IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 1, 2017, accepted December 27, 2017, date of publication January 1, 2018, date of current version March 13, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2788639

Traffic State Spatial-Temporal Characteristic Analysis and Short-Term Forecasting Based on Manifold Similarity

QINGCHAO LIU[1,2], YINGFENG CAI[1], HAOBIN JIANG[1], XIAOBO CHEN[1], AND JIAN LU[2]

[1]Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China
[2]Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing 210096, China

Corresponding author: Qingchao Liu (autoits@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0102603, in part by the National Natural Science Foundation of China under Grant U1564201, Grant 61601203, Grant 61403172, and Grant 61773184, in part by China Postdoctoral Science Foundation under Grant 2017M611729, in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 17KJB580003, in part by the Key Research and Development Program of Jiangsu Province under Grant BE2016149, in part by the Key Project for the Development of Strategic Emerging Industries of Jiangsu Province under Grant 2016-1094 and Grant 2015-1084, in part by the Key Laboratory for New Technology Application of Road Conveyance of Jiangsu Province under Grant BM20082061503, and in part by the Jiangsu University Scientific Research Foundation for Senior Professionals under Grant 16JDG046.

ABSTRACT The study on the spatial-temporal characteristics of highway traffic flow is helpful to deeply understand the inherent evolution of highway traffic system and provide a theoretical basis for prediction and control of highway traffic flow. This paper makes an empirical analysis on the spatial-temporal characteristics of highway traffic flow using manifold similarity index and manifold learning technology. The time series of highway traffic flow is converted into the distance series containing manifold features to calculate the manifold distance between multi-section traffic flow data points, which are highly similar to spatial-temporal distribution of traffic flow speed parameters, and then, the levels calibration of traffic state is carried out according to the manifold distance, so as to reveal the distribution rule of spatial-temporal characteristics of highway traffic flow. Its prediction error is obviously lower than the traditional distance measurement method, which has higher accuracy. The research of this paper can provide new ideas and methods to reveal the highway traffic flow evolution and traffic state prediction.

INDEX TERMS Traffic state, spatial-temporal characteristics, prediction, manifold similarity.

I. INTRODUCTION

Aiming at improving safety, high efficiency and comfort of road transportation system, the Intelligent Transportation System (ITS) [1] takes full advantage of available road elementary facilities to solve problems such as traffic jam, the frequent occurrence of traffic accidents, and environmental pollution. As an integral part of ITS, short-term traffic state forecasting is defined as predicting the traffic flow of a target road segment in the next time interval and can be used for traffic signal control, route guidance, congestion mitigation, adaptive ramp metering, and so on. For example, with reliable forecasting data, traffic managers can detect potential risks of unstable traffic conditions early and take the necessary steps to ensure that traffic is functioning properly. For travelers, they can receive the real time and dynamically estimated results about future traffic condition and make a decision on departure time or adjust travel routes before jam formation. Due to the important role of short-term traffic state forecasting in ITS field and extensive application, it has become an interesting topic, attracting more and more researchers.

The goal of short-term traffic state forecasting is to predict the evolution of traffic over a span of time from a few seconds to a few hours [2]. Estimation of traffic state prediction models that are very consistent with actual traffic data are more preferable and valuable in practical applications. In general, traffic state forecasting can be viewed as a learning problem. First, a predictive model is constructed by learning basic traffic patterns from given historical traffic data and then predicting future conditions based on real-time traffic data. Over the past decade, a variety of traffic flow forecasting methods have been proposed. However, accurate and reliable traffic flow forecasting remains a challenging issue because

transport systems are a time-varying and complex system whose current and future evolution largely depends on the interaction between traffic flows [3]. The combination of the availability of a large amount of space-time traffic data with the advances in data analysis techniques has created reasonable predictive accuracy and shorter processing time for short-term traffic forecasts. The analysis and extraction of traffic flow characteristics is an important part of self-organization rules, which is of great significance to traffic flow modeling, prediction, and control [4].

According to Vlahogianni *et al.* [2] and Kong *et al.* [5], the current methods of traffic state prediction can be divided into four categories: Naïve, non-parametric, parametric, and hybrid. Naïve methods are to provide a simple estimation of future traffic, such as the historical average. Parametric methods refer to model-based that require a fixed set of parameter values as part of the statistical equations or mathematical they use, such as macroscopic models, time series analysis [6]. Most of these methods are influenced by their assumption of parametric models and prove to be relatively poor performance under nonlinear and complex traffic conditions [2]. Non-parametric methods are mainly data-driven and combine empirical algorithms to provide predictions. This method is advantageous because of the uncertainty of any estimate model parameters and assumptions they have not involved on the developed model. Many researchers say, the performance of non-parametric model is better than parametric models, because they are more suitable for learning from more complex data and adapt to these data. For example, Van Lint and Van Hinsbergen [7] argued that in the context of traffic prediction, non-parametric methods are the first choice because the input and output traffic variables are noisy and the relationship between them is nonlinear and difficult to understand. Other methods of short-term traffic prediction have implemented a mix of the above approaches [8]. Smith *et al.* [9], Lin *et al.* [10], Lippi *et al.* [11] have provided a comparative analysis of several models. Most studies that make short-term traffic forecasts use standard statistical techniques such as simple, smooth, complex time series analysis and filtering.

Application of smoothing for traffic forecast includes KS (kernel smoothing) [12], SES (simple exponential smoothing) [13], and HES (hybrid exponential smoothing) and neural networks [14]. Some investigations treat historical traffic flow for the target site as a time series process. They employ time series analysis theory to stimulate the temporal evolution of traffic flow and predict the future trends. For example, the ARIMA model (Autoregressive Integrated Moving Average) [15], [16]. Seasonal ARIMA (of SARIMA) model has been implemented in a number of studies [8], [10], [11], [17].

Szeto *et al.* [8] proposed a method that a combination of cell transmission and SARIMA models. Filter models are also applied to short-term traffic prediction [18], such as Kalman filtering. Chen and Rakha [19] proposed a traffic prediction method based on particle filters. Another research direction on short-term traffic state forecasting is the application of neural networks and pattern recognition methods. A method based on pattern recognition (a subset of the non-parametric method) seems to be more appropriate since they effectively define similar traffic conditions for producing predictions. Some studies have focused on the realization of neural network and its variants [20]–[26]. Pattern recognition methods were also applied for short-term traffic forecast, e.g., k-nearest neighbor [27], support vector machines (SVM) [28], cluster analysis [29].

In recent years, many researchers have paid more and more attention to the construction of prediction models that integrate more space-time features and even the entire traffic network information. Wang *et al.* [30] proposed an improved synergetic traffic state recognition method based on manifold learning, in which the geometrical structure in high dimensions can be well maintained. Lu *et al.* [31] present a graph embedding algorithm that strikes a balance between local manifold structures and global discriminative information for traffic sign recognition. Lee *et al.* [32] present a method to identify the trajectories of moving vehicles from various viewpoints using manifold learning to be implemented on an embedded platform for traffic surveillance. During training, the extracted features of the training data are projected on to a 2D manifold and feature corresponding to each trajectory are clustered into k clusters, each represented as a Gaussian model. In order to extract the features of people's collective behaviors, Yang and Zhou [33] utilized manifold learning technique referred to as locally linear embedding (LLE) is used to computing the K coefficients in fitting every traffic data point with its K nearest neighbors in the high dimensional space and then the local features of the data points are summarized by using principal component analysis (PCA) to obtain a global feature to represent the traffic data. Yang *et al.* [34] used the boundary model to represent the traffic flow time series data, and the historical time sequence highly similar with current traffic flow time sequence is searched in the historical database. Li *et al.* [35] based on the change trend of traffic flow, five distance metrics of time series of traffic data are designed. According to the similarity of traffic parameter data, Zhang *et al.* [36] proposed a method based on weighted Euclidean distance is used to classify the traffic state. Fractal theory is used to study the self-similarity of traffic flow. In order to show better characteristics in traffic condition on the road network [37], [38], Fu *et al.* [39] divide traffic flow into three levels, and the prediction algorithm is realized based on the neural fuzzy system. By influencing the spatial-temporal characteristic of traffic condition in advance, Dong *et al.* [40] realizes the traffic state level prediction methods based on the maximum entropy model. The existing research shows that there is a correlation between traffic flow data at each collection site, especially for the manifold fluctuation, it needs further research [41], [42].

