levels, i.e., mapping the gray values between 1 and 8. Next, GLCM is derived from the resultant scaled matrix as an 8 × 8 matrix by calculating how often a pixel with the gray-level value i is horizontally adjacent to a pixel with the value j according to Eq. (1). Intuitively, it seems to discard arbitrary bits of information because of discretization. However, the GLCMs enable to characterize and maintain the spatial relationships underlying the pixels gray of an image, namely, the evolution of traffic states. All these operations were implemented in MATLAB. The derived GLCM will be utilized to match similar traffic patterns.

$$p(i,j) = \#\{(x_1,y_1),(x_2,y_2) \in M \times N | f(x_1,y_1) = i, f(x_2,y_2) = j)\} \tag{1}$$

where $p(i, j)$ is the value of element (i, j) in the GLCM, $f(x, y)$ is the value of element (x, y) in the scaled matrix, i and j are gray values, M is the height of the speed diagram and N is the width of the speed diagram.

3.2. Matching similar traffic patterns

The critical issue of pattern matching is designing a measurement of similarity. To match similar traffic patterns to the historical dataset, an appropriate measure should be selected as it will substantially affect how the historical dataset is treated. For time series data, the common measures of similarity include correlation coefficient, the Euclidean distance, and the weighted Euclidean distance (Mori et al., 2015). Instead of time series, the data utilized for pattern matching in our context are based on converted GLCMs. Matching the current GLCM to historical GLCMs is analogous to template matching in digital image processing (Hisham et al., 2015). The current GLCM, which is referred to as a template, is searched in the image of historical GLCMs to measure the degree of similarity.

Two similarity measures have been commonly utilized to quantify the pixel-by-pixel intensity differences between two images, i.e., squared differences (SD) and cross-correlation (CC). SD is simple and requires less computation cost since it only involves square operation and pixel subtraction between the template and the original image. Conversely, CC involves numerous multiplication, division and square root operations. Thus, SD is implemented in this study considering the requirement of the computation efficiency by the data-driven approach. SD is defined in Eq. (2)

$$SD = \sum_{T,I} (T(c,L) - I(h,L))^2 \tag{2}$$

where c and h denote the current time and the historical time in the dataset when searching similar patterns, respectively; L is the temporal size of searching; $T(c, L)$ is the GLCM derived from the speed matrix $D(n,m)$ as in Eq. (1) and represents the current traffic pattern; $I(h, L)$ derived from speed matrix $H(n,m)$ as in Eq. (1) represents the historical GLCMs. Specifically, $T(c, L)$ covers the temporal intervals of $[c-L+1, c-L+2,\ldots, c]$ and $I(h, L)$ $[h-L+1, h-L+2,\ldots, h]$.

When implemented, the SD is usually normalized, i.e., normalized squared differences (NSD), as shown in Eq. (3), which can help accelerate the process of pattern matching between the template and the image.

$$NSD = \frac{\sum_{T,I} (T(c,L) - I(h,L))^2}{\sqrt{\sum_{C,L} T(c,L)^2} \cdot \sqrt{\sum_{H,L} I(h,L)^2}} \tag{3}$$

A perfect match has an NSD of 0 and bad matches have large NSDs. Our analysis will show that NSD can be a superior measure of similarity for traffic pattern matching. In the process of dynamic searching, all the traffic patterns will be screened and matched from the historical dataset. By defining an acceptable threshold of NSD, the computation will yield a group of matrices that have the same dimensional size as the template. These matrices, which are referred to as candidates, will be prepared for the prediction of experienced travel time.

3.3. Weight assignment to the candidates

Consequently, the K best matched patterns or candidates are selected according to the ascending order of NSD. Note that NSD helps scale the similarity between the target profile and the candidates, but cannot guarantee the similarity for subsequent traffic states. There are instances that the subsequent traffic patterns of selected candidates are not similar to that for the target profile even though NSD is smaller. In such a situation, it may result in large errors for future travel time prediction by simply aggregating the experience travel times for each candidate with simple weighted average.

To tackle it, a screening process based on the negative exponential similarity function is adopted by referring to the study by Kasai and Warita (2014). It enables to vary the weight as time passes on the discrepancy between historical candidates and the target pattern in the matching period. Being different from Kasai and Warita (2014), experienced travel times at each interval, instead of instantaneous travel times, are used to measure the discrepancy. The reasons are provided below. Experienced travel times can be intuitively viewed as the results of line integrals over the spatiotemporal speed diagram. They serve as the ground truth of the actual realized travel times and can reflect the evolution of subsequent traffic states to a large extent. Similar experienced travel times in the matching period are supposed to imply similar subsequent traffic states. Thus, the weight assignment is conducted to adjust the contribution of each candidate to the prediction value according to the discrepancy of the experienced travel times at each interval within the matching period. An example is provided in Fig. 5 to illustrate the process of weight assignment.

Step 1: Calculate RMSEs between historical candidates and the target profile within a matching period

Assume that two candidates have been selected in the historical dataset to predict the target profile at a future time, as shown in Fig. 5. Experienced travel times at each interval within the matching period are used to measure the similarity between the historical profile and the target profile. In the module, the matching period contains several intervals, which correspond to a number of experienced travel times. As suggested by Kasai and Uchiyama (2010), the similarity between the historical candidates and the target

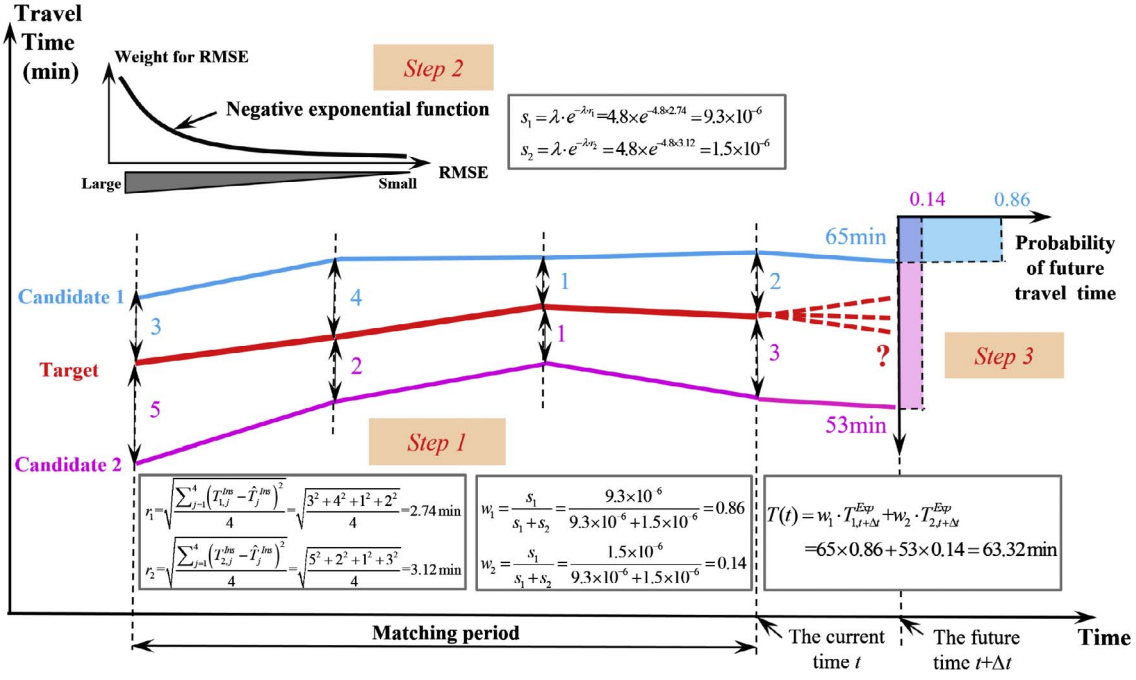

Fig. 5. Simple example of screening process.

profile can be calculated in terms of the root mean square error (RMSE), as shown in Eq. (4).

$$RMSE(r_i) = \sqrt{\frac{\sum_{j=1}^{m} (T_{i,j}^{Exp} - \hat{T}_j^{Exp})^2}{m}} \tag{4}$$

where r_i denotes the RMSE for i^{th} historical candidate, m denotes the number of time intervals Δt in the matching period, $T_{i,j}^{Exp}$ represents the experienced travel time at the jth time interval for the ith historical candidate, and $\hat{T}_j^{Exp}$ represents the experienced travel time at the jth time interval for the target profile.

