IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 2, 2019, accepted December 23, 2019, date of publication December 27, 2019, date of current version January 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962619

A Spatiotemporal Apriori Approach to Capture Dynamic Associations of Regional Traffic Congestion

DONG-FAN XIE[1], **MEI-HONG WANG**[2], **AND XIAO-MEI ZHAO**[1]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[2]Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Xiao-Mei Zhao (xmzhao@bjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1600900, and in part by the National Natural Science Foundation of China under Grant 71621001, Grant 71671014, Grant 71771012, and Grant 71631007.

ABSTRACT Due to the interactions among adjacent roads in urban road networks, traffic congestion gradually propagates to neigboring roads, resulting in regional congestion. To develop advanced regional traffic control strategies, it is necessary to clearly understand the characteristics of regional congestion evolution. To this end, this paper proposes a data-driven approach to mine the spatiotemporal associations of regional traffic congestion. By introducing both time and space attributes, the intra-transaction spatiotemporal Apriori (IntraT-ST-Apriori) algorithm is developed to address the static features of regional traffic congestion; while the inter-transaction spatiotemporal Apriori (InterT-ST-Apriori) algorithm is developed to capture the dynamic characteristics of regional traffic congestion. Case studies are carried out for the urban road network in Tianjin, China, based on empirical data. The results indicate that the Intra-ST-Apriori algorithm can excavate the underlying associations of regional traffic congestion. Furthermore, the congestion propagation trajectories can be clearly revealed based on the InterT-ST-Apriori algorithm. It is expected that the proposed approach can support the regional traffic management and control, significantly relieving traffic congestion.

INDEX TERMS Traffic congestion, association rule, congestion propagation, apriori algorithm.

I. INTRODUCTION

As the increasing number of vehicles, traffic congestion, particularly in rush hours, becomes one of the most essential issues in most large cities all over the world. For an urban traffic system, the roads connect with each other, forming a complex road network. Traffic flows of adjacent roads interact with each other. Traffic congestion on a road can gradually spread to adjacent roads, leading to regional traffic congestion. To this end, it is very necessary to explore the underlying characteristics of traffic congestion association and propagation, which are the basis for regional traffic management and control.

Due to the lack of empirical data, primitive studies were conducted based on traffic simulations in terms of various traffic flow models, such as the link transmission model [1], [2], the grid transmission model [3], [4], cellular automaton models [5], [6], and so on. On the basis, the occurrence, propagation and dissipation of traffic congestion were extensively investigated on urban road networks [7], [8]. Although the characteristics and mechanism of network traffic congestion can be easily obtained based on simulation models, there are some drawbacks for these approaches. For instance, most of the models are not calibrated and validated based on empirical traffic data; Simulation models cannot take into account some detail and essential factors in real traffic, such as traffic interaction of vehicles at intersections and dynamic origin-destination; In particular, most simulation models cannot be applied on large urban networks due to the requirement of huge amount of computation.

With the rapid development of new technologies, such as sensoring, information communication, and computer, large amounts of traffic data can be collected in terms of various detectors and mobile equipment, which provide new approaches to investigate network traffic flow based on empirical traffic data.

Based on network traffic data, early works employed statistical methods to investigate associations of traffic congestion

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

in terms of adjacent road sections [9]–[12]. However, most of the statistical indicators requires specific distributions of empirical traffic data, such as linear or normal distribution. Unfortunately, urban traffic flow data are generally noisy, unstable and non-linear, resulting in lower reliability of the methods.

Recently emerging machine learning approaches provide us with an alternative opportunity to deeply understand various traffic phenomena. In recent years, machine learning approaches have attracted much attention, which has been extensively applied in traffic flow modeling and prediction due to the capability to address high dimension big data [13]. Based on various data sources, i.e., floating car data, fixed sensor data, camera data, some studies were conducted to mine the characteristics of urban traffic congestion by applying such mechanism learning methods as causality trees [14], associate rule learning [15], deep learning network [16], [17], and so on. However, most of these studies addressed urban traffic congestion with the consideration of neither spatial nor temporal associations. Few works were conducted for spatiotemporal congestion associations of urban road networks.

Since traffic congestion may propagate along road networks, there exist associations among different congested regions. To this end, many studies have been dedicated to this topic, attempting to uncover the underlying mechanism of congestion association and propagation. Nevertheless, there are still some open questions closely related to this study: (1) How to recognize traffic state based on large amount of empirical traffic data? (2) What are the underlying characteristics of spatiotemporal associations for regional traffic congestion? (3) How to reveal the spatiotemporal trajectory of regional congestion propagation?

To address the abovementioned problems, this study develops a data-driven approach based on the Apriori algorithm, and case studies are carried out based on empirical data of Tianjin, China. The results indicate that the proposed algorithm can well capture the spatiotemporal characteristics of regional congestion, which expected the proposed approach would be helpful for the regional traffic management and control, and significantly relieve traffic congestion. The motivations and contributions of this study are as follows.