In summary, KNN approach has been previously applied for the purpose of forecasting traffic state. A number of researchers have applied KNN to forecast traffic flow

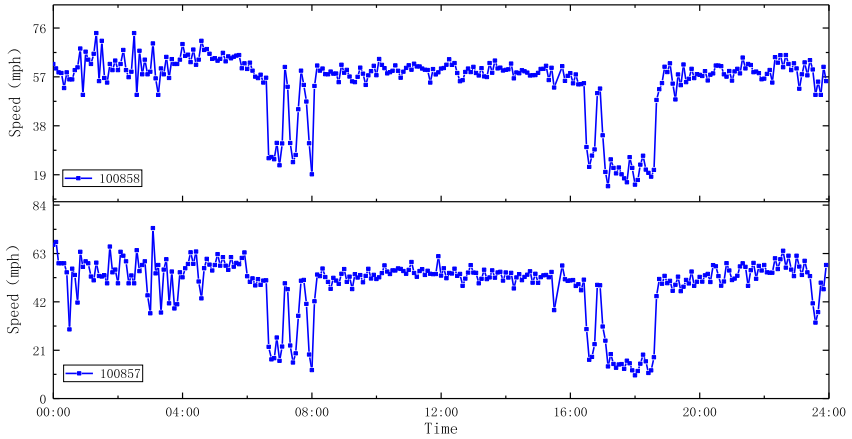

FIGURE 1. The time series of traffic flow data collected by different detectors.

rates [9], [43]. Similarly, other researchers applied KNN for forecasting travel times [44]. The main limitations of the works of these authors are that the simplest form of KNN was used and inclined to predict traffic volume. Lin *et al.* [45] pointed out ''forecasting the traffic volume is unsuitable, because the same volume may correspond to different traffic state.'' Therefore, it is necessary to find a new metric to describe the traffic state. In addition, most of these studies are more ideally premised traffic conditions, they do not consider the manifold features in the extraction and prediction of traffic flow, and there are some difficulties when applied to the actual road. This is a missing component from most of the existing studies in the literature.

Motivated by the above discussions, we design a novel traffic state forecasting model by manifold similarity dealing with the manifold distance metric of KNN. Specifically, we first propose a manifold distance metric of traffic flow data, which can choose candidate data points by the manifold features of traffic flow. And then we construct an improved KNN, which can choose the nearest neighbors by manifold distance. An elaborately designed distance metric is presented to implement the extraction of a spatial-temporal characteristic of traffic flow, it is possible to find the optimal solution for traffic state forecasting problems. In such a way, we can not only determine the suitable nearest neighbors of the target data and the optimal KNN but also fully excavate crucial spatiotemporal information contained in the section of road by training different site data. To evaluate the effectiveness of the proposed method, real-world traffic flow data are collected from 3 observation sites spreading over a freeway called US26 in Portland, OR, USA.

The main contributions of the work are listed as follows. First, we propose the manifold distance in defining the similarity of traffic flow data and design a function to calculate the distance between any two points, the distances between points on the observation manifold are measured along geodesic paths. Second, we utilize the dimensionality reduction strategy to fit every data point with its K nearest neighbors in the manifold feature space and improve KNN that the Euclidean distance is replaced by manifold distance in the search of neighbors. Finally, extensive experiments are conducted to test and compare different metrics, and a traffic model for the short-term forecast is given. The experimental results show that the proposed traffic model can achieve better forecasting accuracy. Moreover, comparing the works of this paper with others which applied KNN approach of the forecast, the details presented in terms of measuring the manifold similarity of traffic flow is new.

The remainder of this paper is organized as follows. Section 2 describes the spatial-temporal characteristic and manifold characteristic of highway traffic flow, which accounts for the flow behavior by a manifold distance metric. Section 3 presents the proposed traffic state forecasting method based on manifold similarity. In this section, we show how traffic data with different spatial-temporal information can be calculated by manifold distance and predicted the future traffic state. Section 4 presents the real traffic data, extensive experiments, and detailed results analysis. Finally, we draw some conclusions in section 5.

II. SPATIAL-TEMPORAL CHARACTERISTIC AND MANIFOLD CHARACTERISTIC OF TRAFFIC FLOW

The time series of traffic flow show some regularity and similarity and present a similar trend in the whole. As shown in Figure 1, the flow data collected by the two detectors of 100857 and 100858 present the evolution of the bimodal, which can be divided into three periods (morning peak, evening peak, and small flow period). In the morning and evening peak, speed decreases significantly while small flow time period is large. The traffic flow data has spatial-temporal similarity, which is the basis of the similarity analysis.

The time series of traffic flow show strong fluctuation as time goes on. Figure 2 (a) depicts the fluctuation and the manifold curve of traffic flow. These manifold features reflect the spatial-temporal characteristics of the traffic flow. In Figure 2 (b) plots traffic flow data collected by detectors

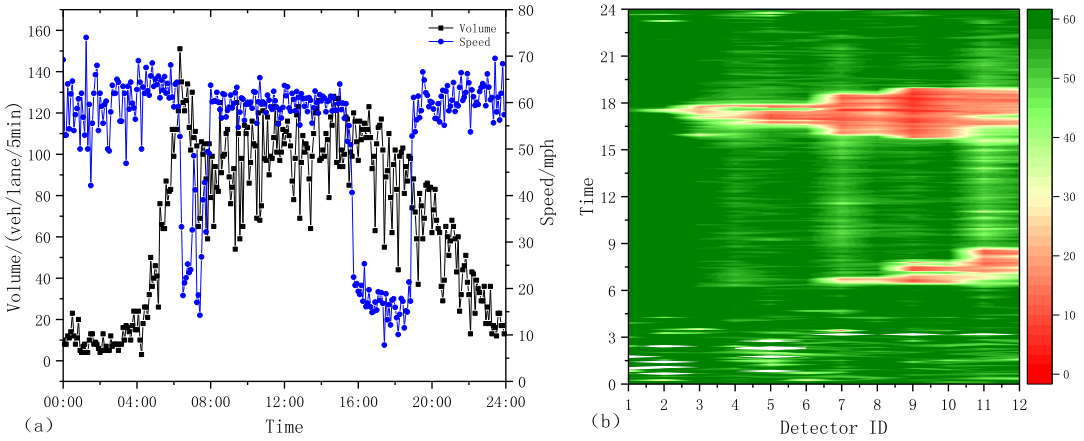

FIGURE 2. A single-section and multi-sections traffic flow diagram.