Step 2: Calculate the similarity function for the screening process

The similarity is assumed to be monotonically decreasing with an increase in the RMSE, for which a negative exponential function is appropriate to describe this relationship, as shown in Eq. (5).

$$s_i = \lambda \cdot e^{-\lambda \cdot r_i} \tag{5}$$

where s_i denotes the similarity between the historical candidate i and the target profile and λ is the coefficient of the negative exponential function.

Step 3: Calculate the weights of each candidate for the future travel time forecast

The weight w_i of the ith candidate for the prediction of the future travel time is defined based on all similarities s_i ($i = 1, 2, \ldots, K$), as shown in Eq. (6).

$$w_i = \frac{s_i}{\sum_{i=1}^{K} s_i} \tag{6}$$

The future travel time $\hat{T}(t + \Delta t)$ can be predicted as a combination of each candidate's experienced travel time $T_{i,t+\Delta t}^{Exp}$ ($i = 1, 2, \ldots, K$) at the interval $t + \Delta t$, as shown in Eq. (7).

$$\hat{T}(t + \Delta t) = \sum_{i=1}^{K} w_i \cdot T_{i,t+\Delta t}^{Exp} \tag{7}$$

4. Case study

To demonstrate the performance of the proposed approach, the probe data that were collected on the urban expressway named Ring 2 in Beijing in a clockwise direction were employed. Ring 2 has a length of approximately 32 km, and the prediction of the total travel time will be conducted every 2 min. The spatiotemporal speed data for 45 days from Jan. 1 to Feb. 14, 2015, which contain 31 weekdays and 14 weekends, are available. To completely utilize the data, two parts of the data are divided for different purposes. In the first part, 43 days of data are allocated to establish the historical database, calibrate the model parameters and adjust the methods. The second part allocates two days, i.e., one typical weekday (Tuesday on Jan. 6) and one typical weekend (New Year's Day on Jan. 1), to evaluate the proposed approach.

4.1. Parameter calibration

The proposed prediction model involves several parameters that affect the prediction. In principle, these parameters, including the temporal size of the pattern (L), the radius of the time constraint windows (R), the number of the best matched candidates (K), and the parameter λ of the negative exponential function, are called Hyperparameters in the context of machine learning. Their optimization is the problem of tuning a set of optimal parameters for a learning algorithm. The parameters are calibrated to identify their optimal combination. The loss function of the model is evaluated in terms of the mean absolute percentage error (MAPE) in the parameter calibration. The MAPE reflects the relative errors of the models, as shown in Eq. (8).

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{T(t) - \hat{T}(t)}{T(t)} \right| \times 100\% \tag{8}$$

where $T(t)$ is the ground truth of the experienced travel time at interval t, $\hat{T}(t)$ is the predicted travel time, and n is the number of observations.

Since MAPE is not an analytical function of these parameters, we apply a simple explanative solving approach for nonlinear optimization. First, the multivariate grid search method is adopted to tune the hyperparameters for part of testing data. It relates to an exhaustive searching through a manually specified subset of the Hyperparameter space of a learning algorithm and aims to avoid the arbitrary and capricious behavior. Specifically, the L parameter, which represents the temporal size of the pattern, varied between 10

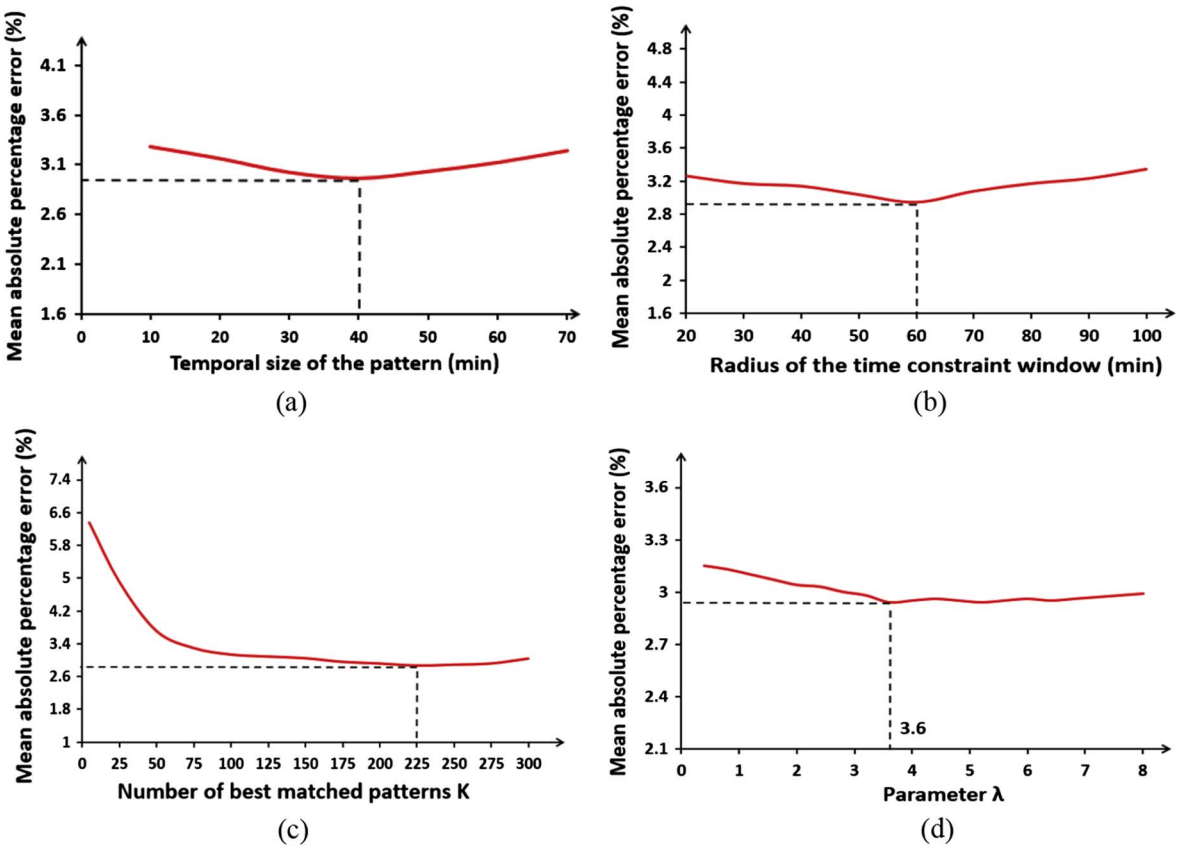

Fig. 6. Influence of the parameters on the proposed method.

to 70 min at 10-min intervals. The size of L should be sufficiently long to replicate the evolution of the traffic states but should not be excessively long, which may cause unnecessary details that mislead similar patterns to be included. To decrease the computation complexity of space and time, R is another parameter representing the radius of the time constraint when searching for best-matched candidates from the historical dataset. The size of R varied between 20 and 100 min at 10-min intervals. The domains of the parameter K (number of best matched patterns) and parameter λ (in the negative exponential function) were also tested from 25 to 300 and 0.4 to 8.0 at 25 and 0.4 increments, respectively.

The number of the combination of all these parameters is $7 \times 9 \times 12 \times 20 = 15{,}120$ in total. To conduct multivariate grid search for all the dataset in a rigorous way is almost impossible due to the enormous computation complexity of space and time. Thus, the calibration of these parameters for optimal prediction performance was first conducted for a weekday and a weekend, respectively, and all of these assignments were implemented with MATLAB in parallelism on the server. However, the calibration results revealed that there is no significant difference between the parameters for the selected weekday and weekend, and the MAPEs of model performance are not that sensitive to calibrated parameters except for K (number of best matched patterns).

To further calibrate the multivariate parameters as well as investigate the impact of individual parameters on the model, we used the control variable method to examine the best combination of parameters based on all the dataset. Every time one parameter was examined by fixing the other three parameters as optimal values according to the grid search results for two selected days as aforementioned. The experiments, which only involve a few combinations of model parameters, were repeated effectively for the dataset of different dates. Such an approach helps relieve the enormous computational efforts of multivariate grid search if directly applied in our case. Specifically, the calibration results are illustrated in Fig. 6 and the detailed analysis is depicted as follows.

