Research Article

Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction

Hua-pu Lu, Zhi-yuan Sun, and Wen-cong Qu

Institute of Transportation Engineering, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Zhi-yuan Sun; sun-zy12@mails.tsinghua.edu.cn

Received 17 July 2014; Accepted 11 September 2014

Academic Editor: Wei Guan

Copyright © 2015 Hua-pu Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of urban informatization, the era of big data is coming. To satisfy the demand of traffic congestion early warning, this paper studies the method of real-time traffic flow state identification and prediction based on big data-driven theory. Traffic big data holds several characteristics, such as temporal correlation, spatial correlation, historical correlation, and multistate. Traffic flow state quantification, the basis of traffic flow state identification, is achieved by a SAGA-FCM (simulated annealing genetic algorithm based fuzzy *c*-means) based traffic clustering model. Considering simple calculation and predictive accuracy, a bilevel optimization model for regional traffic flow correlation analysis is established to predict traffic flow parameters based on temporal-spatial-historical correlation. A two-stage model for correction coefficients optimization is put forward to simplify the bilevel optimization model. The first stage model is built to calculate the number of temporal-spatial-historical correlation variables. The second stage model is present to calculate basic model formulation of regional traffic flow correlation. A case study based on a real-world road network in Beijing, China, is implemented to test the efficiency and applicability of the proposed modeling and computing methods.

1. Introduction

Real-time traffic flow state identification and prediction is one of the critical components of intelligent transportation system. It is of practical significance to identify and predict traffic flow state quickly, precisely and timely. As the first problem that needs to be solved, traffic flow state can be measured by level of service (LOS), which is first introduced in the 1965 Highway Capacity Manual (HCM 1965) [1]. The latest version of Highway Capacity Manual (HCM 2010) [2] divided LOS into six levels. Sasaki and Iida [3] divided LOS into three levels. China GB 50220-1995 [4] divided LOS into four levels. Existing studies of LOS evaluation can be classified into four categories in transportation literature: subjective evaluation based models [5], statistical analysis based models [6], artificial intelligence based models [7], traffic flow theory based models [8, 9].

Many previous research works concerning traffic flow state identification especially for Beijing have been carried out. Guan and He [10] analyzed the statistical features of speed distribution at different density and divided traffic

flow state into four levels based on flow-rate-density plane. Liao et al. [11] studied traffic state identification based on perceptual experiment. Xia et al. [12] built traffic state rapid identification model based on fuzzy theory. Qu et al. [13] determined the relation between traffic state and travel speed. Moreover, Beijing Traffic Management Bureau built Beijing Regional Traffic Conditions and LOS Evaluation System mainly based on fixed detectors [14], and Beijing Municipal Commission of Transport built Traffic Performance Index System mainly based on floating car data [15].

Real-time traffic flow prediction aims at evaluating anticipated traffic flow state at a future time. Many studies focused on traffic flow state and parameters prediction. Liu et al. [16] analyzed multidimensional parameters and developed the traffic prediction models of different dimensions based on the support vector machine. Dong et al. [17] proposed a preselection space time model to estimate the traffic condition at poor-data-detector, especially nondetector locations. Canaud et al. [18] presented a probability hypothesis density filtering based model for real-time traffic flow state prediction. Pan et al. [19] put forward a modified stochastic

cell transmission model to support short-term traffic flow state prediction. Antoniou et al. [20] proposed an approach for local traffic flow state estimation and prediction based on data-driven computational approaches. Zhang et al. [21] designed a traffic flow state estimator based on extended-Kalman-filter method.

Traffic big data, massive and multisource, brings both opportunities and challenges to effective traffic management and control. During data processing, traffic big data meets the same difficulties with the general big data, such as capture, storage, search, sharing, analytics, and visualization. The differences between traffic big data and traditional data in the field of traffic flow state identification and prediction are obvious. For traffic flow state identification, the advantage of traffic big data is mainly manifested in the full coverage. That is, traffic big data can represent traffic flow characteristics as much as possible. However, the difficulties of data processing are enhanced with the increasing of the size of data. Therefore, big data-driven methods should have a stronger capacity of data processing. For traffic flow state prediction, the advantage of traffic big data is mainly manifested in the multisource. That is, the traffic big data has a higher accuracy to describe the relationships between traffic flow state of section s at time t and the others. However, the relationships are not easy to be found. So, big data-driven methods should have a clearer description of physical significance of traffic flow. Traffic big data can improve the efficiency of real-time traffic flow state identification and prediction. However, it may face more challenges.

Taking into account all the present researches in this field, there is still a lack of consideration of traffic big data. Further researches remain to be conducted on the direction of traffic flow state analysis. In this paper, the method of real-time traffic flow state identification and prediction, which perceives ability to handle big data, is put forward in detail. The remainder of this paper is organized as follows. Section 2 presents the basic characteristics of traffic big data. In Section 3, the methodology of real-time traffic flow state identification and prediction is proposed. A case study based on a real-world road network is carried out in Section 4 to demonstrate the performance and applicability of the proposed method. Finally, conclusions are drowned in Section 5.