- An alternative clustering algorithm is developed to recognize regional traffic state by integrating the classical k-means++ and FCM algorithms. In detail, the clustering result of the k-means++ algorithm is set to be the initial clustering centers of the FCM algorithm. Therefore, the proposed algorithm can overcome the drawbacks of the FCM algorithm, i.e., large computational amount, low efficiency, and local convergence.
- By introducing both time and space attributes, the intra-transaction spatiotemporal Apriori (IntraT-ST-Apriori) algorithm is developed to address the static association features of regional traffic congestion within a specific period. The inter-transaction spatiotemporal association (InterT-ST-Apriori) algorithm is developed to capture the trajectories of regional traffic congestion propagations based on empirical data. Namely, the InterT-ST-Apriori algorithm takes into account both spatial and temporal associations of regional congestion, which has not been deeply studied in existing literature.
- Both IntraT-ST-Apriori and InterT-ST-Apriori are data-driven algorithms based on machine learning, which have the capability to excavate underlying characteristics of regional congestion based on large amount of empirical data. In particular, the success of machine learning is one of the motivations. That is, whether machine learning approaches could improve the understand of regional congestion associations?
- Based on large amount of empirical data from Tianjin, China, case studies are carried out. The results validate the capability of the proposed approaches.

The rest of the paper is organized as follows: In Section II, the literature review is presented. In Section III, the IntraT-ST-Apriori and InterT-ST-Apriori Algorithms are developed to address the association and propagation of regional traffic congestion. In Section IV, the empirical data used in this study is introduced, and the approach for traffic state recognization is proposed. In Section VI, case studies are carried out to validate the proposed approaches. Finally, conclusions are made in Section VII.

II. LITERATURE REVIEW

Since network traffic congestion is a representative and essential issue in real traffic, many studies have been made on this topic with various approaches, including simulation based approaches, statistical methods, machine learning methods, and so on. In the following part of this section, brief review on network traffic congestion will be conducted.

A. MODELING CONGESTION PROPAGATIONS BASED ON SIMULATIONS

Traffic simulation is a classical approach to investigate traffic-related phenomena. So far, various traffic flow models have been developed for urban network traffic flow [1]–[4], [6], and some of them have been conducted to capture the dynamic characteristics of network traffic congestion.

Macroscopically, the cell transmission model (CTM) is one of the most prevalent models to address network traffic flow dynamics. Based on the CTM, various models have been proposed to investigate traffic congestion propagation and the underlying mechanisms. Reference [7] employed the CTM to model both link and node traffic flow, and proposed a method to estimate average journey velocity. On the basis, network traffic flow can be simulated to identify network congestion bottlenecks. Reference [8] applied the CTM to simulate the formation and dissipation of congestion in a two-way rectangular grid network. Reference [18] developed a traffic state estimation method by combining the CTM and extended Kalman filter (EKF) recursive algorithm. The CTM is used to predict traffic density, and the EKF recursive algorithm is used to deal with empirical traffic data from sensors.

As a typical microscopic traffic flow model, cellular automaton (CA) model takes advantages in computational efficiency and extendibility. As a consequence, CA has been extensively applied in network traffic flow modeling to discover characteristics of network traffic dynamics. Reference [6] developed the primary two-dimension CA model to simulate traffic flow on a grid network. Reference [19] proposed a two-dimension CA model by combining the NaSch model [5] and the BML model, attempting to investigate traffic state transition from free flow to completely jam on a grid network. Likewise, several two-dimension CA models have been proposed to investigate traffic state evolution with the consideration of various impacting factors, such as origin-destination trips [20], traffic lights [21], network operation reliability [22], and so on.

B. STATISTICAL METHODS FOR CONGESTION ASSOCIATION

Based on empirical traffic data, early works employed statistical methods to investigate congestion association among different road segments. Reference [11] introduced the entropy plot to measure the correlation, and further proposed a statistical model to capture the spatiotemporal correlation of traffic states. Reference [9] developed a Bayesian based methodology to model the correlation of travel times between links along a corridor. Reference [10] proposed a statistical model to study link speed correlation based on empirical data. The model is simple and efficiency. However, normal distributed data is assumed. Reference [12] utilized the Pearson correlation coefficient to measure the speed correlation of different roads in time and space, respectively. In summary, most of the statistical methods requires specific distributions of empirical traffic data, such as linear or normal distribution, which limits the application and reliability of these methods. Reference [23] proposed a method to analyze spatiotemporal correlations of road traffic states in terms of both Pearson and Entropy based methods. The results demonstrated Entropy can be applied in more general cases of distribution of data, while Pearson correlation can only address linear distribution of data.

Based on statistical methods, the correlation of congestion among different roads can be clearly illustrated. However, these methods cannot address the dynamic evolution of congestion on urban network. Moreover, most of the statistical indicators requires linear or normal distribution of empirical traffic data, while empirical traffic data is generally noisy, unstable and non-linear and thus cannot satisfy the requirements.

C. MACHINE LEARNING APPROACHES FOR TRAFFIC CONGESTION

In recent years, various algorithms have been conducted to investigate spatiotemporal correlations among different datasets based on machine learning, such as the fuzzy method [24], the rule-mining algorithms [25], [26], the topic-based temporal mining approach [27], various clustering methods [28], [29], and so on. Since the machine learning performs well in big data mining, it is regarded to be helpful to promote traffic-related studies.