4.1.1. Temporal size of the pattern

The temporal size of the pattern (L) is a key parameter in the proposed pattern matching method and it will directly influence the search space in the historical dataset. Fig. 6(a) depicts the influence of L on the proposed method. The MAPEs rapidly decrease with an increase in L and gradually increase when L exceeds 40 min. Thus, the optimal L is determined to be 40 min considering the variability of the MAPEs.

4.1.2. A time constraint window

Traffic patterns are recurrent in nature, however, unlikely recur at the same time. To decrease the computation complexity of space and time, a time constraint window (Zheng and Su, 2014) is introduced in the process of neighbor selection to ensure that the

selected candidates are restricted within a certain time domain. It is assumed that the starting time of a neighbor candidate is reasonably close to the current time. R is the radius of the time constraint window, which represents the range of matching traffic patterns. Fig. 6(b) depicts the relationship between R and the MAPEs. The model performance is optimized when $R = 60$ min. Specifically, it indicates that for the target traffic states at 9:00 similar patterns will be searched on all historical days from 8:00 to 10:00. The changing tendency of the MAPEs also suggests that neither a restrictive R nor a prolonged R yields the desirable prediction accuracy.

4.1.3. Number of best-matched patterns

In the process of pattern matching, an important issue is to determine the appropriate number of best-matched patterns K. Different K values have a significant impact on the method performance. Fig. 6(c) depicts the relationship between the different numbers of K and the MAPEs, which indicates that the prediction accuracy rapidly decreases when K increases to 225 but gradually increases when K exceeds this value. Thus, K is set as 225 in the search space.

4.1.4. Parameter λ of the negative exponential function

To predict the future travel time with K selected candidates, the unknown parameter λ of the negative exponential function needs to be calibrated in the process of weight assignment. Fig. 6(d) depicts the relationship between λ and the MAPEs. When λ increases, the MAPEs rapidly decrease and gradually increase. The method performance is not sensitive to parameter λ when λ is larger than 3.6. To balance the prediction accuracy, λ is set as 3.6.

Overall, it was found that there is only slight deviation between the optimal combination of the parameters derived by the control variable method based on all the dataset and the grid search method based on the dataset of two selected days. It helps demonstrate the effectiveness of the explanative solving approach for nonlinear optimization. The calibrated parameters serve as a good basis for prediction of travel times on urban expressways as will be demonstrated below.

4.2. Intermediate procedures

In principle, the pattern-matching method is regarded as a data-driven prediction approach. One constant criticism of using this type of approach is that it may be efficient but not insightful for control or management because the causal mechanism, i.e., why is the effectiveness of traffic texture characterization, how the model arrives to a prediction, remains unknown. Thus, in this subsection, we focus on unveiling the traffic spatiotemporal features by constructing the stochastic congestion map for bottleneck identification, and visualizing the intermediate procedures of the methodology by providing an in-depth quantitative analysis on the speed pattern matching and examples of matched speed contour plots.

4.2.1. Bottleneck identification

To validate the application of GLCM matrices, we constructed the stochastic congestion map of the Ring 2 in Beijing in a clockwise direction, and analyzed the daily traffic and the recurrent congestion as shown in Fig. 7. The map represents the likelihood of the occurrence of congestion based on the observations for 45 days. To construct the map, we calculated the probability of recurrent congestion by defining the congested links where the corresponding speeds are lower than a threshold (30 km/h). If the probabilities are greater than 0.8, thereby the links are identified as recurrent bottlenecks. Fig. 7(a) shows that the congestion always occurred at the specific spatial locations of the road ring. After filtering by probabilities below 0.8 as illustrated in Fig. 7(b), it is obvious that the bottlenecks occurred at the kilometers of 6–10, 16–20 and 22–24. Due to the recurrent congestions and almost fixed bottlenecks, the essential structure of the constructed spatiotemporal speed map maintained relatively stable and map shifting in the spatial domain rarely occurred. It supports the application of GLCMs to effectively extract traffic patterns.

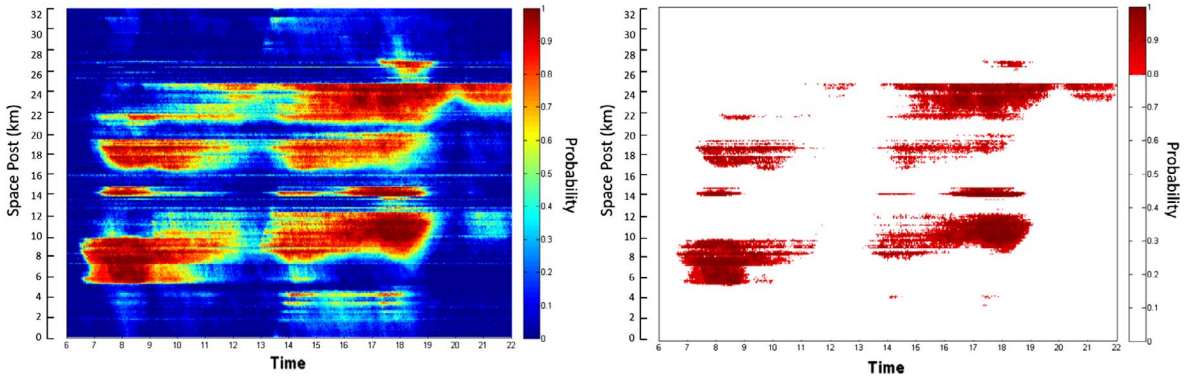

Fig. 7. Stochastic congestion maps of all 45 workdays.

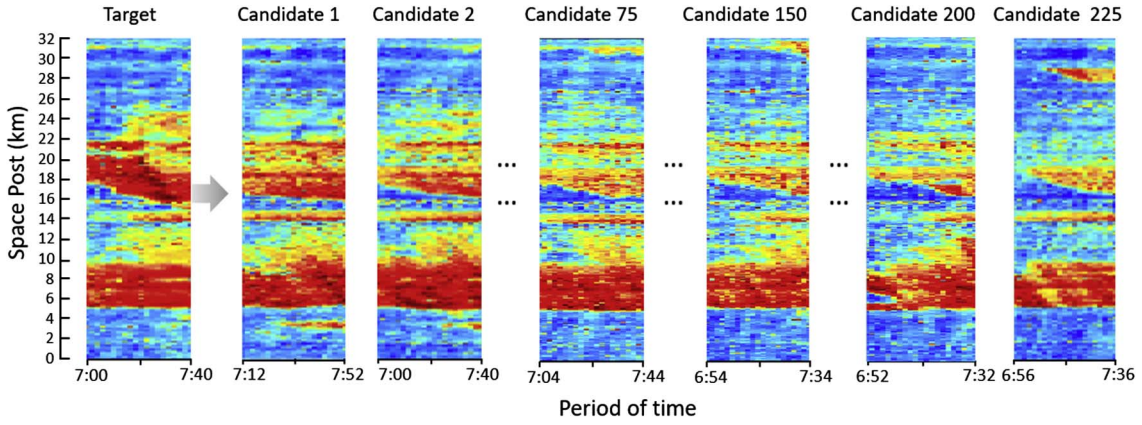

Fig. 8. Illustration of matched traffic patterns.

4.2.2. Quantitative analysis of traffic patterns

The quantitative analysis is conducted as follows. Firstly, we take a peak time period, i.e., 7:00–7:40 on Jan. 6, 2015, as a target profile for pattern matching. In the order of the similarity measurement as proposed in Section 3.2, 225 candidates, i.e. matched speed contour plots, were extracted from the historical database. It indicates that Candidate 1 is most similar to the target profile, Candidate 2 is second similar and so on. For illustration purpose, Candidates 1, 2, 75, 150, 200, 225 and their corresponding traffic patterns are shown in Fig. 8. In general, it can be identified that the selected candidates by matching similar spatiotemporal traffic patterns from the historical database are alike. This also indicates that NSD can be a superior measure of similarity for traffic pattern matching.