2. Traffic Big Data Analysis

2.1. Temporal Correlation. By laying fixed and mobile traffic flow detectors, dynamic traffic data of a section can be obtained. Dynamic traffic flow data V are time-series data, which continuously change over time t with a certain trend; namely,

$$V_t = T(V_{t-1}, V_{t-2}, \dots), \quad (1)$$

where V_t is the traffic flow parameter at time t , representing flow q_t , speed v_t , or occupancy ρ_t . T is a function that describes trends of traffic flow time-series data.

2.2. Spatial Correlation. A regional transportation network consists of multiple intersections and roads. It exists a spatial association between traffic flow data of neighboring junctions or sections and that of target junctions or sections; namely,

$$V_{s,t} = S(V_{1,t}, V_{2,t}, \dots), \quad (2)$$

where $V_{s,t}$ is the traffic flow parameter of section s at time t , representing flow $q_{s,t}$, speed $v_{s,t}$, or occupancy $\rho_{s,t}$. S defines the relationship between the data of upstream and downstream sections.

2.3. Historical Correlation. For different weekdays or weekends, trip distribution of inhabitants in the same period shows similar characteristics. The law of traffic flow cycle is especially evident. Correlation between dynamic traffic flow data and historical data is as follows:

$$V_{D,t} = H(V_{D-1,t}, V_{D-2,t}, \dots), \quad (3)$$

where $V_{D,t}$ is the traffic flow parameter on day D at time t , representing flow $q_{D,t}$, speed $v_{D,t}$, or occupancy $\rho_{D,t}$. H defines the relationship between the data of different workdays or holidays at the same time.

2.4. Multistate. Traffic flow state can be determined by the traffic flow parameters. The correspondence between traffic flow states and traffic flow parameters is as follows:

$$A = MS(v, q, \rho), \quad (4)$$

where A is LOS. v , q , and ρ are the traffic flow parameters. MS defines the relationship between traffic flow states and traffic flow parameters.

3. Methodology

3.1. Overall Framework. The overall traffic flow state identification and prediction framework is outlined in Figure 1.

(a) *SAGA-FCM Based Traffic Clustering Model.* As a random social phenomenon, traffic flow state holds fuzzy characteristics. For example, “congested” and “uncongested” are the fuzzy descriptions of traffic flow. It is difficult to describe these states quantitatively. Since membership function can be used to explain the fuzzy phenomenon, fuzzy c -means algorithm (FCM) is designed to evaluate traffic flow state [22]. However, FCM is sensitive to its initial clustering center and it is easy to fall into the local optimum, especially for big data [23]. SAGA-FCM algorithm integrates the strong local search ability of simulated annealing algorithm (SA) and the strong global search ability of genetic algorithm (GA) to overcome the drawbacks of traditional FCM algorithm [24].

(b) *DFT Based Historical Data Fusion Model.* Road traffic flow state information is periodic [25]. Each object to predict can be associated with a set of historical data. Actual intelligent traffic system has accumulated massive historical data. However, it is difficult to achieve the integration of historical data. During the on-going development of traffic prediction

FIGURE 1: Overall traffic flow state identification and prediction framework.

system in Beijing, discrete Fourier transform (DFT) is found effective in achieving the average of historical data and noise elimination [26]. Traffic parameters are discrete time series sequence; therefore DTF is applied to integrate massive historical data [27].

(c) *Regional Traffic Flow Correlation Model.* Temporal correlation, spatial correlation and historical correlation are the characteristics of traffic big data. Based on this principle, regional traffic flow correlation model can be established to predict traffic flow parameters. Taking simple calculation and predictive accuracy into consideration, a bi-level optimization model is proposed to simplify the regional traffic flow correlation model.

(d) *Correction Coefficients Optimization Model.* Considering the amount of temporal, spatial, and historical variables, each of which corresponds to an unknown parameter, the model is too difficult to calibrate. Therefore, a two-stage model for correction coefficients optimization is put forward to reduce the number of variables and ensure accuracy simultaneously.

3.2. SAGA-FCM Based Traffic Clustering Model

3.2.1. *Basic Traffic Clustering Model.* Based on HCM 2010 [2], LOS (A) can be divided into six types, as $A = [A_1, A_2, \dots, A_6]$. Considering the monotonicity between traffic flow parameters and LOS, speed (v) and occupancy (ρ) are chosen as clustering parameters. Therefore, data sample used for traffic clustering model can be written as