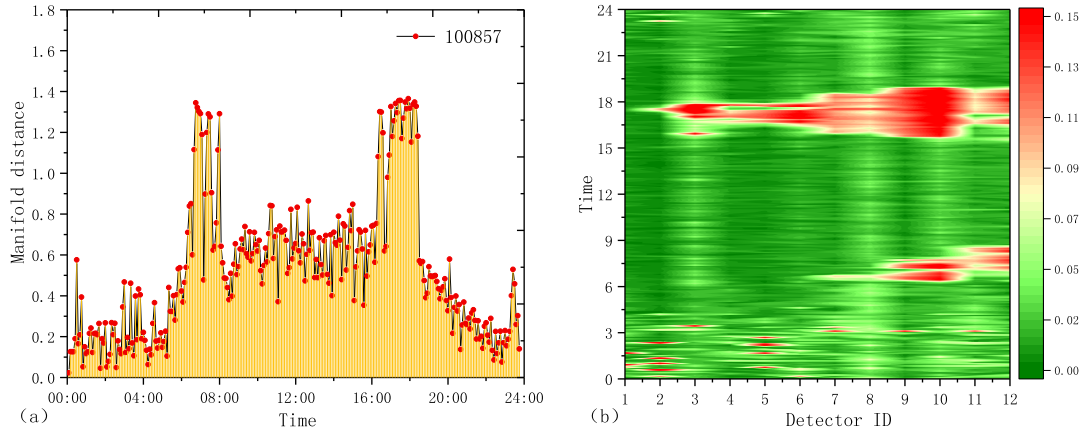

FIGURE 3. The (a) A single section manifold distance; (b) Multi-sections manifold distance.

at different locations, which can show the congestion and dispersion process of traffic flow. Although traffic flow are nonlinear and time-varying, there are some rules in the spatial-temporal evolution, which are mainly shown in two aspects: firstly, the spatial-temporal distribution of traffic flow has the characteristics of aggregation. Secondly, the spatial-temporal evolution of traffic flow is continuous.

Traffic flow data were collected at intervals of 5 minutes for each detector, so the indicator of speed are discrete. In order to describe the spatial-temporal characteristics of traffic flow, this paper adopts manifold distance as the similarity metric (the specific formula is shown in section 3). Figure 3 (a) is detector 100857, which is used to calculates the manifold distance at each 5min interval for the detector 100857 with 0:00 as starting point. From figure 3 (a) and Figure 1, it can be seen that the greater the distance, the more serious the traffic congestion and the lower the speed of vehicles.

Figure 3 (b) plots the manifold distance collected by detectors at different locations. In Figure 3 (b) and Figure 2(b), take the morning and evening peak periods as an example. As shown in Figure 2 (b), when the morning peak reaches the maximum capacity of the location at detector 6, the congestion occurs and gradually dissipates downstream. The congestion condition is easy to dissipate and the congestion level is low. Around 15:30, after congestion occurs at detector 9, capacity falls. With the increasing of vehicle queue moving upstream, congestion occurs and is more severe. Congestion is prone to occur at upstream and dissipates slowly to downstream gradually. The velocity at the time of dissipation is around 30mph, the corresponding manifold distance is about 0.08, and the congestion concentration and dissipation process of Figure 3 (b) is similar to that of Figure 2 (b). The manifold distance intuitively describes the process of traffic flow congestion generation, concentration, dissipation and so on, which is helpful to analyze the spatial-temporal evolution rules of different sections.

III. MANIFOLD SIMILARITY

Similarity measures generally include Euclidean distance, Manhattan distance and so on. However, the traffic flow has

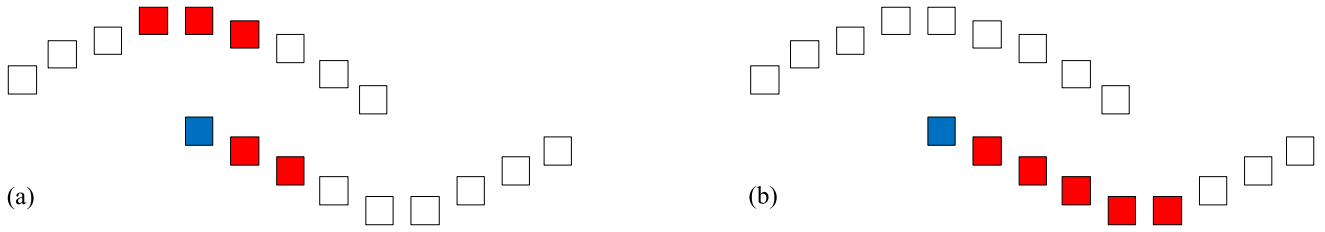

FIGURE 4. (a) the Euclidean nearest neighbors; (b) the manifold nearest neighbors.

the characteristics of manifold fluctuation, common indicators can not accurately describe manifold features. For the selection of the neighborhood points on the manifold, we introduce the manifold distance.

A. MANIFOLD DISTANCE

Definition 1: The length of (x_i, x_j) between two points $L(x_i, x_j)$ in traffic flow is defined as the following

$$L(x_i, x_j) = e^{\frac{d(x_i, x_j)}{\sigma}} - 1 \tag{1}$$

In which $d(x_i, x_j)$ indicates the Euclidean distance between the two points (x_i, x_j) and σ is adjustable parameters (After performing tests in the data preprocessing stage many times, $\sigma = 5$ result is better). The distance between any two points is defined as follows:

Definition 2: If the traffic flow data points are considered as the peak on the graph, and P_{ij} represents all path centration of the connection data points in figure $p \in P_{ij}$, where each path contains k sections, the manifold distance $MD(x_i, x_j)$ between the two point (x_i, x_j) is the minimum path in the graph, which connecting the two points.

$$MD(x_i, x_j) = \min_{p \in P_{ij}} \sum_{k=1}^{|p|-1} L(p_k, p_{k+1}) \tag{2}$$

The manifold distance between any two points of traffic flow data set can be calculated according to the above definition, and the suitable adjacent point can be selected. As shown in Figure 4, blue is the target point, red is for the adjacent point of the target point. Adjacent selection results for Euclidean distance and manifold distance are respectively Figure 4 (a) and (b). Figure 4 (b) selects adjacent points in the direction of the manifold curve according to the local feature of traffic flow.

B. CHARACTERISTICS ANALYSIS OF TRAFFIC STATE BASED ON MANIFOLD SIMILARITY

Traffic state can be reflected through indicators such as speed, flow, occupancy, for example, reflecting the traffic state by speed, and indicating congestion degree that the higher the speed is the smoother of traffic state, whereas the less congestion. Based on the speed index, this paper starts from the manifold similarity of traffic flow to study how the manifold distance reflects the characteristics of traffic state.

Traffic state 1: as shown in figure 3 (b), manifold distance is between 0∼0.05, speed is above 50 mph, the speed difference between upstream and downstream detectors is small, and the whole road covered by the detector is in smooth.

Traffic state 2: manifold distance is between 0.05∼0.067, speed range is 40∼50 mph, the speed difference between upstream and downstream detectors increases and the whole road covered by detectors are basic smooth.

Traffic state 3: manifold distance is between 0.067 ∼ 0.083 and the speed range is 30∼40 mph. In traffic state 3, the traffic flow shows the alternation between smooth and the congestion.

Traffic state 4: manifold distance range is between 0.083∼0.11, and the speed ranges from 15∼30 mph. Traffic state 4 is aggravated by traffic congestion, and the road sections covered by detectors are in moderate congestion.

Traffic state 5: manifold distance range is 0.11∼0.15, and the speed is lower than 15 mph. Queue of vehicle increases and moves upstream, and the whole road covered by the detectors are in severe congestion.

Take morning and evening peaks as an example, traffic state can be divided into four patterns and five traffic state levels. Figure 5 (a) is the morning peak, 5 (b) is evening peak. The white area is for traffic state 1, grey area is for traffic state 2, and in this analogy, 3, 4 and 5 are black, maroon and red. The five colors represent five traffic states, which are smooth, basic smooth, lightly congestion, moderate congestion, and severe congestion.

Pattern 1 indicates the section at this period is smooth, and the adjacent road sections in adjacent period tends to be smooth, in which the traffic state is relatively simple and its corresponding manifold distance and contained traffic state is shown in Table 1; Pattern 2 indicates the section at this period is congestion, but the adjacent road section in adjacent period is smooth, in which congestion occurs because the flow is over the allowable capacity and it moves downstream gradually and such state is easy to dissipate, the congestion state easy to dissipate, and includes traffic state 2, 3; Pattern 3 indicates that this section of the road is congested, and its adjacent space-time is also congested. The traffic condition is mixed with modes 3, 4, 5. Pattern 4 indicates that this section of the road is smooth, but the adjacent road sections are congested in the adjacent period, the traffic state is mixed with 1, 2 and 3, and the congestion level is low and easy to dissipate.