Next, Fig. 9 reveals how 225 candidate traffic patterns are distributed in the historical database. It was found that almost all candidate traffic patterns belong to weekdays except Jan. 4, which falls on a Sunday and involves one day shift of work due to the New Year's Day holiday. It is also worthy of noting that the time periods of all 225 traffic patterns are between 6:52–7:52, which is close to the target profile (i.e., 7:00–7:40). This helps verify the effectiveness of the time constraint window as well as the validity of the fundamental assumption that traffic patterns are recurrent in nature.

In this study, the travel time forecast is made by assigning weights to the experienced travel times of each candidate. The details of the process of weight assignment are visualized in Fig. 10 The brown line up and the gray line below represent the experienced travel

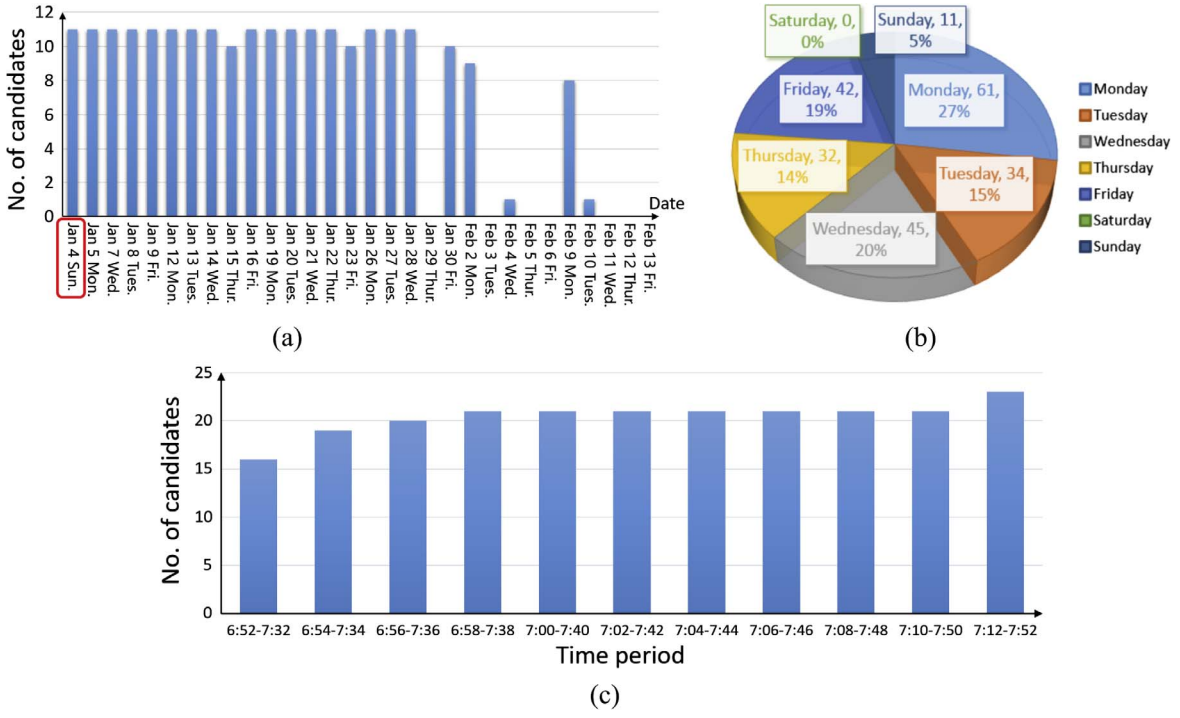

Fig. 9. Distribution of candidate traffic patterns.

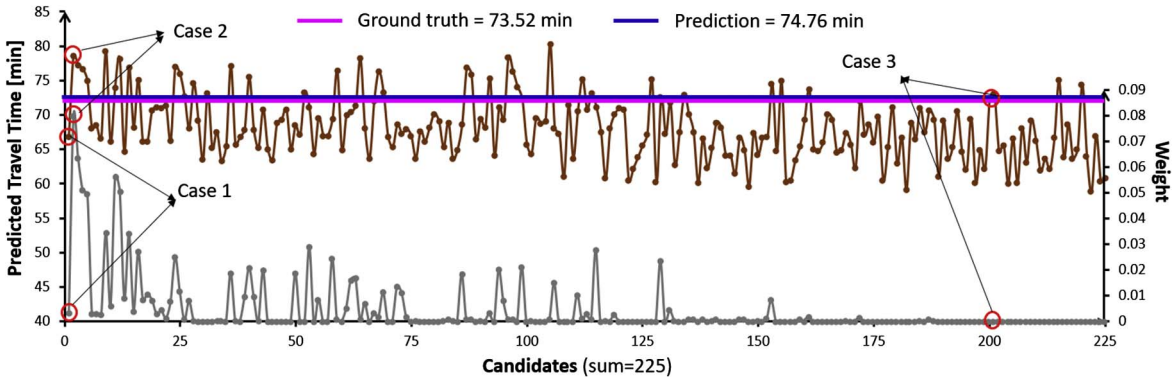

Fig. 10. Details of weight assignment.

times and the corresponding weights of 225 candidates, respectively. It can be found that the weights have an overall trend of gradually decreasing to 0 as the similarity decreases. It is also worthy of noting that the weights are not inversely proportional to the similarity strictly, and the experienced travel times of each candidate fluctuate around the consequence travel time (blue line in Fig. 10). For instance, Candidate 1 is most similar to the target profile but has a lower weight; Candidate 2, which is second similar to the target profile, has a larger weight; Candidate 200, whose corresponding experienced travel time is almost the same to the ground truth, yet has a lower weight. The reasons are explained below.

Since the spatiotemporal traffic states are volatile, the similar traffic patterns to the target profile at current intervals cannot guarantee similar traffic states at subsequent intervals. Thus, there is a need to adjust the contribution of each candidate to the prediction value. As stated in Section 3.3, the discrepancy of the experienced travel times at each interval within the matching period is used as a measurement to examine the similarity of subsequent traffic states. A small discrepancy indicates relatively similar subsequent traffic states. Only the candidates with similar traffic patterns at both current and subsequent intervals will be assigned larger weights and greater contributions to the final prediction results.

To further characterize the potential causality of this issue and provide insights into the principle of weight assignment, three candidates, i.e., Candidates 1, 2, 200, are taken as examples and their subsequent traffic states are visualized as shown in Fig. 11. For illustration purpose, the projected experienced trajectories and corresponding experienced travel times at the beginning of each departure time are also provided. It can be identified that the subsequent traffic pattern of Candidate 1 is not similar to that of the target profile, though they are similar at current intervals as shown in Fig. 8. Thus, Candidate 1 was assigned a lower weight and makes little contribution to the final results, as depicted in Fig. 10. For Candidate 2, it has both similar matched and subsequent traffic patterns to the target profile. Fig. 11 shows that Candidate 2 evidently has a more similar subsequent traffic pattern than Candidate 1, thus assigned a larger weight. In addition, it is worth noting that in despite of similar experienced travel time as in Fig. 11, Candidate 200 has a lower weight value. It has dissimilar traffic patterns and subsequent traffic patterns, leading to less influence on the prediction result. The final prediction result of 74.76 min is close to the ground truth, i.e., 73.52 min as shown in Fig. 11. The accuracy of the prediction owes to the effective mechanism of weight assignment.

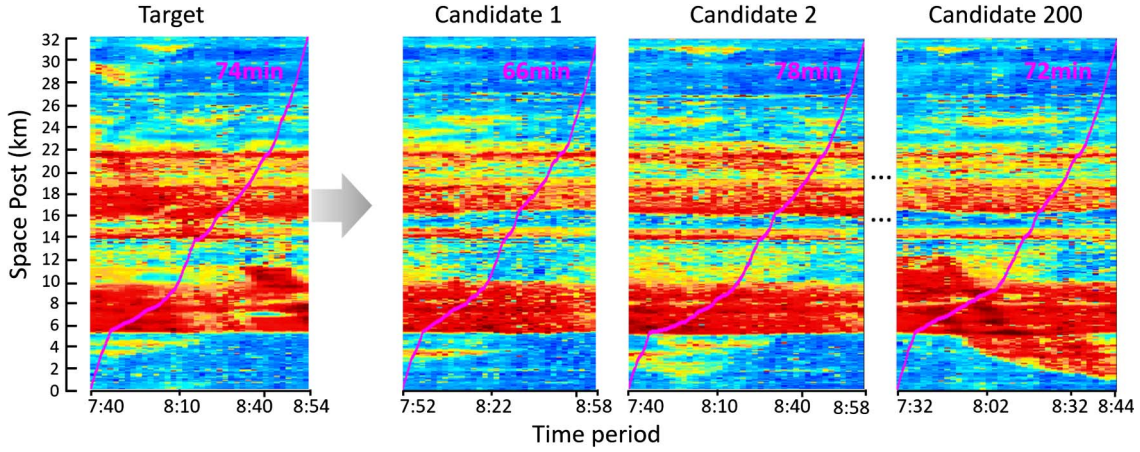

Fig. 11. The subsequent traffic states of representative candidates.