$X = \{x_1, x_2, \dots, x_n\}$, $x_j = (v_j, \rho_j)$. Since traditional FCM algorithm is put forward to find an optimal classification, the objective function is expressed as

$$\begin{aligned} \min \quad & J_b(U, a) = \sum_{i=1}^6 \sum_{j=1}^n (\mu_{ij})^b (d_{ij})^2 \\ \text{s.t.} \quad & \sum_{i=1}^6 \mu_{ij} = 1, \quad 0 \leq \mu_{ij} \leq 1 \\ & i = 1, 2, \dots, 6, \quad j = 1, 2, \dots, n, \end{aligned} \quad (5)$$

where U is membership matrix, $U = [\mu_i(x_j)]_{6 \times n}$. a is the set of clustering centers, $a = \{a_1, a_2, \dots, a_6\}$, $a_i = (v_i, \rho_i)$. d_{ij} is the Euclidean distance that measures distance between data sample (x_j) and clustering center (a_i), $d_{ij} = d(x_j - a_i) = \|x_j - a_i\|$. b is weighted parameter, $b \in [1, +\infty)$. $\mu_i(x_j)$ is the degree of membership of which the sample data x_j belongs to LOS A_i , shorten for μ_{ij} .

a_i can be calculated by the flowing formulation

$$\begin{aligned} a_i &= \frac{\sum_{j=1}^n (\mu_{ij})^b x_j}{\sum_{j=1}^n (\mu_{ij})^b}, \\ \text{s.t.} \quad & \sum_{i=1}^6 \mu_{ij} = 1, \quad 0 \leq \mu_{ij} \leq 1, \quad 0 \leq \sum_{j=1}^n \mu_{ij} \leq 1 \\ & i = 1, 2, \dots, 6, \quad j = 1, 2, \dots, n. \end{aligned} \quad (6)$$

μ_{ij} can be calculated by the flowing formulation

$$\mu_{ij} = \begin{cases} \frac{1}{\sum_{k=1}^6 (d_{ij}/d_{kj})^{2/(b-1)}}, & d_{ij} > 0, \\ 1, & d_{ij} = 0, i = j, \\ 0, & d_{ij} = 0, i \neq j. \end{cases} \quad (7)$$

Modification of the clustering centers and degrees of membership is the core of FCM. When the algorithm is convergent, clustering centers a_i can be found, and degrees of membership μ_{ij} can be calculated. Therefore, fuzzy clustering result can be achieved.

3.2.2. SAGA-FCM Based Solution Algorithm. SAGA-FCM algorithm is used to improve the clustering quality of FCM. The process of SAGA-FCM algorithm can be summarized as in the following steps [28].

Step 1 (parameters initialization). Population size (*sizepop*), maximum genetic evolution algebra (*MAXGEN*), crossing probability (P_c), gene variation probability (P_m), original temperature (T_0), cooling coefficient (l), and final temperature (T_{end}).

Step 2 (clustering centers initialization). Randomly selecting six clustering centers (a), designing the original sequence (*chrom*), calculating degrees of membership (μ_{ij}), and fitness value of each individual (f_i).

Step 3 (genetic evolution algebra initialization). $gen = 0$.

Step 4 (genetic operations). Selection, crossover, and mutation of original sequence (*chrom*), achieving new clustering centers (a), calculating degrees of membership (μ'_{ij}), and fitness value of each individual (f'_i).

Step 5 (fitness value judgment). If $f'_i > f_i$, replace all of the old individual with the new individual; otherwise, replace the old individual with the new individual by probability ($P = \exp(f_i - f'_i)T$).

Step 6 (*MAXGEN* judgment). If $gen < \text{MAXGEN}$, $gen = gen + 1$, turn to Step 4; otherwise, turn to Step 7.

Step 7 (temperature judgment). If $T_i < T_{\text{end}}$, the new solution is outputted, and that is the most optimized assembly sequence; otherwise, $T_{i+1} = lT_i$, turning to Step 3.

3.2.3. Calculation of LOS. Based on SAGA-FCM algorithm, six clustering centers (a) can be found. Thus, μ_{ij} is calculated by formulation (7). Combining the numerical magnitude of μ_{ij} , the LOS of $x_j(A(x_j))$ is achieved:

$$\begin{aligned} A(x_j) &= A_k \\ \text{s.t. } \mu_{kj} &> \mu_{ij}, \quad k \neq i, \end{aligned} \quad (8)$$

where $A(x_j)$ is the LOS of x_j .

3.3. DFT Based Historical Data Fusion Model. Given a discrete temporal series $\{x(0), x(1), \dots, x(N-1)\}$, the length of which is N , its discrete Fourier transform with M sampling points can be calculated:

$$X(k) = \frac{1}{N} \times \sum_{n=0}^{N-1} x(n) \times e^{-j(2k\pi/N)n}, \quad (k = 0, 1, \dots, M-1). \quad (9)$$

The corresponding inverse Fourier transform formulation is

$$x(n) = \sum_{k=0}^{M-1} X(k) \times e^{j(2k\pi/N)n}, \quad (k = 0, 1, \dots, M-1). \quad (10)$$