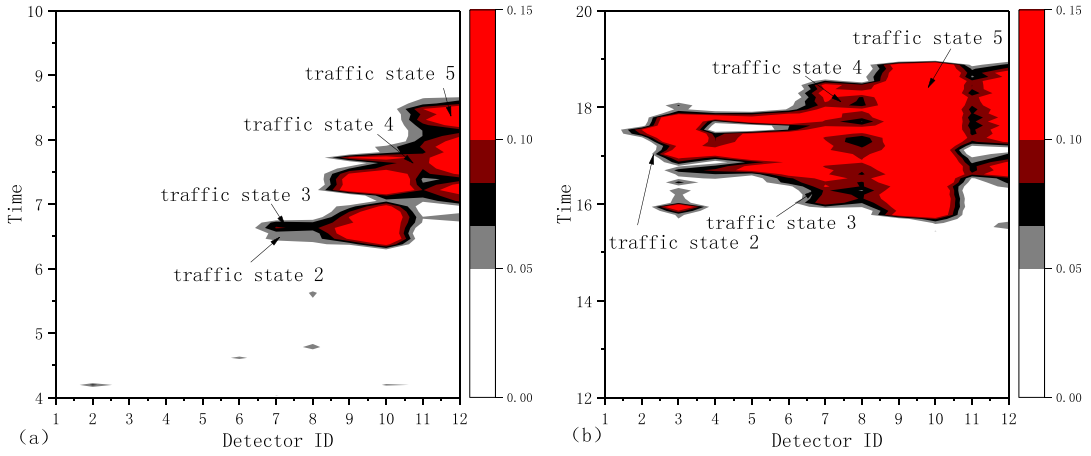

FIGURE 5. Traffic state level corresponding to the range of manifold distance (a) morning peak; (b) evening peak.

TABLE 1. Manifold distance range in four traffic patterns.

Traffic Pattern	Spatial-Temporal Characteristics	Manifold Distance	Traffic State Level
Pattern 1	Space where the detector is smooth while adjacent space is in smooth	0∼0.05	1
Pattern 2	Space where the detector is congestion while adjacent space is in smooth	0.05∼0.083	2,3
Pattern 3	Space where the detector is congestion while adjacent space is in congestion	0.067∼0.15	3,4,5
Pattern 4	Space where the detector is smooth while adjacent space is in congestion	0∼0.083	1,2,3

C. TRAFFIC STATE FORECASTING BASED ON MANIFOLD SIMILARITY

The main ideas of traffic state forecasting based on manifold similarity (TSFMS) are as follows: (1) We analyze the historical traffic flow data and select the characteristic data that can accurately describe the traffic condition, so as to construct Traffic state feature library, which contains multi-section spatial features and various time-varying characteristics of traffic state; (2) The ISOMAP (Isometric Feature Mapping) algorithm is introduced. ISOMAP is used to keep the manifold characteristics of the original data and map it to the reduced data, allowing the neighbor selection to contain the manifold structure of the original traffic flow data. In our model, the multivariate feature of traffic data established in step 1 is reduced to two-dimensional space as model input data; (3) We fit every data point with its K nearest neighbors in the manifold feature space and improve KNN that the Euclidean distance is replaced by manifold distance in the search of neighbors. (4) In the prediction stage, we utilize KNN which is improved by manifold distance to select a group of clusters closest distance of samples to be predicted as the prediction state. Figure 6 shows the technical roadmap of the proposed method.

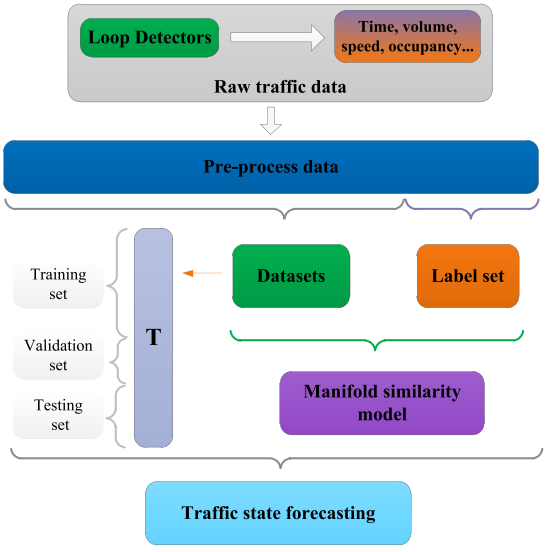

FIGURE 6. The roadmap of TSFMS.

D. MEASURING PERFORMANCE

Three measures are used as indicators of the accuracy of the short-term traffic state prediction method, as shown in

TABLE 2. The link number and data integrity information.

Detection ID	Station ID	Lane	Highway name	Milepost	Data integrity
100857	1085	3	US 26	65.9	Yes
100858	1085	2	US 26	65.9	Yes
100859	*1085*	*1*	*US 26*	*65.9*	*No*
100860	1086	3	US 26	67.4	Yes
100861	1086	2	US 26	67.4	Yes
100862	*1086*	*1*	*US 26*	*67.4*	*No*
100863	1087	3	US 26	68.55	Yes
100864	1087	2	US 26	68.55	Yes

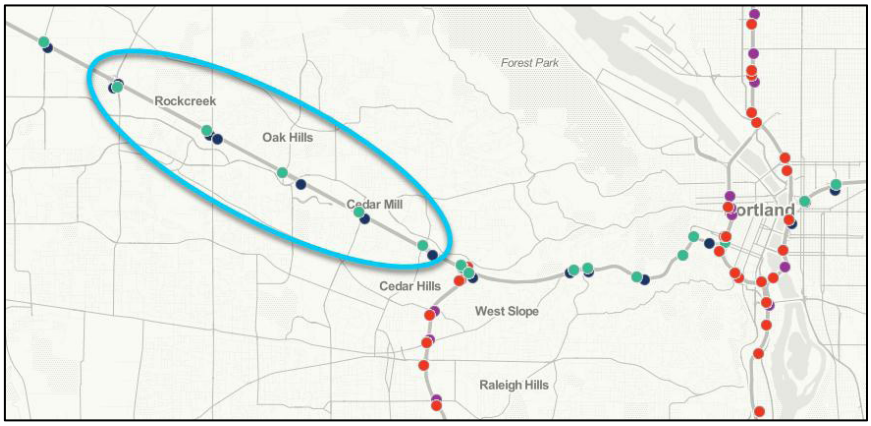

FIGURE 7. Traffic flow data collection location map.

equations (3)–(5): Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Equality Coefficient (EC). These measures of performance provide a deep understanding of the nature of the prediction errors. For example, MAPE provides prediction error based on the percentage difference between observed and predicted flow rates, and RMSE provides prediction error based on the number of vehicles.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|(Y_i - \hat{Y}_i/Y_i)\right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{4}$$

$$EC = 1 - \sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}/(\sqrt{\sum_{i=1}^{n}Y_i^2} + \sqrt{\sum_{i=1}^{n}\hat{Y}_i^2}) \tag{5}$$

Where Y_i is the ith observed value, $\hat{Y}$ is the ith forecast value, n is the number of samples.