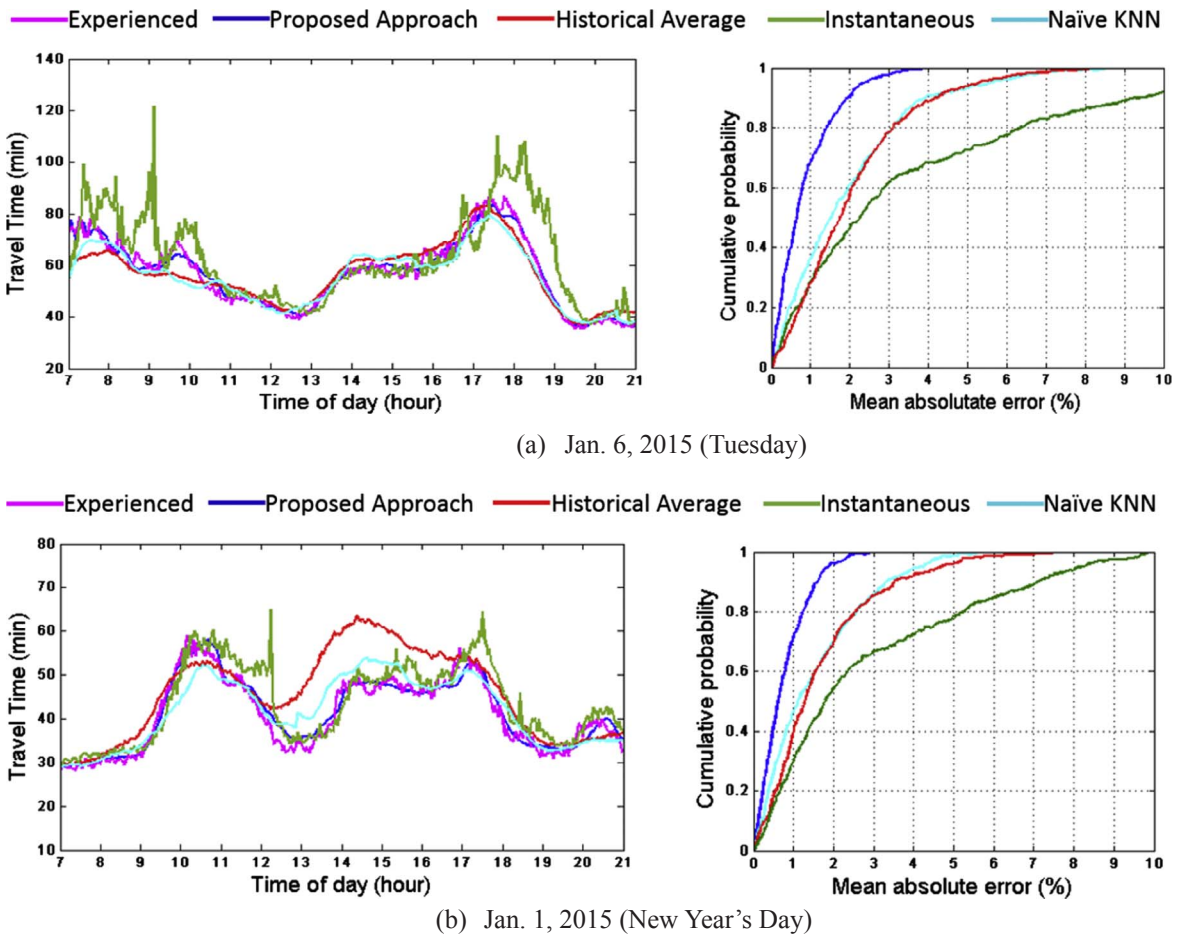

Fig. 12. Comparison of travel time prediction results by different methods.

4.3. Results and discussion

4.3.1. Comparison methods and performance indices

To rigorously evaluate the performance of the proposed approach, several different approaches are considered and compared using the same dataset, including experienced travel times, Historical Average, Instantaneous travel times and Naive KNN. Experienced travel times are considered to be the ground truth. Historical Average travel times are computed with each day of the corresponding historical database, that is, the mean value of the experienced travel times at a given departure time in all days of the corresponding historical database will serve as the historical average prediction. The instantaneous travel time method is also included for comparison. It is the easiest alternative to predict future travel times by assuming that the current traffic speeds along all segments remain constant until the completion of the trip. To demonstrate the advantage of the proposed approach, the Naïve KNN method, as a classic pattern-matching approach, is applied. It also employs the GLCMs and NSD to select candidates or nearest neighbors; however, the simple weighted average of the experienced travel times for each candidate was calculated for prediction.

Fig. 12 presents the travel times predicted by each approach for two selected days. The proposed approach accurately predicts travel time and outperforms the other approaches during all periods. For the Historical Average, it is based on historical data. Though the historical tendency can be revealed to some extent, it has poor flexibility in complex and changeable traffic conditions. For Naive KNN, a simple weighted average of all candidates does not achieve desirable accuracy. Naive KNN predictions experience a temporal lag relative to the ground truth data, especially during the periods of congestion formation and dissipation. The instantaneous travel time method fails to consider the temporal-spatial evolution of the traffic states, significantly underestimates the ground truth when congestion is forming, and overestimates the ground truth when congestion is dissipating.

Specifically, Fig. 12(a) presents the prediction results for a typical weekday. Traffic congestion is significant during the morning and afternoon peak periods. The experienced travel times during peak hours (about 80 min) can be nearly two times those during non-peak periods (about 40 min). The prediction results by the proposed approach track closely to the experienced travel times. In spite of this, there exists small deviations from the ground truth under transient states when congestion is forming or dissipating. Fig. 12(b) presents the prediction results for a typical weekend of the New Year's Day. Unlike traffic characteristics on weekday, the traffic congestion on weekend appears lighter, with the morning peak postponed until 10:00 and the afternoon peak extended from

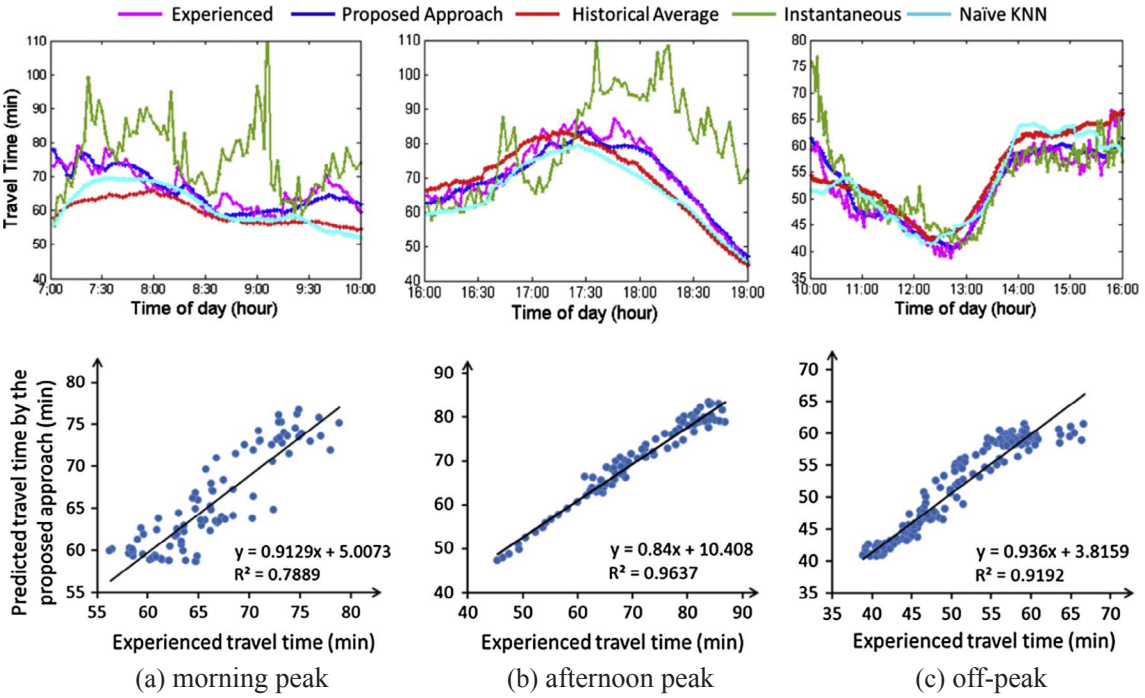

Fig. 13. Comparison of travel time prediction results during different periods on Jan. 6, 2015.