Taking the advantages of machine learning and sufficient amount of empirical traffic data, various data-driven approaches have been developed to mine the underlying characteristics of network congestion. To validate the predictability of urban traffic patterns, Reference [30] proposed a method in terms of taxi trajectory data by mapping the road congestion degree into a time series of symbols and measuring its entropy. Reference [16] employed deep Restricted Boltzmann Machine and Recurrent Neural Network architecture to model traffic congestion propagation based on Global Positioning System (GPS) data of taxis. The proposed method could well reveal the spatial distribution of congestion on urban road networks. However, the spatiotemporal association of congestion cannot be obtained. Reference [31] proposed a data-driven approach based on dictionary-based compression theory to identify spatial and temporal patterns of road networks. However, the method has some limitations. For instance, it cannot work better with small dataset; and it is very sensitive to the outliers. Reference [32] introduced the method of voting and ranking to address the uncertaint performance of the road link based on speed dataset, which can be applied to identify the dynamic characteristics of bottlenecks. Reference [33] used Fuzzy C-means clustering to conduct congestion patterns of road segments, and employed spatial autoregressive moving average (SARMA) model to discover the relationship between built environment and congestion pattern. Nevertheless, the spatiotemporal evolution of congestion cannot be well revealed. Reference [34] proposed a method to predict congestion propagation based on greedy algorithm. The method utilizes large amount of camera data to mine congestion association among adjacent road links. Reference [35] developed a method to model congestion propagation in terms of a space-temporal congestion subgraph, and validated the proposed method based on taxicab data in Shanghai.

Clustering is a typical data mining method which has been conducted in exploring the characteristics of traffic congestion. Reference [36] developed a method by clustering spatiotemporally overlapping episodes to detect non-recurrent congestion events. However, there are some limitations of the method. For instance, it may induce a prediction do not truly have a physical significance, because it employed only the congestion factor to identify congestion levels. Reference [37] developed a dynamic clustering methodology to describe the evolution process of congestion formation and dissolution on urban road networks. However, this method requires a prior knowledge on the number of clusters for partitioning the network; and this method cannot well capture congestion dynamics because it considers vehicles only on links and neglects the impact of other factors such as signalized intersections.

To further improve the understand of spatiotemporal association and dynamics of urban traffic congestion, several multi-step methods have been developed. Reference [14] developed a method to detect spatiotemporal congestion of roads and to mine the causal relationships in terms of empirical datasets. The proposed method comprises of three components: Causal congestion trees construction; Frequent congestion subtree discovery; and Dynamic Bayesian Network based traffic congestion propagation model. Nevertheless, the method requires detailed road traffic information which cannot easily be obtained and the dataset may be not sufficiently accuracy. Reference [38] developed a three-phase framework to explore congestion correlation among road links based on both taxi trajectory data and point of interest (POI) data. Reference [39] proposed an integrated stepwise method to recognize evolution patterns of recurrent regional traffic congestion based on taxi GPS trajectory data. However, the method cannot provide the associations among traffic states in different regions. In addition, the accuracy of the method cannot be guaranteed because they considered only taxi data which is very small part of sampling of the whole traffic data.

Association rule is a typical machine learning algorithm which has been extensively applied in many areas such as, traffic accidence analysis, traffic congestion prediction, excavation of stock exchange rules, patent excavation clinical excavation, and so on. Regarding to urban traffic congestion associations, several rule based approaches have been developed. Reference [40] proposed a method of comparative association rules mining based on genetic network programming (GNP) to uncover association rules among different empirical traffic data. Reference [41] proposed a Apriori based algorithm in terms of social media data to mine spatiotemporal correlation of congestion among road segments.

D. REGIONAL CONGESTION ASSOCIATION

In recent years, it has been validated the existing of well-defined fundamental diagram (MFD) for urban road networks from both theoretical analysis and experimental studies [42], [43]. Then, regional based traffic flow models attracted much attention, which provide an alternative approach to investigate network traffic congestion. Reference [44] proposed causality trees based algorithms to detect spatiotemporal traffic state dynamics of sub-regions of a road network. The proposed algorithm was validated based on taxi trajectory data of Beijing, indicating that it could identify anomalous traffic states as well the corresponding spatiotemporal dynamics. Reference [45] partitioned a city into grid grids, and employed the likelihood ratio test statistic approach to depict traffic patterns based on GPS data from taxis. However, the method cannot be extensively applied in practice since it needs appropriately sized datasets. Reference [46] proposed a dynamic method to identify traffic congestion of heterogeneous urban road networks. The method is mainly composed of two steps. First, it generates a directed weighted network in terms of network connectivity and traffic load of the network; Then, it performs a detection algorithm consisting of three sub-steps: generation of congestion regions, expansion and regression of congested regions, and merging adjacent congestion regions.

III. METHODOLOGY FOR REGIONAL CONGESTION ASSOCIATION AND PROPAGATION

In this paper, the data-driven approach is applied to extract the underlying spatiotemporal characteristics of regional congestion. In brief, the proposed methodological framework consists of the following components:

- Regionalism and regional traffic state recognition. First, the urban area is divided into grids. Then, traffic variables (i.e., traffic flux, speed) are averaged for all the roads within each grid. Accordingly, traffic state for each grid can be recognized.
- Association of regional traffic congestion. The IntraT-ST-Apriori algorithm is developed to uncover the correlations of traffic states among different grids. In this case, the association algorithm is static. Namely, the IntraT-ST-Apriori algorithm is developed for a specific time interval. Therefore, the correlations between congestion grids can be excavated at each time interval.
- regional traffic congestion propagation. To further understand the dynamic evolution of regional traffic congestion, the InterT-ST-Apriori algorithm is developed by introducing both spatio and temporal attributes. To this end, congestion propagations can be excavated, which is regarded to be the basis for development of regional traffic control strategies.

The spatiotemporal state of event (traffic congestion) is closely related to the length of time interval. A too small time interval may results in large amounts of detail rules which would not be necessary; while a too large time interval may neglect some of the important association rules. Therefore, it is necessary to choose an appropriate time interval to balance the accuracy and efficiency of mining process. In terms of the characteristics of urban traffic congestion, the time interval is set to be 15 minutes in the following case studies.