Normally, take $M = N$. To facilitate the program, the natural logarithm complex power form above is converted to trigonometric form:

$$\begin{aligned} x(i) &= a_0 + \sum_{f=1}^{N-1} \left(\alpha_n \cos\left(\frac{2\pi}{N} f_i\right) + \beta_n \sin\left(\frac{2\pi}{N} f_i\right) \right) \\ &\quad \times (f_i = e^{j(2k\pi/N)i} \quad (i = 0, 1, \dots, N-1)). \end{aligned} \quad (11)$$

Suppose that the traffic flow varies in one week cycle, which includes 7 days with different characteristics. Each day contains 288 pieces of data with a time interval of 5 minutes. Apply DFT to average massive historical data and to eliminate noise, so as to achieve the integration of historical data. Consider

$$V_D^H = \{V_{D,1}^H, \dots, V_{D,t}^H, \dots, V_{D,288}^H\}, \quad (12)$$

where V_D^H is the collection of historical data on day D , $D = [1, 2, \dots, 7]$. $V_{D,t}^H$ is the traffic data on day D at time t , $V_{D,t}^H = [q_{D,t}^H, v_{D,t}^H, \rho_{D,t}^H]^T$, $t = [1, 2, \dots, 288]$.

For example, the process of DFT of traffic volume (q) can be summarized as in the following steps.

Step 1. Enter the discrete temporal series $\{q(0), q(1), \dots, q(287)\}$.

Step 2. Apply DFT and acquire Fourier coefficients sequence $\{\alpha_1, \beta_1, \alpha_2, \beta_2, \dots, \alpha_{287}, \beta_{287}\}$.

Step 3. Filter Fourier coefficients series to excise all coefficients less than the specified amplitude threshold of coefficient. Record the rest of the coefficients. The value of the threshold needs to be chosen by analyzing the historical data.

Step 4. Restore data at any point in the sequence $q(i)$.

3.4. Regional Traffic Flow Correlation Model

3.4.1. Basic Model Formulation. The basic assumption of the regional traffic flow correlation model is that traffic flow has a strong temporal-spatial-historical correlation, namely:

- (i) in the temporal series, the traffic flow of the last moment can be regarded as the continuation of the current traffic flow;
- (ii) in the spatial series, downstream sections of the traffic flow can be seen as a continuation of the upstream traffic flow;
- (iii) in the historical series, traffic demand characteristics determine that traffic flow characteristics of the same day in the same period are similar.

Basic form of regional traffic flow correlation model is expressed as

$$\begin{aligned} V_s(t) &= \gamma_T V_T(t) + \gamma_S V_S(t) + \gamma_H V_H(t) \\ \text{s.t. } \gamma_T + \gamma_S + \gamma_H &= 1, \quad \gamma_T, \gamma_S, \gamma_H \geq 0, \end{aligned} \quad (13)$$

where $V_s(t)$ is the traffic flow parameter of section s at time t , representing flow $q_s(t)$, speed $v_s(t)$, or occupancy $\rho_s(t)$. $V_T(t)$, $V_S(t)$, and $V_H(t)$ are the estimated value of $V_s(t)$. $V_T(t)$ is calculated by temporal correlation analysis. $V_S(t)$ is calculated by spatial correlation analysis. $V_H(t)$ is calculated by historical correlation analysis. γ_T , γ_S , and γ_H are coefficients of these three variables.

According to formulation (1), T is generally achieved by regression analysis model. Thus, $V_T(t)$ is created by the following equation:

$$\begin{aligned} V_T(t) &= \theta_1 V_s(t-1) + \theta_2 V_s(t-2) + \dots \\ \text{s.t. } \sum \theta &= 1, \quad \theta \geq 0, \end{aligned} \quad (14)$$

where $V_s(t-d)$ is the traffic flow parameter of section s at time $t-d$. d is time delay, $d \geq 0$. $\theta_1, \theta_2, \dots$ are regression coefficients.

According to formulation (2), S is generally achieved by neighbor regression model. Thus, $V_S(t)$ is created by the following equation:

$$\begin{aligned} V_S(t) &= \lambda_1 V_1(t-1) + \lambda_2 V_2(t-2) + \dots \\ \text{s.t. } \sum \lambda &= 1, \quad \lambda \geq 0, \end{aligned} \quad (15)$$

where $V_j(t-d)$ is the traffic flow parameter of section j at time $t-d$. $\lambda_1, \lambda_2, \dots$ are regression coefficients.

According to formulations (12), (13), (14), and (15), simplified equation is achieved:

$$\begin{aligned} V_s(t) &= \sum \sum \omega_{jd} V_j(t-d) \\ \text{s.t. } \sum \omega &= 1, \quad \omega > 0, \quad \eta \in N^*, \end{aligned} \quad (16)$$

where ω_{jd} is regression coefficient of $V_j(t-d)$. η is the number of ω_{jd} .