IV. EMPIRICAL ANALYSIS

A. EXPERIMENTAL DATA

In this study, Traffic flow data collected by eight loop detectors for 31 days, from Dec. 1st to Dec. 31st in the year 2016 is used in the experiments as shown in Figure 7. These loop detectors were installed around the U.S. Route 26 (US-26) in Portland, Oregon, USA. US-26 is a major cross-state state highway in the U.S. state of Oregon, connecting U.S. Route 101 on the Oregon Coast near Seaside with the Idaho state line east of Nyssa. The Portland Oregon Regional Transportation Archive Listing (PORTAL) has been developed by Portland State University in cooperation with the Oregon Department of Transportation (ODOT) and other regional transportation agency partners. The identification numbers (IDs) for these loop detectors are 100857, 100858, 100859, 100860, 100861, 100862, 100863 and 100864. As shown in Table 2, 100859 and 100862 collected traffic flow data are seriously missing, so we did not use. The aggregation period is 5 minutes, thus leading to 288 sample points per day. The total number of sample points is $288 \times 6 \times 31 = 53568$. For each station, the first fifteen days (Dec.1st to Dec.15st) are used as a training period, the following five days (Dec.16st to Dec.20st) are used as a validation period for model selection and parameter tuning, and the remaining eleven days (Dec.21st to Dec.31st) are used as a test period.

B. VARIABLE ESTIMATION FOR TSFMS TRAFFIC STATE FORECASTING

For the proposed TSFMS algorithm, several variables must be predetermined so that the prediction error is as small as

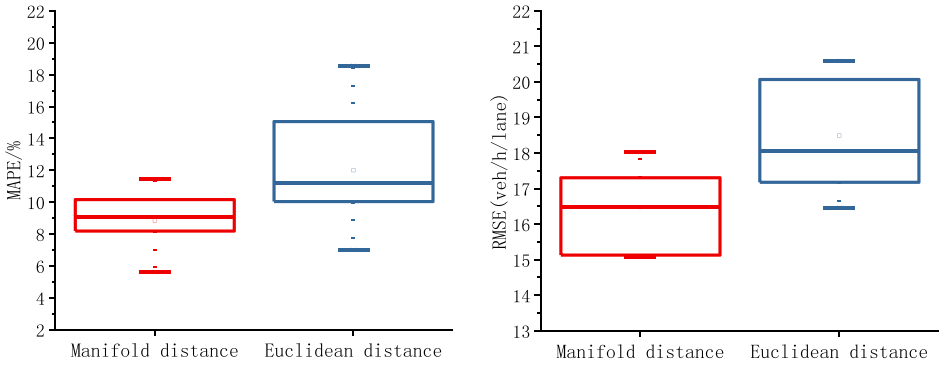

FIGURE 8. Comparison of Manifold and Euclidean distance metric.

possible. These variables include in comparing the appropriate distance measure, lag time, and nearest candidates.

1) COMPARISON OF DISTANCE MEASURE

As discussed previously, two distance measures were considered for searching the nearest candidates. Figure 8 shows the average prediction error corresponding to the above distance measurement, namely: (1) Manifold distance and (2) Euclidean distance. The twenty-one nearest neighbors with two hours' lag time for this purpose (this is discussed in detail later). It can easily be observed that the Manifold distance is better than the Euclidean distance and the prediction error is significantly lower. Therefore, the rest of this article will use the manifold distance measure.

2) CHOOSING LAG DURATION AND NUMBER OF CANDIDATES

It is very important to use the best lag time and the number of candidates to minimize the prediction error. The lag time affects the performance of K-NN based traffic prediction because it is a major variable for searching the similar traffic state. The study considered a series of lag durations ranging from 1 hour to 24 hours. To a certain extent, shorter lag time is suitable for short-term traffic forecasting, while relatively long time is suitable for long-term traffic forecasting. Another variable that affects the accuracy of forecasting methods is the number of candidates. In this paper, a large number of different numbers of candidates are considered, from one candidate to 24 candidates. The impact of time lag and number of candidates on forecast accuracy is shown in Figure 9. When the number of candidates (K) is greater than 3, it can be seen that as the lag duration increases, the prediction error increases and then start to decrease slightly. This shows that the best lag period to identify similar traffic patterns should be relatively short; Likewise, the number of candidates also affects the accuracy of the forecast, but to a lesser extent than the impact of the lag duration. Figure 9 shows the effect of some candidates on the prediction accuracy when the lag duration ranges from 2 hours to 4 hours. In term of MAPE, Given the increase in the number of candidates, the forecast error decreases. In term of RMSE, as the number of candidates increases, the RMSE begins to fluctuate and then the amplitude begins to decrease until the number of candidates is greater than 15. As shown in Figure 9, in our case, the optimum number of candidates to be considered is found to be twenty-one. And two-hours lag duration is found to be most suitable.

C. PREDICTION BY LEVEL OF TRAFFIC STATE AND ERROR DISTRIBUTION

Take stations 1085, 1086 and 1087 as example, Figure 10 visually reflects the manifold distance between the sample to be predicted and the different traffic conditions. Figure 10 (a) and (b) show the traffic flow data collected by the detection station 1085. Figure 10 (c) and 10 (d), 10 (e) and 10 (f) are respectively stations 1086 and 1087. Detectors 100857 and 100858 in early rush hour are in a better condition with no obvious congestion. Detectors 100860, 100861, 100863 and 100868 are more heavily congestion, with the congestion time longer than 1 hour. Among them, the congestion of detectors 100860 and 100861 dissipates quicker than detectors 100863 and 100864. During the evening peak period, six detectors are in a congested condition and the congestion continues for more than 2 hours. Detectors 100860 and 100861 are in the level of traffic state 4 and 5 alternately.

To perform comparison and analysis of Euclidean distance and manifold distance, the distribution of prediction error on the level of traffic states are evaluated by four indicators: Dev_0, Dev_1, Dev_2 and Dev_H, the formulas as (6)–(9). Dev_1 indicates that the predicted traffic state differs from the actual traffic state by one level, Dev_H indicates that the predicted traffic state differs from the actual traffic state by more than two levels, and so on. Error distribution of road traffic state level prediction (Y_i is actual value and $\hat{Y}_i$ is predicted value) is as shown in Table 3 and 4.

$$\text{Dev\_0} = \frac{n_{right}}{n_{total}} \times 100\% \tag{6}$$

$$\text{Dev\_1} = \frac{n_{one-deviation}}{n_{total}} \times 100\% \tag{7}$$

TABLE 3. Error distribution based on euclidean distance.

Detection ID	5 min				15 min			
	Dev_0	Dev_1	Dev_2	Dev_H	Dev_0	Dev_1	Dev_2	Dev_H
100857	89.51%	6.29%	1.75%	2.45%	83.16%	11.58%	4.21%	1.05%
100858	91.29%	5.23%	1.74%	1.74%	87.37%	8.42%	2.11%	2.11%
100860	93.01%	2.09%	1.4%	3.5%	88.42%	5.26%	1.05%	5.26%
100861	95.12%	1.05%	1.39%	2.44%	86.32%	7.37%	6.32%	0
100863	82.17%	14.69%	1.40%	1.75%	77.89%	15.79%	6.32%	0
100864	94.77%	1.74%	2.44%	1.05%	81.05%	16.84%	1.05%	1.05%