15:00 to 17:00. The Historical Averages significantly deviate from the ground truth while the proposed approach accurately predicts the changing tendency of the experienced travel times.

To further demonstrate the performance of the proposed approach, Fig. 13 presents the prediction results separately for peak and off-peak periods. The morning peak and the afternoon peak are 7:00–10:00 and 16:00–19:00, respectively, while the off-peak is 10:00–16:00. Given that the traffic states change abruptly during two peak periods, Historical Average, Instantaneous and Naïve KNN methods perform barely satisfactory. By comparison, the proposed approach successfully captures the traffic dynamics along the site.

To assess the effectiveness of the methods, the performance criteria are specified using two statistics, namely, the MAPE, as shown in Eq. (8), and the Mean Absolute Error (MAE), as shown in Eq. (9).

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |T(t) - \hat{T}(t)| \tag{9}$$

where $T(t)$ is the ground truth of the experienced travel time at interval t, $\hat{T}(t)$ is the predicted travel time, and n is the number of observations.

Table 1 presents the evaluation results for different methods. It can be found the lowest prediction errors are achieved by the proposed approach. The Historical Average and the instantaneous method provide the worst performance since neither of the methods are able to accurately capture the changing tendency of the experienced travel times. The Naive KNN produces marginally higher prediction errors. In addition, the overall errors for off-peak periods are smaller than those for peak periods on weekday. However, it is not the case on the typical weekend, i.e. the New Year's Day. The possible reason is that the traffic states on the New Year's day are somewhat different from those on the normal weekends, thus few similar traffic patterns can be identified in the limited historical database.

Table 1
Evaluation of travel time forecast by different methods.

Date	Period	Index	Proposed	Historical Average	Instantaneous	Naive KNN
Jan 6, 2015 (Tuesday)	Peak	MAE (min)	2.16	5.53	12.23	5.36
		MAPE (%)	3.12	7.95	18.69	7.56
	Off-peak	MAE (min)	1.54	3.19	3.28	3.21
		MAPE (%)	2.93	6.22	6.41	6.05
Jan 1, 2015 (New Year's Day)	Peak	MAE (min)	1.51	4.44	3.97	2.88
		MAPE (%)	3.09	9.54	8.45	5.82
	Off-peak	MAE (min)	1.88	7.98	3.26	3.81
		MAPE (%)	5.12	20.64	8.85	10.08

Table 2
Evaluation of multi-step-ahead travel time forecast by different methods on Jan. 6, 2015.

		Forecasting horizon (min)							
		4	8	10	20	30	40	50	60
Proposed	MAE (min)	2.01	2.29	2.41	2.74	3.21	3.54	3.78	4.06
	MAPE (%)	3.41	3.86	4.06	4.67	5.48	6.11	6.51	6.97
Historical Average	MAE (min)	4.05	4.01	3.99	3.88	3.79	3.64	3.55	3.53
	MAPE (%)	6.97	6.94	6.93	6.81	6.65	6.44	6.31	6.24
Instantaneous	MAE (min)	7.36	7.31	7.32	7.31	7.15	7.29	7.51	7.67
	MAPE (%)	12.47	12.23	12.15	11.83	11.45	11.66	12.24	12.98
Naïve KNN	MAE (min)	3.91	3.95	3.98	4.03	4.09	4.08	4.02	4.04
	MAPE (%)	6.46	6.57	6.63	6.73	6.88	6.95	6.89	6.93

4.3.2. Performance of the proposed approach for multi-step-ahead forecast

The proposed approach can also provide multi-step-ahead travel time forecasts. Predictions that cover a forecasting horizon of 60 min in advance are conducted. Different methods are tested for the weekday of Jan. 6, 2015 based on the same historical database as the initial models that were developed for the single-step travel time forecast. The results are summarized in Table 2 and Fig. 14, which shows MAE and MAPE for the evaluation of multi-step-ahead travel time forecast by different methods. With an increase in the forecasting horizon, the accuracy of the multi-step forecast slightly decreases. Compared with the other three methods, the changing tendency of MAE and MAPE indicates that the proposed approach evidently provides a higher forecast accuracy within a short forecasting horizon, e.g., 30 min. By efficiently excavating the historical datasets, the proposed approach enables the adaption of the majority of the given traffic states on the site.

4.3.3. Performance of the proposed approach with low resolution data

The forecast accuracy of the proposed approach is highly related to the resolution of the data. Low resolution data sources create a significant challenge for achieving desirable accuracy. Thus, the robustness of the prediction approach with low resolution data is important. In this study, the probe data were processed into a spatiotemporal traffic speed diagram with a resolution of every 2 min. These high resolution data can accurately reflect the real traffic states, which serves as a solid basis for future travel time forecasts. To investigate the robustness of the approach with low resolution data, four scenarios are established. The original dataset on Tuesday, Jan. 6, 2015 were aggregated into resolutions of 4 min, 6 min, 8 min and 10 min, respectively, by computing the average speeds of all GPS data within each cell along the expressway. Note that the forecasting horizons for multi-step-ahead are different due to different data resolutions. For example, for the data resolutions of 4 min and 6 min, the forecasting horizons for two-step-ahead will be 2×4 min $= 8$min and 2×6 min $= 12$ min, respectively.

Table 3 and Fig. 15 provide the forecast results by different methods for four scenarios. The forecast accuracy by each method slightly decreases as the data resolution decreases. However, even under a data resolution of 10 min, the proposed approach achieves a desirable accuracy of 2.45 min and 4.45% in terms of MAE and MAPE, respectively. Among the methods for comparison, the proposed approach outperforms the other three methods. When the proposed approach is extended to multi-step-ahead forecast under different data resolutions, the results are presented in Table 4 and Fig. 16, which help demonstrate the effectiveness and robustness of the method.

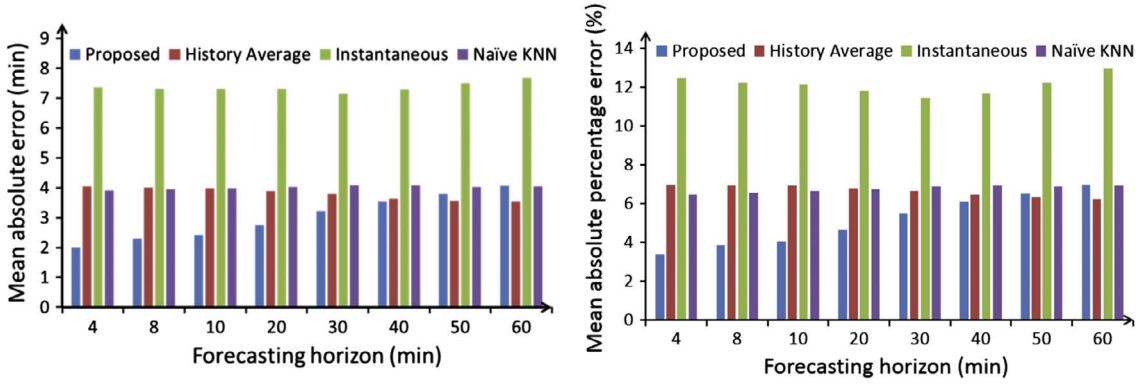

Fig. 14. Comparison of multi-step-ahead travel time forecast by different methods on Jan. 6, 2015.