3.4.2. Effectiveness Proof of Proposed Formulation. The general form of formula (16) is

$$\begin{aligned} Y &= \sum_{i=1}^n w_i Y_i \\ \text{s.t. } \sum_{i=1}^n w_i &= 1, \quad w_i \geq 0, \end{aligned} \quad (17)$$

where Y is the state of physical system with the measurement of multisensors. Y_i is the measured value of i th sensor, and they are independent. w_i is a weighting factor.

The total mean square error δ^2 is a multivariate quadratic function of w_i :

$$\delta^2 = \sum_{i=1}^n w_i^2 \delta_i^2, \quad (18)$$

where δ_i is the mean square error of Y_i .

According to extreme value theorem of multivariate quadratic function, weighting factor can be calculated as follows:

$$w_i = \frac{1}{\delta_i^2 \sum_{i=1}^n (1/\delta_i^2)}. \quad (19)$$

Assume that the detection accuracy varies with sensors, the mean square error of the measurements from the least accurate sensor is $\delta_{\min}^2$ and that from the most accurate sensor is $\delta_{\max}^2$. Then,

$$\begin{aligned} \text{esp}[Y] &= \left[\frac{1}{\delta_{\max}^2} + \frac{1}{\delta_{\min}^2} + \sum_{i=1}^{n-2} \frac{1}{\delta_i^2} \right]^{-1} \\ &< \left[\frac{1}{\delta_{\min}^2} + \sum_{i=1}^{n-2} \frac{1}{\delta_i^2} \right]^{-1} < \delta_{\min}^2. \end{aligned} \quad (20)$$

The formula above shows that the measurement accuracy of the system can be improved by multisensor measurement and that participation of an inaccurate sensor also helps improve the accuracy of the system.

3.4.3. Bilevel Optimization Model. For big date-driven analysis, the speed and accuracy of data processing are both important. The number of unknown parameters in formulation (16) is large and uncontrolled. Therefore, taking simple calculation and predictive accuracy into consideration, a bilevel optimization model is proposed.

The upper level model is formulated to optimize the speed of date processing by minimizing the number of temporal-spatial-historical correlation variables. The upper level formulation is

$$\begin{aligned} \min \quad & \eta \\ \text{s.t. } \quad & \eta \in N^*. \end{aligned} \quad (21)$$

The lower level model is articulated in accordance with predictive accuracy. Therefore, the lower level formulation is proposed in order to minimize predictive encoding. The lower level formulation is

$$\begin{aligned} \min \quad & |V_s(t) - \widehat{V}_s(t)| \\ \text{s.t. } \quad & V_s(t) = \sum \sum \omega_{jd} V_j(t-d), \quad \sum \omega = 1, \quad \omega \geq 0, \end{aligned} \quad (22)$$

where $\widehat{V}_s(t)$ is actual value of $V_s(t)$.

In the bilevel optimization model, the purpose of objective function is to achieve a best combination, which holds both fast calculation speed and high prediction accuracy.

3.5. Correction Coefficients Optimization Model. The more the temporal-spatial-historical correlation variables are, the slower the program calculates and the higher the accuracy is. The bilevel optimization model above is difficult to solve. To put it in another way, we can define a threshold of computing speed and therefore derive the maximum of acceptable number of temporal-spatial-historical correlation variables. In addition, we can put forward a new variable which can be calculated to replace the variable ω_{jd} . Since the alternative process will bring some errors, which are likely to be systematic, a linear correction method is put forward. Therefore, the correction coefficients optimization model is a two-stage model.

3.5.1. The First Stage Model. The first stage model is mainly used to determine the number of unknown parameters η . It can be determined through the following steps.

Step 1 (determination of the range of η). The number of unknown parameters should not exceed a threshold value. After several data tests, it is found that when η is less than 4, the correlation coefficients are not very different from relatively high values; when η is more than 8, the rapid decay of the correlation coefficients is observed with relatively low values. Therefore η ranges from $[4, 8]$.

Step 2 (calculation of the correlation coefficients). Calculate the correlation coefficients (r_{jd}) between the studied section s and the related section j , solve the max correlation coefficient (r_j) corresponding to each section j , and determine the corresponding time delay (d); namely, $r_j = \max r_{jd}$. r_{jd} can be solved by

$$r_{jd} = \left(\sum_{t=1}^{n-d} V_s(t) V_j(t-d) - \frac{1}{n-d} \sum_{t=1}^{n-d} V_s(t) \sum_{t=1}^{n-d} V_j(t-d) \right) \times \left(\left[\sum_{t=1}^{n-d} V_s^2(t) - \frac{1}{n-d} \left(\sum_{t=1}^{n-d} V_s(t) \right)^2 \right] \times \left[\sum_{t=1}^{n-d} V_j^2(t-d) - \frac{1}{n-d} \left(\sum_{t=1}^{n-d} V_j(t-d) \right)^2 \right] \right)^{-1/2}$$

(23)

Step 3 (selection of the relative section). Sort r_j in ascending order, and take the first η section as a range of temporal and spatial correlation.