In this study, the approach is developed based on the classical Apriori algorithm. Therefore, the fundamental concepts will be briefly introduced.

A. BASIC CONCEPTS

Apriori algorithm, primally developed by [25], is a classical approach to investigate association rules.

Definition 1 (Item and Itemset): Itemset $I = (I_1, I_2, \ldots, I_m)$ is a set of items consisting of m items. An item $I_j (j = 1, 2, \ldots, m)$ denotes one of the items in itemset I. I_k is a subset of itemset I consisting of k items, which is named as $k - itemset$.

In this study, an item is regarded as the congestion state of a grid at a time interval. Totally, the number of items can be calculated as $N_G \times N_{tstep}$. Here, N_G denotes the number of grids; and N_{tstep} is the number of time intervals.

Definition 2 (Transaction): Each transaction $T_i (i = 1, 2, \ldots, n)$ $(T_i \subseteq I)$ is comprized of a set of items, relating to an unique identifier *TID*. The transaction set $D = \{T_1, T_2, \ldots, T_n\}$ is a set of transactions.

Specifically, a transaction is comprised of congestion items of some grids. A transaction is a subset of itemset I.

Definition 3 (Support): For itemsets X and Y belonging to transaction set D, support $sup(X \Rightarrow Y)$ indicates the frequency of the association rule in the transaction set D, and it is expressed as,

$$\sup(X \Rightarrow Y) = \frac{frq(X \cup Y)}{N} \tag{1}$$

where $frq(X \cup Y)$ represents the number of transactions containing both X and Y; N denotes the total number of transactions in transaction set D.

In applications, the association rules should be larger than the user-specified value, namely the minimum support *minsup*. In general, the larger the minimum support is, the fewer the association rules are.

Definition 4 (Confidence): Assume that transaction set D contain X ($X \subseteq D$) and Y ($Y \subseteq D$). Confidence is the conditional probability of Y appearing in a transaction containing X, which is formulated as,

$$conf(X, Y) = \frac{\sup(X \cup Y)}{\sup(X)} \tag{2}$$

where $X \subseteq I, Y \subseteq I$, and $X \cup Y = \Phi$. The minimum confidence *minconf* should be satisfied for an association rule.

Definition 5 (Lift): Lift indicates the relationship between X and Y. It reads,

$$Lift(X, Y) = \frac{\sup(X, Y)}{\sup(Y)^{*} \sup(X)} \tag{3}$$

When the lift is larger than 1, the rule $X \Rightarrow Y$ is strong association rule, and the two itemsets X and Y are strongly dependent with each other; while if the lift is less than 1, the existence of item X has a reverse inhibition effect on Y. In particular, if the lift equals 1, the itemsets X and Y are independent with each other. Mathematically, lift can be expressed as follows:

$$Lift(X, Y) = \frac{P(Y|X)}{P(Y)} \tag{4}$$

For instance, the rule $\{Rail-station, Congestion\} \Rightarrow \{Bus-station, Congestion\}$ with lift 1.1 indicates that the congestion of rail station may result in the occurance of congestion of the bus station.

Definition 6 (Frequent Itemset): A frequent itemset is a non-empty subset of itemset I whose support value is equal to or larger than the minimum support *minsup*. In general, a frequent itemset with item number of m is called as frequent $m-itemset$, and it is denoted as L_m.

Definition 7 (Strong Association Rule): A strong association rule $X \Rightarrow Y$ indicates the association rule to be mined, which satisfies both the following criterions:

$$sup(X \Rightarrow Y) \geq minsup \tag{5}$$

and

$$conf(X \Rightarrow Y) \geq minconf \tag{6}$$

Definition 8 (Itemset Property): Any infrequent $(k-1)-itemset$ cannot be a subset of the frequent $k-itemset$.

On the basis, association rule mining can be concluded as follows: excavate the strong association rules (for example, $X \Rightarrow Y$) in the transaction set D. Here, X and Y are itemsets belonging to transaction set D, which are named as antecedent and consequent, respectively. In general, it is difficult to directly excavate strong association rules from the transaction set D due to the large amount of possible rules. For instance, for a transaction set with d transactions, the number of possible rules is $AR = 3^d - 2^{d+1} + 1$, which increases rapidly as d increases. To address this issue, the Apriori algorithm has been developed [25], which divides the problem into two sub-processes:

- Searching for frequent itemsets whose supports are equal to or larger than the user-specified minimum support.
- If the confidences of frequent itemsets are equal to or larger than the minimum confidence, the association rules are generated by disassembling frequent itemsets.

B. INTRA-TRANSACTION SPATIOTEMPORAL APRIORI ALGORITHM

Different from the basic issues addressed by most association rule approaches, traffic congestion may propagate upstream or downstream with time. That is, traffic congestion is closely related to time and space, which should be added as attributes to transactions. To this end, the intra-transaction spatiotemporal (IntraT-ST) association rule is defined; See Definition 9. According to the definition, both spatio and temporal attributes are embedded into the transactions. As well, the support and confidence should be re-defined; See Definitions 10 and 11, respectively.

Definition 9 (IntraT-ST Association Rule):

$$\begin{aligned} &P_1 \cap P_2 \cap \ldots P_m, time \\ &\qquad \Rightarrow Q_1 \cap Q_2 \cap \ldots Q_n, time, (minsup, minconf) \end{aligned} \tag{7}$$

where P_i and Q_i denote the locations of congestion grids; *time* denotes the time interval. Equation 7 indicates that the correlations between $X = [P_1 \cap P_2 \cap \ldots P_m, time]$ and $Y = [Q_1 \cap Q_2 \cap \ldots Q_n, time]$ under the constraints of minimum support and confidence.