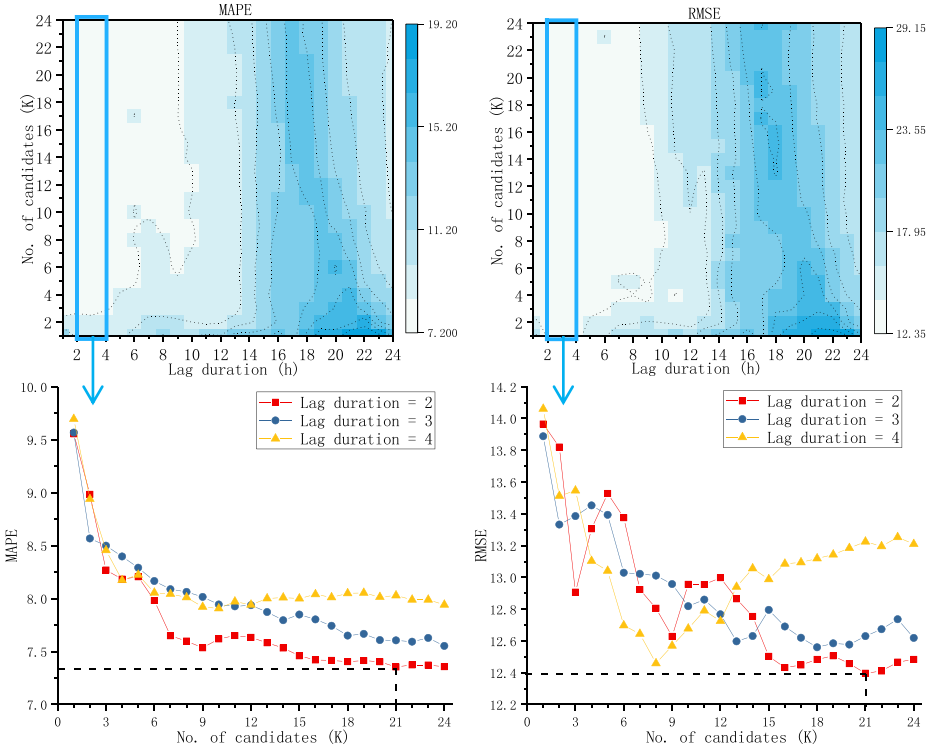

FIGURE 9. Optimum number of candidates given the lag duration in 2-4 h.

TABLE 4. Error distribution based on manifold distance.

Detection ID	5 min				15 min			
	Dev_0	Dev_1	Dev_2	Dev_H	Dev_0	Dev_1	Dev_2	Dev_H
100857	92.66%	3.5%	1.75%	2.09%	86.32%	9.47%	2.11%	2.11%
100858	95.82%	0.7%	1.39%	2.09%	89.47%	6.32%	2.11%	2.11%
100860	93.71%	1.4%	1.05%	3.85%	88.42%	6.32%	0	5.27%
100861	92.33%	3.14%	1.39%	2.44%	88.42%	5.26%	2.11%	4.21%
100863	83.57%	11.89%	2.1%	1.74%	85.26%	8.42%	6.32%	0
100864	95.12%	0.35%	4.53%	0	82.11%	15.79%	1.05%	1.05%

$$\text{Dev\_2} = \frac{n_{two-deviation}}{n_{total}} \times 100\% \quad (8)$$

$$\text{Dev\_H} = \frac{n_{total} n_{right} n_{one-deviation} n_{two-deviation}}{n_{total}} \times 100\% \quad (9)$$

We conducted short-term traffic state forecasting experiments over two kinds of span: 5 minutes and 15 minutes. By 5 minutes interval prediction error, the error Dev_0 distribution at 6 detectors for prediction method of manifold

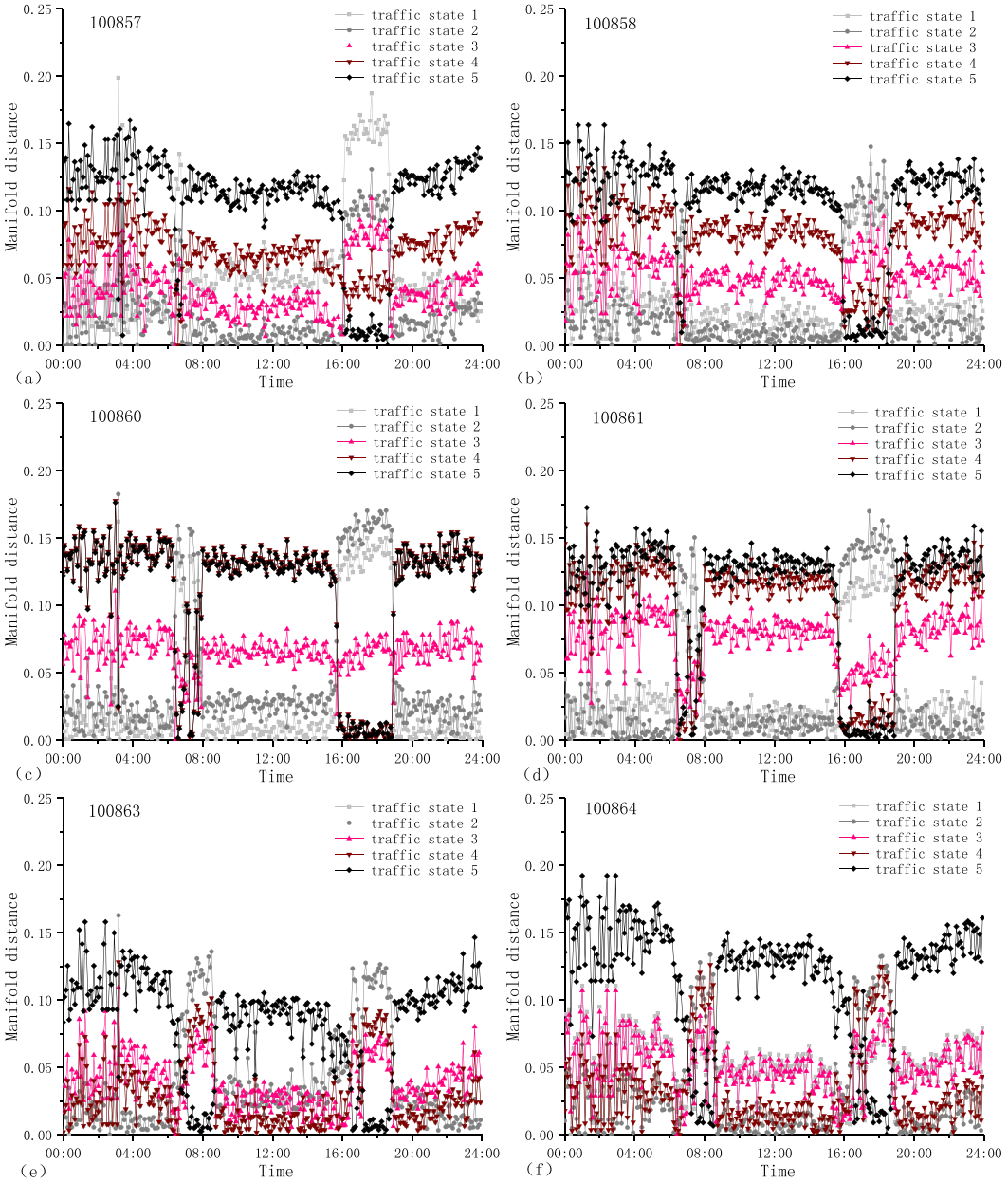

FIGURE 10. Multi-sections manifold distance time series diagram.

distance has five times better than that for prediction method of Euclidean distance. Only for 100863, the prediction accuracy is lower than prediction method based on Euclidean distance. However, the two prediction methods of the Dev_2 rate and Dev_H are almost equal, which means that the prediction precision can be improved by Dev_1 reduction. Look at 15 minute intervals, in error comparison of 6 detectors, prediction method based on manifold distance has five wins and a draw, so it has about 3% higher in average prediction accuracy for 5 wins and the overall accuracy fluctuates within 82%∼89%, reflecting the effectiveness of the traffic state prediction method based on manifold similarity.