3.5.2. The Second Stage Model. The second stage model is mainly used to reduce parameters and determines the base model. Through the first stage model, the number of unknown parameters is reduced. But the model is still difficult to solve and can be further simplified.

FIGURE 2: Spatial location of research object.

Step 1 (calculation of the normalization factor). Take normalized r_j as the parameters of $V_j(t-d)$:

$$\omega_j = \frac{r_j}{\sum_{j=1}^{\eta} r_j}, \quad \sum_{j=1}^{\eta} \omega_j = 1, \quad \omega_j > 0. \quad (24)$$

Step 2 (solving the system error). Introduce the correction parameters φ and ε to calibrate, in order to reduce the system error caused by parameter substitution:

$$\begin{aligned} \min \quad & |V_s(t) - \widehat{V}_s(t)| \\ \text{s.t.} \quad & V_s(t) = \varphi \sum_{j=1}^{\eta} \omega_j V_j(t-d) + \varepsilon. \end{aligned} \quad (25)$$

4. Case Study

4.1. Data Characteristics. Taking a section of the Second Ring Road (Section 1, as shown in Figure 2) and its surrounding roads in Beijing, China, as the object of study, it verifies the effectiveness and feasibility of the proposed method. Basic traffic flow data are detected by microwave detectors.

As shown in Figure 2, Section 1, Section 2, Section 3, and Section 9 are all Second Ring Road. Section 4 and Section 5 are Expressway Tie Line. Section 6, Section 7, Section 8, and Section 10 are Secondary Road. Section 2, 3, ..., 10 are the temporal-spatial correlated sections of Section 1.

For traffic flow state identification, it needs enough data to divide LOS. The historical data of Section 1 (one week, five minutes for the interval) are used to build SAGA-FCM based traffic clustering model. The amount of data is large enough to represent nearly all the traffic flow characteristics.

TABLE 1: Clustering centers.

LOS	Speed (km/h)	Occupancy (%)
A	68.1	2.1
B	58.3	3.1
C	51.2	11.4
D	46.0	14.9
E	38.0	20.5
F	23.4	35.1

FIGURE 3: Traffic flow state evaluation result.

For traffic flow prediction, it needs enough data to describe the temporal-spatial-historical correlation characteristic, that is, to build correction coefficients optimization model. The historical data of Section 1 (one week, five minutes for the interval) are used for temporal-spatial correlation analysis. The historical data of Section 1, 2, ..., 10 (one week, five minutes for the interval) are used for temporal-spatial-historical correlation analysis. The historical data of Section 1 (one month, five minutes for the interval) are used for historical correlation analysis.

4.2. Traffic Flow State Identification

4.2.1. Offline Training. The main task of offline training is to establish the evaluation method of traffic flow state. Case study of traffic flow state identification is done based on the two clustering methods (FCM and SAGA-FCM). The value of objective function calculated by SAGA-FCM is better than the one by FCM. Different traffic flow states are respected by different colors. As shown in Figure 3, traffic flow state is divided into six levels, as A to F. Clustering centers are shown in Table 1.

4.2.2. Online Application. The main task of online application is to realize traffic flow state identification. Traffic flow state identification result is shown in Figure 4.

FIGURE 4: Traffic flow state identification.

TABLE 2: Temporal-spatial-historical correlation coefficients.

j	d (min)	r_{jd}	ω_j
9	0	0.9970	0.2677
1 (temporal correlation)	5	0.9797	0.2630
1 (historical correlation)	0	0.9167	0.2461
2	0	0.8312	0.2232

4.3. Traffic Flow State Prediction

4.3.1. Offline Training. The offline training mainly summarizes the prior knowledge based on historical data.

(1) *Parameter Initialization.* Choose 9 of the upstream and downstream sections of Section 1 based on their attributes and transport network characteristics. Let $\eta \in [4, 8]$.

(2) *Search of Temporal-Spatial-Historical Correlative Sections.* Calculate the correlation coefficient between Section 1 and its relative sections, time series data, and historical trend data. The top 6 of the calculation result are shown in Figure 5, where the abscissa is the time delay d and the vertical axis is the value of the correlation coefficient r_{jd} .

Compare r_{jd} , decide the value of corresponding d , and then get r_j . According to the distribution characteristics of r_j , let $\eta = 4$. The search results of temporal-spatial-historical correlative sections are Section 9, Section 1 (time series data), Section 1 (historical trend data), and Section 2.

Results of temporal-spatial-historical correlation coefficients are shown in Table 2.

(3) *Calculation of Normalization Parameter.* Calculate the normalization parameters ω_j , as shown in Table 2.

(4) *Solving Correction Parameters.* Apply regression analysis to solve correction parameters φ and ε :

$$v_1(t) = 1.029 \times \left[0.2677 \cdot v_9(t) + 0.2630 \cdot v_1(t-5) + 0.2461 \cdot v_1^H(t) + 0.2232 \cdot v_2(t) \right] + 1.8173, \quad (26)$$

FIGURE 5: Temporal-spatial-historical correlation analysis based speed.