Definition 10 (IntraT-ST-Support): IntraT-ST-Support indicates the frequency of association rules appearing in the transaction set with all transactions taking spatio and temporal attributes. It is expressed as,

$$sup\,(I\,(o_i, t)) = \frac{frq\,(I\,(o_i, t))}{N[t]} \tag{8}$$

where $frq(I(o_i, t))$ is the number of transactions containing itemset $I(o_i, t)$; o_i denotes the grid location; and t denotes the

TABLE 1. Transaction set.

TID	*TF*	Characteristic Value	*Location*
1	t_1	v_1	o_1
2	t_2	v_2	o_2
...	...	...	...
n	t_n	v_n	o_n

time interval. $N[t]$ represents the total number of transactions at t.

Definition 11 (IntraT-ST-Confidence): Based on the Definition 10, IntraT-ST-Confidence is defined as follows,

$$conf\left(I\left(o_i,t\right),I\left(o_j,t\right)\right)=\frac{sup\left(I\left(o_i,t\right),I\left(o_j,t\right)\right)}{sup\left(I\left(o_i,t\right)\right)} \tag{9}$$

where $o_i \neq o_j$.

Lift, as in Definition 5, is an extensively used index for association rules. Nevertheless, lift may be not confident due to the impact of zero-sum transactions (i.e., the transactions with non items), and the performance is asymmetric near the critical value 1. As a consequence, *Kulc* and *IR* will be employed together as the performance index of association rules; See Equations 10 and 11 for details.

Definition 12 (Association Index): For itemsets X and Y belonging to transaction set D, *Kulc* and *IR* can, respectively, be formulated as follows:

$$Kulc = 0.5\times(conf(X,Y)+conf(Y,X)) \tag{10}$$

$$IR=\frac{|sup(X)-sup(Y)|}{sup(X)+sup(Y)-sup(X\cap Y)} \tag{11}$$

In terms of Equation 10, X and Y are positively correlated as *Kulc* is close to 1, indicating that the occurrence of X may lead to the occurrence of Y. According to Equation 11, the equilibrium state can be obtained as *IR* is close to 0, which indicates that X and Y may impact on each other.

To investigate the spatio and temporal characteristics of traffic congestion, the spatiotemporal congestion itemsets should be designed. Since the association rules approach to discrete problems, the characteristic values have to be discretized, resulting in discrete transaction itemset $D=\{Time\ Interval\ (TF), Characteristic\ Value, Location\}$, as shown in Table 1. All the time intervals have the same time length. The characteristic value v_i, $(i=1,2,\ldots,N_I)$ represents the traffic state in grid i, where N_I denotes the number of possible traffic states.

Based on the above definitions, the IntraT-ST-Apriori algorithm can be presented, which is mainly consist of two components, i.e., generation of frequent intmsets and generation of association rules; See Algorithm 1 for details.

In Algorithm 1, Step 4 illustrates the generation of frequent itemsets, which contains two major steps, namely the join and prune of the itemsets. In detail, the generation of frequent itemsets are conducted as follows:

```
Algorithm 1 IntraT-ST-Apriori Algorithm
  Step 1: Divide the urban road network into grids, and generate the transaction set based on grid traffic states.
  Step 2: Add temporal attribute into congestion transaction set, resulting in spatiotemporal congestion transaction set.
  Step 3: Let $k = 1$. Calculate the lift of candidate $k-itemset$, and generate the frequent $k-itemset$ for each time interval in terms of the user-specified minimum support. Then, let $k = k + 1$.
  Step 4: For frequent $(k-1)-itemsets$ in each time interval, generate $k-itemsets$ based on join step. Then, perform the prune step based on the Apriori algorithm. Scan the transaction set again to obtain the support of candidate $k-itemsets$ for each time interval. Generate frequent $k-itemsets$ in terms of the user-specified minimum support.
  Step 5: Repeat Step 4 until new frequent candidate sets cannot be generated or the generated candidate sets do not satisfy the specified minimum support. Then, the generation of spatiotemporal frequent $k-itemsets$ is completed.
  Step 6: Calculate the confidence of each frequent $k-itemset$ for each time interval. Obtain the strong association rules, whose confidences are equal to or larger than the minimum confidence.
  Step 7: Calculate the association index based on Equations 10 and 11 to present the correlation between the antecedents and consequents of the association rules.
```

- Join step of Intra-T-ST-Apriori Algorithm: Join step is applied to generate frequent $k-itemsets$, denoted as L_k. The input of join step is $(k-1)-itemsets\ L_{k-1}$. Let l_i $(i=1,2,\ldots,k-1)$ represents the ith item in L_{k-1}. Join any two l_i and l_j $(i\neq j)$ of L_{k-1} to generate candidate $k-itemsets\ C_k$.
- Prune step of Intra-T-ST-Apriori Algorithm: C_k is a superset of L_k. That is, C_k contains all frequent $k-itemsets$, and additional $k-itemsets$ which are not frequent *itemsets*. Obviously, L_k can be obtained by calculating the support of each $k-itemset$ in C_k and discarding the ones whose supports are smaller than the user-specified minimum support. Nevertheless, the large size of C_k may result in huge amount of calculation. Therefore, Definition 8 is applied to compress the size of C_k. That is, if the $(k-1)-itemset$ of a candidate $k-itemset$ is not in L_{k-1}, the candidate $k-itemset$ is not frequent, which should be removed from C_k. Generally, this process is regarded as prune.