Table 3
Evaluation of travel time forecast under different data resolutions.

Data resolution (min)		Proposed	Historical Average	Instantaneous	Naïve KNN
2	MAE (min)	1.70	4.10	7.41	3.85
	MAPE (%)	2.89	7.02	12.73	6.33
4	MAE (min)	1.77	4.12	6.51	4.31
	MAPE (%)	3.01	6.85	11.01	7.06
6	MAE (min)	2.07	4.11	6.44	4.49
	MAPE (%)	3.47	6.85	10.91	7.42
8	MAE (min)	2.23	4.01	6.33	4.61
	MAPE (%)	3.81	6.63	10.79	7.64
10	MAE (min)	2.45	3.98	6.21	4.78
	MAPE (%)	4.25	6.77	10.63	8.04

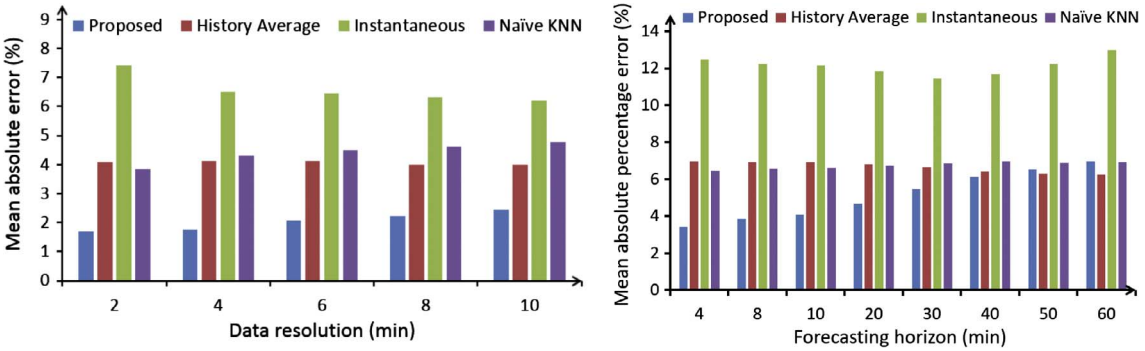

Fig. 15. Comparison of travel time forecasting under different data resolutions.

Table 4
Evaluation of multi-step-ahead travel time forecast under different data resolutions.

Data resolution (min)		Forecasting horizon (steps)							
		1	2	3	4	5	6	7	8
4	MAE (min)	2.06	2.28	2.48	2.61	2.77	2.95	3.12	3.23
	MAPE (%)	3.51	3.91	4.27	4.53	4.84	5.18	5.55	5.72
6	MAE (min)	2.34	2.59	2.89	3.19	3.51	3.74	3.91	4.11
	MAPE (%)	3.98	4.45	5.02	5.57	6.15	6.61	6.94	7.34
8	MAE (min)	2.62	2.91	3.21	3.54	3.78	4.04	4.34	4.63
	MAPE (%)	4.46	4.98	5.49	6.08	6.54	7.01	7.58	8.15
10	MAE (min)	2.87	3.21	3.68	4.08	4.58	4.78	4.88	4.96
	MAPE (%)	5.01	5.67	6.51	7.26	8.29	8.52	8.56	8.61

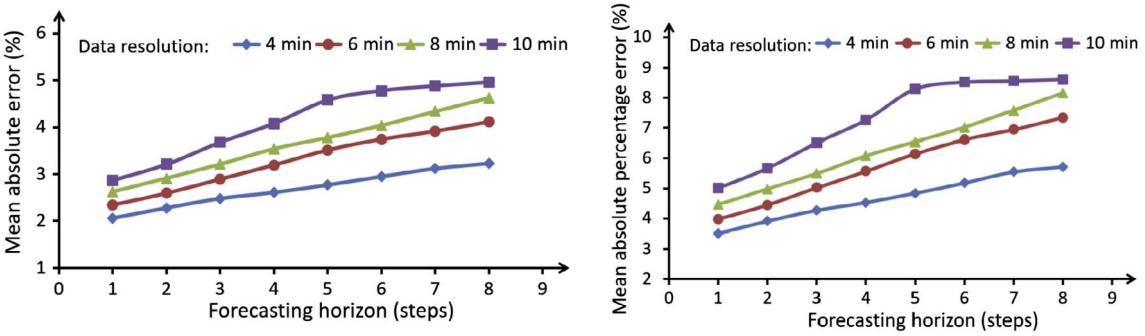

Fig. 16. Comparison of multi-step-ahead travel time forecast under different data resolutions.

5. Conclusions and recommendations for future work

In this study, multiday spatiotemporal speed diagrams were constructed based on vast amounts of probe vehicle data collected in Beijing, China. The GLCM was employed to extract traffic features in both time and space dimensions. Then, the NSD between current and historical GLCMs was calculated as a similarity measure in the matching process to select candidate traffic patterns. Next, a screening process with a time constraint window was implemented for the selection of the best matched candidates. Last, the future travel times for any departure were predicted by aggregating the experienced travel times for each candidate with weight assignment in a negative exponential function. In a case study, the proposed approach was investigated on Ring 2, which is a 32 km urban expressway of Beijing. The results show that the proposed approach, with a minimum number of parameters, provides promising travel time forecasts for varying traffic conditions and different data resolutions, and outperforms the state-of-the-art approaches, such as Historical Average, instantaneous method and Naive KNN. Unlike most of the time series methods, the proposed pattern-matching approach tends to eliminate the lagging behind problems to a large extent by matching large-scale spatiotemporal traffic patterns from the historical database. An in-depth quantitative analysis was provided on the process of speed pattern matching and examples of matched speed contour plots.

The proposed approach provides a framework for utilizing probe data in both temporal and spatial dimensions to predict dynamic travel times on urban expressways. The model can be adjusted in future studies by integrating the data from different sources, e.g., fusing probe data with loop detector data. Because the improvement of forecast accuracy in this study was achieved by changing the weighting rule for pattern extraction and scaling the similarity between historical data and current data, other advanced pattern recognition and data mining techniques are assumed to be incorporated into the proposed approach to effectively extract similar traffic patterns. Note that the application of GLCMs may cause potential problems when the congestions are no longer recurrent in fixed kilometer posts or caused by traffic accidents. To enhance the effectiveness and reliability of the pattern matching approach, the application of GLCMs should be adjusted in future studies by integrating other directions, e.g., fusing the horizontal and diagonal directions with the vertical direction or recognizing and tracking“wide moving jam” and “synchronized flow” in spatiotemporal diagrams (Kerner et al., 2004). It will support characterizing the spatiotemporal relationships on any place of a diagram. Other research priorities include predicting travel time distributions for different departure times to improve the travel planning and route choice of road users. Since historical datasets only include 45 days of previous traffic states, the model performance can be tested based on increasing amounts of data.

Acknowledgements

The authors appreciate the National Natural Science Foundation of China (No. 51308475 and No. U1564212) for support of this research.