FIGURE 6: Fitting curve.

where $v_j(t)$ is the velocity of section j at time t and $v_1^H(t)$ is the historical relative data of Section 1.

Apply the same method to analyze flow and occupancy:

$$\begin{aligned}
 q_1(t) = & 1.0629 \times [0.2246 \cdot q_9(t) + 0.2151 \cdot q_1(t-5) \\
 & + 0.2145 \cdot q_1^H(t-5) + 0.1787 \cdot q_2(t) \\
 & + 0.1670 \cdot q_6(t-10)] + 22.652,
 \end{aligned} \quad (27)$$

$$\begin{aligned}
 \rho_1(t) = & 1.0148 \times [0.2791 \cdot \rho_9(t) + 0.2747 \cdot \rho_1(t-5) \\
 & + 0.2294 \cdot \rho_2(t) + 0.2168 \cdot \rho_1^H(t)] \\
 & + 0.4793,
 \end{aligned} \quad (28)$$

where $q_j(t)$ is the flow of Section j at time t , $q_1^H(t)$ is the historical relative data of Section 1, $\rho_j(t)$ is the occupancy of

Section j at time t , and $\rho_1^H(t)$ is the historical relative data of Section 1.

Fitting curve is shown in Figure 6.

4.3.2. Online Application. Online calculations mainly deal with real-time data based on prior knowledge. Figure 7 shows the predicted results. The prediction errors of speed, flow, and occupancy are controlled around 10%. Traffic state deviation is shown in Table 3. Two-level-deviation and Three-level-deviation are considered as errors, which are controlled in 10%.

5. Conclusions

In the era of big data, real-time traffic flow state identification, and prediction may face many challenges. This difficulty is decided by the characteristics of big data. This paper uses speed and occupancy to build the traffic flow state clustering

FIGURE 7: Modification effective analysis.

TABLE 3: Comparative analysis.

Deviation	Non-level-deviation	One-level-deviation	Two-level-deviation	Three-level-deviation
Accuracy	75.37%	15.80%	8.46%	0.37%

model. And real-time traffic flow state identification is based on speed and occupancy. Traditional fuzzy c -mean clustering does not meet the requirements of big data analysis. To improve the feasibility of traffic flow state clustering, this paper uses the simulated annealing genetic algorithm based fuzzy c -means (SAGA-FCM). Case study shows that the value of objective function calculated based on SAGA-FCM is better.

Traffic flow big data strongly shows temporal, spatial, and historical correlations. The regional traffic flow correlation model is established for real-time traffic flow prediction. The characteristics of big data make it difficult to resolve the model. In order to reduce parameters and ensure calculation speed and calculation accuracy, the correction coefficients optimization model, which can be divided into two stages, is put forward. Effectiveness of the method has been validated by the case study.

The core of this paper is to present a traffic temporal-spatial-historical correlation model, which comprehensively considers the temporal, spatial, and historical correlations of traffic flow big data. Compared with the model based on a single nature, the accuracy of proposed model is relatively high. Besides, this model quantitatively solves, from

the perspective of spatiotemporal correlation analysis, the weight selection problem of different analytical methods in the traditional combined model. The traffic temporal-spatial-historical correlation model can be applied in other researches such as identification and correction of abnormal data and traffic congestion mechanism analysis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are grateful to the editor and anonymous reviewers for their valuable suggestions. This research was funded by Beijing Science and Technology Plan (no. Z121100000312101).

References

- [1] Transportation Research Board, 1965 *Highway Capacity Manual*, Transportation Research Board of the National Academies, Washington, D C, USA, 1965.