Finally, the generation of association rules in Step 6 of Algorithm 1, which is the major purpose of the IntraT-ST-Apriori algorithm, can be detailed as,

- Find all the non-empty subsets I_i $(i=1,2,\ldots,k)$ of L_k, and divide them into two subsets $I_i(t_i,o_i)$ and $I_j(t_j,o_j)$ $(i\neq j)$ with a total of $k-1$ cases. For $I_i(t_i,o_i)$ and $I_j(t_j,o_j)$, calculate the confidence based on Equation 9. The strong association rule can thus be obtained if the confidence is equal to or larger than the specified minimum confidence.

C. INTER-TRANSACTION SPATIOTEMPORAL APRIORI ALGORITHM

The dynamic evolution of congestion is a typical characteristic of network traffic flow, which may seriously impact on the efficiency of road networks. To this end, the Inter-transaction spatiotemporal Apriori (InterT-ST-Apriori) algorithm is developed to capture the propagation trajectories of regional traffic congestion.

Definition 13 (Time-Series Congestion Set): Let $S = \{s_1, s_2, \ldots, s_n\}$ denotes the time-series set; $T_i = \{s_1(i), s_2(i), \ldots, s_n(i)\}$ represents the set of observations to S at time interval i. Multiple time-series congestion set can be defined as $D = \{T_1, T_2, \ldots T_n\}$, where each set of observations in D is regarded as a transaction set, with an identifier of *TID*

Definition 14 (InterT-ST Association Rule): Let $\sum = \{e_1(0), e_1(1), \ldots, e_1(\omega-1), e_2(0), e_2(1), \ldots, e_2(\omega-1), \ldots, e_u(0), \ldots, e_u(\omega-1)\}$ denote the congestion transaction set, with each element representing the attribute of observations in time series transaction set D. ω is the length of sliding time window of D. Let $s(1 \leq s \leq n-\omega+1)$ represent the initial reference time of D. If congestion transaction e_i takes place at time $s+x$ $(0 \leq x \leq \omega-1)$, denote $e_i(x) \in T_s$, and set an unique identifier *TID* to it. Then, the InterT-ST association rule can be expressed as $X \Rightarrow Y$, which satisfy the following criterions:

(1) $X \in \sum$, and $Y \in \sum$;
(2) $\exists e_i(0) \in X, 1 \leq i \leq u$;
(3) $\exists e_j(q) \in X, 1 \leq j \leq u, ((i \neq j) \wedge (1 \leq q \leq \omega-1))$;
(4) $\exists e_i(p) \in Y, 1 \leq i \leq u, \max(q) < p \leq \omega-1$;
(5) $X \cap Y = \varnothing$;
(6) $X = [t_i, o_i, v_1]$, $Y = \left[t_j, o_j, v_1\right]$; X and Y satisfy with Definition 16.

To mine the association rules, it is necessary to construct transaction sets to be mined. Based on transaction set of IntraT-ST association rules and the specified time window, a number of sliding time window transaction sets with different starting time are formed in terms of Definition 15.

Definition 15 (Sliding Time-Window Transaction Set): Arrange the transaction sets in chronological order, and the sliding time-window transaction sets are composed of itemsets of consecutive ω time intervals.

In this study, the time series is $t = [7:00,\ 7:15,\ 7:30, \ldots,\ 22:00]$. The time series set S contains m elements as well as the time series t. Then, based on Definitions 13 and 14, $\sum$ can be obtained and the sliding time-window transaction sets can be generated with various initial time.

In applications, an appropriate time-window should be determined. A too large time-window may lead to lower computational efficiency, while a too small time-window may neglect the temporal impact on the concerned issues. In general, the time-window size ω can be selected in terms of research purpose and the size of transaction sets. Intuitively, one can see Figure 1 for the sketch of sliding time-window transaction set with initial time of i.

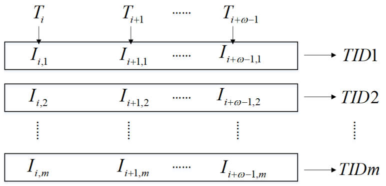


FIGURE 1. Sliding time-window transaction set with initial time of *i*. Here, T_i represents the congestion grids at time *i*; *m* is the number of time series; ω denotes the length of sliding time-window; and *TID* is the identifier.

The support and confidence of InterT-ST association rules are similar to those of IntraT-ST association rules. Nevertheless, each transaction should have attributes of time t_i and location o_i. Therefore, for the calculation of these parameters, the items can be regarded to be the same only when they have the same spatio and temporal attributes.

Traffic congestion in a grid may propagate upstream or downstream, and thus impacts on traffic state of surrounding grids. Therefore, when mining strong spatiotemporal association rules of congestion grids, the antecedents and consequents of rules need to satisfy both the temporal continuity and spatial connectivity. According to the characteristics of congestion evolution, the temporal constraint, the spatio constraint and attribute constraint are proposed, as described in Definitions 16, 17 and 18, respectively.

Definition 16 (Temporal Constraint):

$$\Delta t = t_i^d - t_j^k = \begin{cases} (i-j) \times \lambda, & (d=k, i>j) \\ (j-i) \times \lambda, & (d=k, i<j) \\ \infty, & d \neq k \end{cases} \tag{12}$$

where Δt denotes difference of the time interval; t_i^d (t_j^k) denotes the time interval i (j) in day d (k); and λ is the size of time interval, representing the continuity of time.