D. COMPARISON OF THE RESULTS WITH CLASSIC MODELS

In order to evaluate the performance of TSFMS, detector 100860 is taken as an example. The results are compared with classic models. The models they employed include:

1) MULTILAYER FEEDFORWARD NEURAL NETWORK (MLFNN)

The MLFNN model for traffic state prediction consists of an input layer, a hidden layer, and an output layer. The number of neurons in the hidden layer was set to 10; one output neuron is used. The learning rate was set at 0.3, the momentum rate

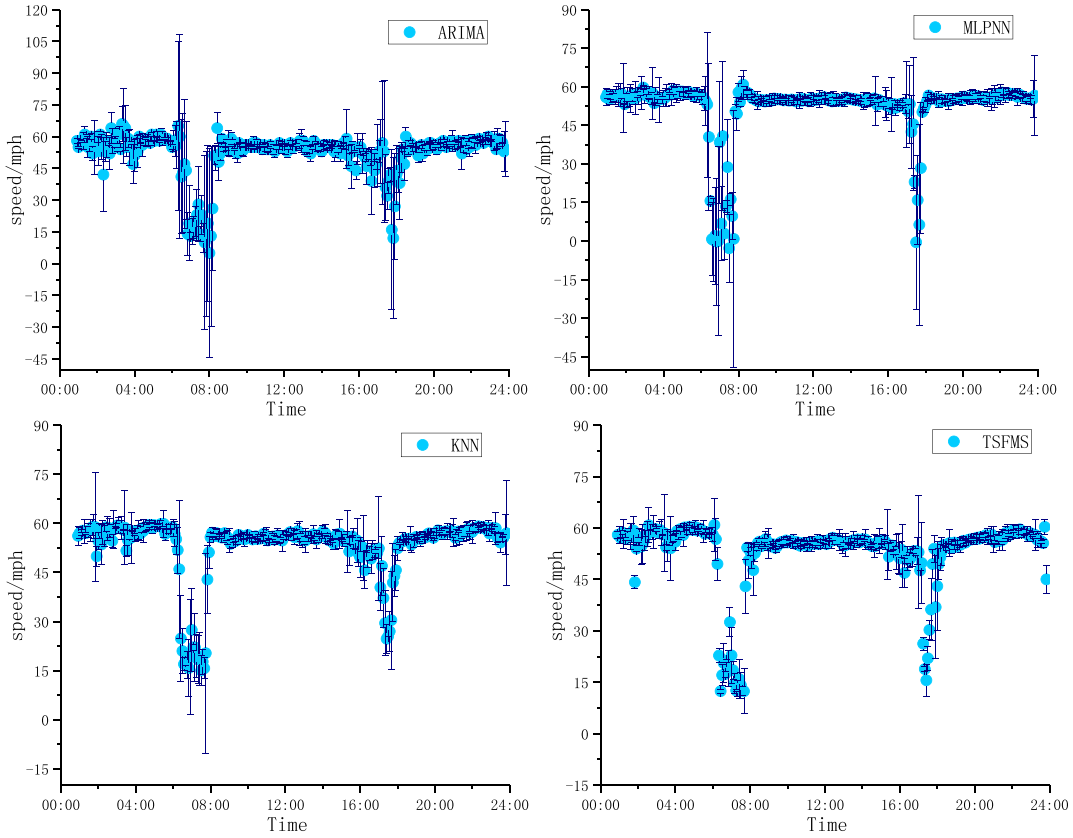

FIGURE 11. Prediction error bars obtained by KNN, ARIMA, MLPNN and TSFMS.

was fixed at 0.2, and the number of training times was set at 1000.

2) ARIMA (p, d, q)

p: the number of lag observations included in the model, is set to 1. d : the number of times that the raw observations are differenced, also called the degree of differencing, is set to 0, q: the size of the moving average window, also called the order of moving average, is set to 1. When two out of the three terms are zeros, the model may be referred to base on the non-zero parameter, dropping "AR", "I" or "MA" from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA (0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

For comparison, MLFNN and ARIMA are performed on the same training set and testing set. MLFNN is implemented in WEKA, which is a popular software tool for machine learning. ARIMA is implemented in Python, which has ARIMA model package.

To intuitively illustrate the prediction results of different methods, we show the predicted traffic flow and the associated error bars obtained by each method in Figure 11. Error bars are graphical representations of the difference of actual data and used on graphs to indicate the error in traffic prediction. Error bars can be expressed in a plus-minus sign ($\pm$), plus the upper limit of the error and minus the lower limit of the error. As shown in Figure 11, The prediction range of ARIMA fluctuates in the range of $(-45, 120)$, the prediction range of MLPNN fluctuates in the range of $(-45, 90)$, the prediction range of KNN fluctuates in the range of $(-15, 90)$ and the range of fluctuation of the prediction of TSFMS at $(0, 90)$. The predicted range of ARIMA, MLPNN and KNN fluctuate more than TSFMS. In the term of error bars, the predicted error bars of ARIMA, MLPNN and KNN are smaller during the small flow period, but the prediction error increase when the morning and evening peak. The prediction error bars of TSFMS during rush hours are less than ARIMA, MLPNN and KNN. From these results, we can observe that the proposed TSFMS achieves smaller prediction error than the other competitors, indicating this method can better capture the traffic flow pattern.

Table 5 lists six kinds of classic prediction algorithm, in which autoregressive models (AR), moving average model (MA), autoregressive moving average model (ARMA), and ARIMA are commonly used by scholars in recent years as prediction methods of traffic state while the MLFNN is commonly used by scholars as prediction method of neural network. Comparing the six methods, TSFMS had advantages in MAPE and EC. According to the above experiments, the prediction results of TSFMS can well reflect the

TABLE 5. Comparison of six traffic state prediction methods.

Evaluation Index	Traffic State Prediction Methods					
	AR(1)	MA(1)	ARMA(1,1)	ARIMA(1,0,1)	MLFNN	TSFMS
MAPE (%)	29.726	22.069	18.048	17.246	19.615	**5.285**
EC (%)	96.396	96.681	96.852	97.265	96.731	**98.936**

trend and regularity of the traffic flow with high prediction accuracy, which can be used in traffic state prediction and has strong competitiveness.

V. CONCLUSION

The ability to predict timely and accurate the evolution of traffic is important to proactive traffic management and to provide travelers with reliable travel times. In this paper, we propose a short-term traffic forecasting method, which is an algorithm TSFMS using manifold distance to predict traffic state. In addition, the TSFMS algorithm variables are optimized, applied to different data sets collected from different sites and compared with other models to estimate the performance of multiple steps to prove the robustness of the method. This study provides the following findings:

- Manifold features of traffic flow that exist within the archived datasets can be used to provide reliable and accurate short-term traffic state forecasts.
- Manifold distance-based KNN algorithm is very effective in identifying the trend and regularity of the traffic flow from large sets of archived data.

One of the limitations of TSFMS is based on the speed data collected only on freeway corridors and does not take into account the ramp. Another limitation is that the impact of factors affecting traffic operations, such as weather conditions, the proportion of heavy vehicles, and the occurrence of incidents, have not been addressed. Future work will focus on traffic state forecasts at the network level as well as forecasting the length of the corridor on the expressway and include external factors that influence traffic operations into the forecasting model.

ACKNOWLEDGEMENT

The authors would like to make a grateful acknowledgment to PORTAL (the official transportation data archive for the Portland-Vancouver Metropolitan region), for providing traffic flow data.

REFERENCES

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[2] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 3–19, Jun. 2014.

[3] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[4] F. Crawford, D. P. Watling, and R. D. Connors, "A statistical method for estimating predictable differences between daily traffic flow profiles," *Transp. Res. B, Methodol.*, vol. 95, pp. 196–213, Jan. 2017.

[5] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generat. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.

[6] Y. Wang, M. Papageorgiou, and A. Messmer, "RENAISSANCE—A unified macroscopic model-based approach to real-time freeway network traffic surveillance," *Transp. Res. C, Emerg. Technol.*, vol. 14, no. 3, pp. 190–212, 2006.

[7] H. van Lint and C. van Hinsbergen, "Short-term traffic and travel time prediction models," *Artif. Intell. Appl. Critical Transp.*, vol. 22, pp. 22–41, Nov. 2012.

[8] W. Y. Szeto, B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate traffic forecasting technique using cell transmission model and SARIMA model," *J. Transp. Eng.*, vol. 135, no. 9, pp. 658–667, Sep. 2009.

[9] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 4, pp. 303–321, 2002.