References

Bajwa, S.I., Kuwahara, M., 2003. A Travel Time Prediction Method Based on Pattern Matching Technique. Publication of ARRB Transport Research Limited.
Bajwa, S.I., Chung, E., Kuwahara, M., 2005. Performance evaluation of an adaptive travel time prediction model. Intell. Transport. Syst. Proc. 2005, 1000–1005.
Billings, D., Yang, J.S., 2006. Application of the ARIMA models to urban roadway travel time prediction – a case study. IEEE Int. Conf. Syst., Man Cybern. 2006, 2529–2534.
Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., Sun, J., 2016. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. Transport. Res. Part C: Emerging Technol. 62, 21–34.
Celikoglu, H.B., 2013a. Flow-based freeway travel-time estimation: a comparative evaluation within dynamic path loading. IEEE Trans. Intell. Transp. Syst. 14 (2), 772–781.
Celikoglu, H.B., 2013b. Reconstructing freeway travel times with a simplified network flow model alternating the adopted fundamental diagram. Eur. J. Oper. Res. 2 (228), 13–21.
Chen H., Rakha H., Sadek S., and Katz B., 2012. Particle filter approach for real-time freeway traffic state prediction. Transportation Research Board 91st Annual Meeting, Washington, D.C.
Chen, P., Ding, C., Lu, G., Wang, Y., 2016. Short-term traffic states forecasting considering spatial-temporal impact on an urban expressway. Transport. Res. Record: J. Transport. Res. Board 2594, 61–72.
Chen, M., Yu, G., Chen, P., Wang, Y., 2017a. A copula-based approach for estimating the travel time reliability of urban arterial. Transport. Res. Part C: Emerging Technol. 82, 1–23.
Chen, P., Sun, J., Qi, H., 2017b. Estimation of delay variability at signalized intersections for urban arterial performance evaluation. J. Intell. Transport. Syst.: Technol., Plann., Oper. 21 (2), 94–110.
Coifman, B., 2002. Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transport. Res. Part A: Policy Practice 36 (4), 351–364.
Fei, X., Lu, C., Liu, K., 2011. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transport. Res. Part C: Emerging Technol. 19 (6), 1306–1318.
Habtemichael, F.G., Cetin, M., 2015. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. Transport. Res. Part C: Emerging Technol. 66, 61–78.
He, Z., Zheng, L., Chen, P., Guan, W., 2017. Mapping to cells: a simple method to extract traffic dynamics from probe vehicle data. Computer-Aided Civil Infrastruct. Eng. 32 (3), 252–267.
Hisham M.B., Yaakob S.N., Raof R.A.A., and Nazren A.B.A., 2015. Template matching using sum of squared difference and normalized cross correlation. IEEE Student Conference on Research and Development.
Jenelius, E., Koutsopoulos, H.N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. Transport. Res. Part B: Methodol. 53, 64–81.
Kamarianakis, Y., Prastacos, P., 2005. Space-time modeling of traffic flow. Comput. Geosci. 31 (2), 119–133.
Karlaftis, M.G., Vlahogianni, E.I., 2009. Memory properties and fractional integration in transportation time-series. Transport. Res. Part C: Emerging Technol. 17 (4), 444–453.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transport.

Res. Part C: Emerging Technol. 19 (3), 387–399.
Kasai M. and Uchiyama H., 2010. A study on estimation of probabilistic changing travel time based on bayesian statistics. In: Proceedings of the 12th World Conference on Transport Research (WCTRs).
Kasai, M., Warita, H., 2014. Refinement of pattern-matching method for travel time prediction. Int. J. Intell. Transport. Syst. Res. 13 (2), 1–11.
Kerner, B.S., Rehborn, H., Aleksic, M., Haug, A., 2004. Recognition and tracking of spatial-temporal congested traffic patterns on freeways. Transport. Res. Part C: Emerging Technol. 12 (5), 369–400.
Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, D., Van Lint, J.W.C., 2011. Prediction intervals to account for uncertainties in travel time prediction. IEEE Trans. Intell. Transport. Syst. 12 (2), 537–547.
Kobrin, D., 2011. Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system. Transport. Res. Record: J. Transport. Res. Board 2256, 51–59.
Lam, S.H.M., Toan, T.D., 2008. Short-term travel time prediction using support vector regression. Transportation Research Board 87th Annual Meeting, Washington D.C.
Li, R., Rose, G., Sarvi, M., 2006. Evaluation of speed-based travel time estimation models. J. Transport. Eng. 132 (7), 540–547.
Lint, J.W.C.V., 2010. Empirical evaluation of new robust travel time estimation algorithms. Transport. Res. Record: J. Transport. Res. Board 2160, 50–59.
Lint, J.W.C.V., Zijpp, N.J.V.D., 2003. Improving a travel-time estimation algorithm by using dual loop detectors. Transport. Res. Record: J. Transport. Res. Board 1855, 41–48.
Lint, J.W.C.V., Hoogendoorn, S.P., Zuylen, H.J.V., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transport. Res. Part C: Emerging Technol. 13 (5–6), 347–369.
Lint, J.W.C.V., 2008. Online learning solutions for freeway travel time prediction. IEEE Trans. Intell. Transp. Syst. 9 (1), 38–47.
Liu, H., 2008. Travel time prediction for urban networks. TRAIL dissertation Series no. T2008/12, The Netherlands TRAIL Research School.
Mori, U., Mendiburu, A., Álvarez, M., Lozano, J.A., 2015. A review of travel time estimation and forecasting for advanced traveller information systems. Transportmetrica A: Transport Sci. 11 (2), 119–157.
Nixon, M., Aguado, A.S., 2008. Feature Extraction and Image Processing, Second Edition. Elsevier.
Oh, S., Byon, Y., Jang, K., Yeo, H., 2015. Short-term travel-time prediction on highway: a review of the data-driven approach. Transport Rev. 35 (1), 4–32.
Rahmani, M., Koutsopoulos, H.N., 2013. Path inference from sparse floating car data for urban networks. Transport. Res. Part C: Emerging Technol. 30, 41–54.
Song, G., Zhang, F., Liu, J., Yu, L., 2015. Floating car data-based method for detecting flooding incident under grade separation bridges in Beijing. IET Intell. Transport Syst. 9 (8), 817–823.
Sun, L., Yang, J., Mahmassani, H., 2008. Travel time estimation based on piecewise truncated quadratic speed trajectory. Transport. Res. Part A: Policy Practice 42 (1), 173–186.
Theodoridis, S., Koutroumbas, K., 2008. Pattern recognition, Fourth ed. Elsevier.
Treiber, M., Kesting, A., 2013. Traffic flow dynamics: data. Elsevier, Models and Simulation.
Tu, H., van Lint, J., van Zuylen, H.J., 2008. Travel time reliability model on freeways. Transportation Research Board 87th Annual Meeting, Washington, D.C.
Wang, Y., Papageorgiou, M., Messmer, A., 2008. Real-time freeway traffic state estimation based on extended Kalman filter: adaptive capabilities and real data testing. Transport. Res. Part A: Policy Practice 42 (10), 1340–1358.
Weijermars, W., Van Berkum, E., 2005. Analyzing highway flow patterns using cluster analysis. Intell. Transport. Syst. Proc. 308–313.
Xia, J., Huang, M.C.W., 2011. A multistep corridor travel-time prediction method using presence-type vehicle detector data. J. Intell. Transport. Syst. 15 (2), 104–113.
Xiong, Z., Rey, D., Mao, T., Liu, H., 2014. A three-stage framework for motorway travel time prediction. IEEE Int. Conf. Intell. Transport. Syst. 816–821.
Yeon, J., Elefteriadou, L., Lawphongpanich, S., 2008. Travel time estimation on a freeway using discrete time Markov chains. Transport. Res. Part B: Methodol. 42 (4), 325–338.
Yildirimoglu, M., Geroliminis, N., 2013. Experienced travel time prediction for congested freeways. Transport. Res. Part B: Methodol. 53 (4), 45–63.
Zeng, X., Zhang, Y., 2013. Development of recurrent neural network considering temporal-spatial input dynamics for freeway travel time modeling. Computer-Aided Civil Infrastruct. Eng. 28 (5), 359–371.
Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transport. Res. Part C: Emerging Technol. 58, 308–324.
Zhang, X., Rice, J.A., 2003. Short-term travel time prediction. Transport. Res. Part C: Emerging Technol. 11 (3–4), 187–210.
Zhao, N., Yu, L., Zhao, H., Guo, J., Wen, H., 2009. Analysis of traffic flow characteristics on ring road expressways in Beijing using floating car data and remote traffic microwave sensor data. Transport. Res. Record: J. Transport. Res. Board 2124, 178–185.
Zheng, Z., Su, D., 2014. Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transport. Res. Part C: Emerging Technol. 43, 143–157.
Zhu, Y., Li, Z., Zhu, H., Li, M., Zhang, Q., 2013. A compressive sensing approach to urban traffic estimation with probe vehicles. IEEE Trans. Mob. Comput. 12 (11), 2289–2302.
Zou, Y., Zhu, X., Zhang, Y., Zeng, X., 2014. A space-time diurnal method for short-term freeway travel time prediction. Transport. Res. Part C: Emerging Technol. 43 (1), 33–49.