Definition 17 (Spatio Constraint): Spatio associations mainly include topology, distance and orientation, which can be depicted by spatio predicates. Spatio topological associations are adjacent, connected and overlapping relationships among grids. Distance describes grids with predicates such as "far away" and "adjacent". Orientation represents grids with predicates such as upper and lower, left and right, front and back, east, west, South and north. The congestion in a grid can only spread to the surrounding grids, i.e., spatio adjacent grids, as shown in Figure 2.

Definition 18: Attribute constraint

$$(o_i, t_i, 0) \Rightarrow (o_i, t_{i+1}, 1) \tag{13}$$

where o_i $(i = 1, 2, \ldots, n)$ denotes the grid location; 0 represents non-congestion, while 1 represents congestion. Equation 13 indicates that traffic state of grid o_i turns to congestion from t_i to t_{i+1};

Based on the aforementioned definitions, the InterT-ST-Apriori algorithm can be developed with three components: (1) Generate sliding time-window transaction set based on

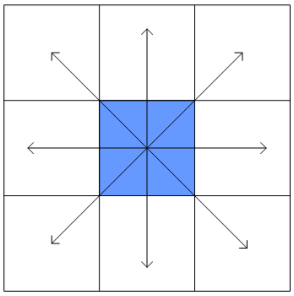


FIGURE 2. Spatio relationships.

Algorithm 2 InterT-ST-Apriori Algorithm

Step 1: Input congestion transaction set; Set the minimum support, the minimum confidence, the length of time interval λ, and the size of sliding time-window ω.

Step 2: Generate the sliding time-window transaction set according to Definition 15, as shown in Figure 1. Applied InterT-ST association rules to the generated sliding time-window transaction set.

Step 3: Calculate the support of each itemset of the transaction set, and obtain the frequent $1 - itemsets$ in terms of the specified minimum support.

Step 4: Apply the join and prune steps to frequent $1 - itemsets$. Thus, frequent $2 - itemsets$ can be obtained to explore the interaction among transactions along the time.

Step 5: Calculate the confidence of each frequent itemset, and generate the strong association rules whose confidences are equal to or larger than the minimum confidence. Integrate the strong association rules within ω time intervals. Namely, combine the strong association rules if the consequent of a rule at time t is the same as the antecedent of a rule at time $t + 1$. Then, the multi-item association rules can be obtained.

datasets; (2) Mine the association rules; and (3) Select the valuable association rules with the consideration of real traffic. Among them, components (1) and (2) are the most important ones, which are detailed in Algorithm 2.

In Algorithm 2, Step 4 illustrates the generation of frequent itemsets, which contains two major steps, namely the join and prune of the itemsets. The prune process is similar to that of Algorithm 1. The join step is detailed as follows:

- Join step of Inter-T-ST-Apriori algorithm: Let L_k denote the frequent $k - itemsets$ which are joined by itemsets in L_{k-1}. For the join step based on the frequent $(k - 1) - itemsets$ L_{k-1}, it is necessary to take into account the spatiotemporal constraints, as in Figure 2 and Equation 12. In addition, such attribute constraint as Equation 13 should also be considered. On the basis, the candidate $k - itemsets$ C_k can be generated.

IV. DATA DESCRIPTION

The data employed in this study is from Tianjin, China. The data is acquired by the surveillance camera at the intersections of most urban roads in Tianjin. The surveillance camera are widely distributed with the total number of more than 800, as shown in Figure 3. The surveillance camera take photos of the passing vehicles whose information can thus be recorded, including photographing time, the license plate number, and the driving direction. As well, the data contains the information of each surveillance camera, including the ID and the location (i.e., GPS coordinate).

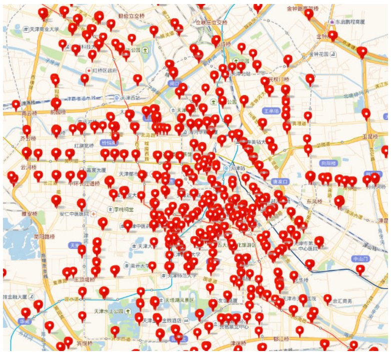


FIGURE 3. Distribution of surveillance camera in Tianjin, China.

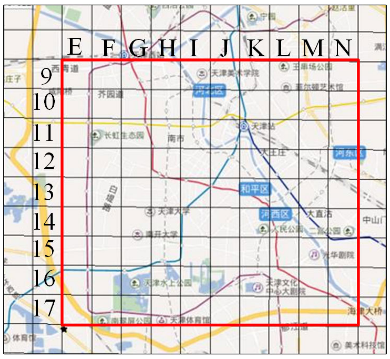


FIGURE 4. Urban grid.

This study concerns the traffic within the outer ring road of Tianjin. The data is collected from June 2, 2017 to June 30, 2017. Totally, more than 30 billion data samples are extracted.

A. REGIONAL GRID GENERATION

To mine the spatiotemporal associations of regional traffic congestion, the selected area is divided into square grids with size of $L_c \times L_c$ m^2. In practice, the length L_c is an essential parameter should be carefully addressed. Regarding to the fact that the average distance between two adjacent surveillance camera is approximate 300 m, L_c is set to be 1 km to ensure the sufficient road links are included in a grid.

Furthermore, case studies are conducted in the downtown areas of Tianjin with a dense distribution of surveillance camera, as is the surrounded area by the red solid line in Figure 4. Therefore, traffic states of each grid can be well recognized due to sufficient data.