[10] L. Lin, Y. Li, and A. Sadek, "A k nearest neighbor based local linear wavelet neural network model for on-line short-term traffic volume prediction," *Proc. Social Behav. Sci.*, vol. 96, pp. 2066–2077, Nov. 2013.

[11] M. Lippi, M. Bertini, and P. Frasconi, "Short-Term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 871–882, Jun. 2013.

[12] N.-E. El Faouzi, "Nonparametric traffic flow prediction using kernel estimator," in *Proc. Int. Symp. Trans. Traffic Theory*, 1996, pp. 41–54.

[13] P. Ross, "Exponential filtering of traffic data," *Transp. Res. Rec.*, vol. 869, pp. 43–49, Mar.1982.

[14] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 644–654, Jun. 2012.

[15] M. Cetin and G. Comert, "Short-term traffic flow prediction with regime switching models," *Transp. Res. Rec.*, vol. 1965, no. 3, pp. 23–31, 2006.

[16] M. Cools, E. Moons, and G. Wets, "Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: Assessing the effect of holidays on two site locations," *Transp. Res. Rec.*, vol. 2136, no. 7, pp. 57–66, 2009.

[17] J. Guo, B. Williams, and B. Smith, "Data collection time intervals for stochastic short-term traffic flow forecasting," *Transp. Res. Rec.*, vol. 2024, no. 3, pp. 18–26, 2008.

[18] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 50–64, Jun. 2014.

[19] H. Chen and H. Rakha, "Agent-based modeling approach to predict experienced travel times," in *Proc. Transp. Res. Board 93rd Annu. Meet.*, Washington, DC, USA, 2014, pp. 1–20.

[20] S. Innamaa, "Short-term prediction of travel time using neural networks on an interurban highway," *Transportation*, vol. 32, no. 6, pp. 649–669, 2005.

[21] C.-S. Li and M.-C. Chen, "Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks," *Neural Comput. Appl.*, vol. 23, no. 6, pp. 1611–1629, 2013.

[22] E. I. Vlahogianni, "Prediction of non-recurrent short-term traffic patterns using genetically optimized probabilistic neural networks," *Oper. Res.*, vol. 7, no. 2, pp. 171–184, 2007.

[23] E. Vlahogianni, "Short-term predictability of traffic flow regimes in signalised arterials," *Road Transp. Res., J. Austral. New Zealand Res. Pract.*, vol. 17, no. 2, p. 19, 2008.

[24] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on Chaos–wavelet analysis-support vector machine theory," *Transp. Res. C, Emerg. Technol.*, vol. 27, pp. 219–232, Feb. 2013.
[25] S. A. Zargari, S. Z. Siabil, A. H. Alavi, and A. H. Gandomi, "A computational intelligence-based approach for short-term traffic flow prediction," *Expert Syst.*, vol. 29, no. 2, pp. 124–142, 2012.
[26] W. Zheng, D.-H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach," *J. Transp. Eng.*, vol. 132, no. 2, pp. 114–121, 2006.
[27] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An improved *k*-nearest neighbor model for short-term traffic flow prediction," *Proc.-Social Behav. Sci.*, vol. 96, pp. 653–662, Nov. 2013.
[28] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6164–6173, 2009.
[29] J. Xia, W. Huang, and J. Guo, "A clustering approach to online freeway traffic state identification using ITS data," *KSCE J. Civil Eng.*, vol. 16, no. 3, pp. 426–432, 2012.
[30] W. Wang, L. Pan, and B. Liu, "Synergetic method of traffic state recognition based on manifold learning," in *Proc. IEEE Int. Conf. Autom. Logistics*, Shenyang, China, Aug. 2009, pp. 587–591.
[31] K. Lu, Z. Ding, and S. Ge, "Sparse-representation-based graph embedding for traffic sign recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1515–1524, Dec. 2012.
[32] G. Lee, R. Mallipeddi, and M. Lee, "Identification of moving vehicle trajectory using manifold learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2012, pp. 188–195.
[33] S. Yang and W. Zhou, "Anomaly detection on collective moving patterns: Manifold learning based analysis of traffic streams," in *Proc. IEEE 3rd Int. Conf. Soc. Comput.*, Boston, MA, USA, Jun. 2011, pp. 704–707.
[34] Z. Yang, Q. Bing, X. Zhou, M. Ma, and X. Li, "A short-term traffic flow prediction method based on simiarity search of time series," *J. Transp. Inf. Safety*, vol. 32, no. 6, pp. 22–26, 2014.
[35] Z. Li, Z. Huang, and Y. Zhang, "Fractal property analysis for freeway traffic flow," *China J. High Transp.*, vol. 13, no. 3, pp. 82–85, 2000.
[36] L. Zhang, Y. Jia, Z. Niu, and C. Liao, "Traffic state classification based on parameter weighting and clustering method," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 14, no. 6, pp. 147–151, 2014.
[37] Q.-C. Liu, J. Lu, and S.-Y. Chen, "Traffic state prediction based on competence region," *Acta Phys. Sin.*, vol. 63, no. 14, p. 140504, 2014.
[38] Q.-L. Ma, W.-N. Liu, and D.-H. Sun, "Multi-parameter fusion applied to road traffic condition forecasting," *Acta Phys. Sin.*, vol. 61, no. 16, p. 169501, 2012.
[39] H. Fu, L.-H. Xu, G. Hu, and Y. Wang, "Traffic flow state-forecasting algorithm based on Sugeno neural fuzzy system," *Control Theory Appl.*, vol. 27, no. 12, pp. 1637–1640, 2010.
[40] H. Dong, L. Jia, X. Sun, C. Li, Y. Qin, and M. Guo, "Traffic state forecasting to urban freeway based on the maximum entropy model," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 10, no. 2, pp. 112–116, 2010.
[41] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.
[42] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2016.
[43] S. Clark, "Traffic prediction using multivariate nonparametric regression," *J. Transp. Eng.*, vol. 129, no. 2, pp. 161–168, 2003.
[44] W. Qiao, A. Haghani, and M. Hamedi, "A nonparametric model for short-term travel time prediction using bluetooth data," *J. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 165–175, 2013.
[45] W.-H. Lin, Q. Lu, and J. Dahlgren, "Dynamic procedure for short-term prediction of traffic conditions," *Transp. Res. Rec.*, vol. 1783, no. 19, pp. 149–157, 2002.

QINGCHAO LIU received the Ph.D. degree from Southeast University, Nanjing, China, in 2015. He joined the Automotive Engineering Research Institute, Jiangsu University, as a Lecturer. His research interests include driving behavior analysis, path planning, traffic safety, intelligent automobiles, and intelligent transportation systems.

YINGFENG CAI received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China. In 2013, she joined the Automotive Engineering Research Institute, Jiangsu University, as an Assistant Professor. Her research interests include computer vision, intelligent transportation systems, and intelligent automobiles.

HAOBIN JIANG received the B.Sc. degree from Nanjing Agricultural University, Nanjing, China, in 1991, and the M.Sc., and Ph.D. degrees from Jiangsu University, Zhenjiang, China, in 1994 and 2000, respectively, all in mechanical engineering. He is currently a Professor with the Automotive Engineering Research Institute, Jiangsu University. His research interests include vehicle dynamics performance analysis and electronic control technologies for vehicles.

XIAOBO CHEN received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology in 2013. From 2015 to 2017, he served as a Post-Doctoral Research Associate with The University of North Carolina at Chapel Hill, USA. He is currently an Associate Professor with the Automotive Engineering Research Institute, Jiangsu University, China. His research interests include pattern recognition and its applications.

JIAN LU received the B.Sc., M.Sc., and Ph.D. degrees from Southeast University, Nanjing, China, in 1995, 1998, and 2003, respectively, all in traffic engineering. He is currently a Professor with the School of Transportation, Southeast University. His research interests include traffic management, transport planning, traffic safety, and intelligent transportation systems.