- [2] Transportation Research Board, *Highway Capacity Manual 2010*, Transportation Research Board of the National Academies, Washington, DC, USA, 2010.
- [3] T. Sasaki and Y. Iida, *Traffic Engineering*, Ohmsha, Tokyo, Japan, 1992.
- [4] “Code for transport planning on urban road,” GB 50220-1995, CN-GB, Beijing, China, 1995.
- [5] K. Choocharukul, K. C. Sinha, and F. L. Mannering, “User perceptions and engineering definitions of highway level of service: an exploratory statistical comparison,” *Transportation Research Part A: Policy and Practice*, vol. 38, no. 9-10, pp. 677–689, 2004.
- [6] C. C. Xu, P. Liu, W. Wang, and Z. B. Li, “Evaluation of the impacts of traffic states on crash risks on freeways,” *Accident Analysis and Prevention*, vol. 47, pp. 162–171, 2012.
- [7] L. Wang, Y. Z. Fan, and H. Zhang, “A new traffic state evaluation approach based on probe vehicle and multi-classification fuzzy pattern recognition,” *Journal of Highway and Transportation Research and Development*, vol. 24, no. 9, pp. 92–95, 2007.
- [8] B. S. Kerner, “Three-phase traffic theory and highway capacity,” *Physica A*, vol. 333, no. 1–4, pp. 379–440, 2004.
- [9] H. Li, J. Lu, G. Jiang, and Y. Ma, “Traffic incident detection algorithm for expressway,” *Journal of Southeast University*, vol. 43, no. 3, pp. 649–653, 2013.
- [10] W. Guan and S. Y. He, “Phase identification of urban freeway traffic based on statistical properties,” *Journal of Transportation Systems Engineering and Information Technology*, vol. 7, no. 5, pp. 42–50, 2007.
- [11] Y. Q. Liao, L. M. Jia, C. X. Li, H. H. Dong, and Y. Qin, “A perceptual experiment study on level of service for regional road traffic,” *Journal of Beijing Jiaotong University*, vol. 7, no. 5, pp. 42–50, 2007.
- [12] J. Xia, W. Huang, and J. Guo, “A clustering approach to online freeway traffic state identification using ITS data,” *KSCE Journal of Civil Engineering*, vol. 16, no. 3, pp. 426–432, 2012.
- [13] Z. W. Qu, Q. Wei, Y. M. Bie, H. Zhu, and D. H. Wang, “Method for traffic state identification based on fixed detector,” *Journal of Central South University (Science and Technology)*, vol. 44, no. 1, pp. 403–410, 2013.
- [14] X. L. Sun, H. H. Dong, M. Guo, S. W. Lin, J. Qiao, and P. Zheng, “The design and implementation of Beijing regional traffic conditions and LOS evaluation system, road traffic safety,” *Road Traffic & Safety*, vol. 10, no. 4, pp. 36–40, 2010.
- [15] Y. S. Quan, J. F. Guo, J. Z. Guan, L. Yu, and H. M. Wen, “Research on traffic congestion evaluation and empirical analysis of Beijing,” in *Proceedings of the 3rd China Annual Conference on ITS*, pp. 1–6, 2007.
- [16] X.-L. Liu, P. Jia, S.-H. Wu, and B. Yu, “Short-term traffic flow forecasting based on multi-dimensional parameters,” *Journal of Transportation Systems Engineering and Information Technology*, vol. 11, no. 4, pp. 140–146, 2011.
- [17] H.-H. Dong, X.-L. Sun, L.-M. Jia, H.-J. Li, and Y. Qin, “Traffic condition estimation with pre-selection space time model,” *Journal of Central South University of Technology*, vol. 19, no. 1, pp. 206–212, 2012.
- [18] M. Canaud, L. Mihaylova, J. Sau, and N.-E. El Faouzi, “Probability hypothesis density filtering for real-time traffic state estimation and prediction,” *Networks and Heterogeneous Media*, vol. 8, no. 3, pp. 825–842, 2013.
- [19] T. L. Pan, A. Sumalee, R. X. Zhong, and N. Indra-Payoong, “Short-term traffic state prediction based on temporal-spatial correlation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1242–1254, 2013.
- [20] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, “Dynamic data-driven local traffic state estimation and prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 89–107, 2013.
- [21] S. Zhang, Z. X. Huang, and A. W. Kuang, “Estimation of real-time traffic state of city expressway under incomplete information,” *Systems Engineering*, vol. 31, no. 5, pp. 16–21, 2013.
- [22] Y. Guo, X. Zhang, B. Dong, and L. Wu, “Traffic state of rapid identification and transitions shift method based on fuzzy theory,” *Journal of Central South University (Science and Technology)*, vol. 44, no. 1, pp. 1–5, 2013.
- [23] H. B. Shan, S. X. Li, D. G. Gong, and P. Lou, “Genetic simulated annealing algorithm-based assembly sequence planning,” in *Proceedings of the International Technology and Innovation Conference (ITIC '06)*, pp. 1573–1579, November 2006.
- [24] K.-L. Zhou and S.-L. Yang, “An improved fuzzy C-Means algorithm for power load characteristics classification,” *Power System Protection and Control*, vol. 40, no. 22, pp. 58–63, 2012.
- [25] D. W. Xu, L. M. Jia, and X. L. Sun, “Design method of road traffic information template,” *Journal of Highway and Transportation Research and Development*, vol. 30, no. 12, pp. 132–137, 2013.
- [26] L. G. Sun, R. M. Li, S. Dong, and H. P. Lu, “Study on short-term traffic flow combined forecast method,” *Journal Wuhan University of Technology (Transportation Science & Engineering)*, vol. 34, no. 5, pp. 874–876, 2010.
- [27] H. P. Lu, Y. G. Sui, M. Guo, and R. M. Li, *Urban Mixed Traffic Flow Analysis Model and Method*, China Railway Publishing House, Beijing, China, 2009.
- [28] F. Shi, H. Wang, L. Yu, and F. Hu, *Thirty Cases Analysis of Intelligence Algorithm Based on Matlab*, Beihang University Press, Beijing, China, 2010.