For the following convenience, the grids are denoted as

$$GridID = G_xG_y,\ G_x \in [A, B, \ldots, N], \quad G_y \in [1, 2, \ldots, 17] \tag{14}$$

where G_x represents the abscissa value, and G_y represents the ordinate value. For the selected area, the abscissa ranges in $[E, F, \ldots, N]$, and the ordinate ranges in $[9, 10, \ldots, 17]$.

B. REGIONAL TRAFFIC VARIABLES

Based on the extracted information of passing vehicles, as well as the locations of surveillance camera, the basic variables, such as speed and flux, can be calculated, which will be detailed in this subsection.

1) SPEED

For a vehicle passing a road link with surveillance camera at both the beginning and the end, the travel time is the difference of photographing times of the two surveillance camera. The road length can be obtained with *API* of Baidu map based on the GPS coordinates of the surveillance camera. Then, the average speed of a vehicle passing through the road link can be calculated as follows:

$$v_k^i = \frac{l_k}{T_k^i} \tag{15}$$

where v_k^i represents the average speed of vehicle i passing road link k; l_k denotes the length of road link k; T_k^i denotes the travel time of vehicle i passing road link k.

Without loss of generality, set the time interval to be 5 minutes. The average speed of each road link can be calculated for each time interval as follows:

$$v_{t,k} = \frac{\sum_{i=1}^{n} v_k^i}{n} \tag{16}$$

where t denotes the order number of time interval, $t = 0, 1, \ldots, 287$; $v_{t,k}$ denotes the average speed of road link k during the time interval t; and n represents the amount of vehicles passing through road link k during the time interval t.

On the basis, the average speed of a grid can be calculated as,

$$V_{grid,t} = \frac{\sum_{k=1}^{N_{grid,l}} v_{t,k}}{N_{grid,l}} \tag{17}$$

where $V_{grid,t}$ denotes the average speed of grid *grid* during time interval t; and $N_{grid,l}$ is the amount of road links in grid *grid*.

2) FLUX

In this study, the flux of a road is regarded as the number of vehicles passing through the surveillance camera locating at the upstream intersection. Likewise, the flux is accounted for each 5 minutes, yields,

$$q_{t,m} = vehicle_{m,t} \tag{18}$$

where $vehicle_{m,t}$ denotes the number of vehicles passing surveillance camera m within time interval t. Accordingly, the regional flux can be calculated as,

$$Q_{grid,t} = \frac{\sum_{m=1}^{N_{grid,m}} q_{t,m}}{N_{grid,m}} \tag{19}$$

where $Q_{grid,t}$ denotes the average flux of grid *grid* during time interval t; and $N_{grid,m}$ denotes the number of surveillance camera within grid *grid*.

3) LENGTH OF TIME STEP

In general, an appropriate time interval should be determined to match the characteristics of congestion propagation. A too small time interval may result in unrealistic fluctuation of flux due to the impacts of signals; while a too large time interval may neglect some of the dynamic characteristics. In terms of existing traffic practice, the time interval is set to be 15 minutes, and 96 time intervals can be obtained. Based on the average speed and flux of grids within 5 minutes, these variables of 15 minutes can be easily obtained as,

$$V_{grid,t'} = \frac{\sum_{i=0}^{2} V_{grid,3t'+i}}{3} \tag{20}$$

$$Q_{\mathrm{grid},t'} = \sum_{i=0}^{2} Q_{\mathrm{grid},3t'+i} \tag{21}$$

where t' $(t' = 0, 1, \ldots, 95)$ represents the time interval within a day.

C. DATA PREPROCESSING

The original data is obtained by the surveillance camera to take photos of the passing vehicles. Then, the specific information of the vehicles such as the license plate can be extracted based on the image recognition technology. During the process, there are many factors that may lead to data missing or anomalies. To ensure the accuracy and reliability of the following case studies, it is necessary to pre-process the data.

1) MISSING DATA

The missing data can generally be preprocessed with various approaches, such as ignoring the data, filling the data by interpolation, regression, or prediction, and so on. In this study, the missing data are addressed as follows:

- If the data are continuously missed for more than 30 minutes, they are neglected.
- For the isolated missing data, it is filled by averaging the data of two adjacent time intervals.

2) OUTLIER VALUE DATA

In this study, the link travel time is regarded as the time difference for a vehicle passing two adjacent surveillance camera. In accordance with the extracted travel times, one can find some of them are extremely large which may be caused by various issues such as randomly roadside parking, and so on. The abnormal travel times may further result in unrealistic speeds. To this end, the travel time is filtered based on the following criterion:

$$0 < T_k^i < 1h \tag{22}$$

Furthermore, the speed is filtered by:

$$\mu_{speed} - 3\sigma_{speed} \leq v_{t,k} \leq \mu_{speed} + 3\sigma_{speed} \tag{23}$$

where, μ_{speed} and σ_{speed} denote the average and standard deviation of the speeds, respectively. Equation 22 is applied to eliminate the unrealistic link travel times. Equation 23 represents the classical $3\sigma_{speed}$ criterion, which is further used to remove the unrealistic speeds.

V. TRAFFIC STATE RECOGNITION

To investigate the spatiotemporal association rules of regional traffic congestion, it is necessary to develop approaches to recognize regional traffic states based on empirical traffic data.

So far, various criterions have been proposed to recognize traffic states, such as the criterion in HCM (textcolorred HCM 2000). Nevertheless, traffic states themselves are fuzzy, which cannot clearly classified. To this end, the fuzzy c-means (FCM) algorithm is extended to recognize traffic states.

A. CLUSTERING ALGORITHM FOR TRAFFIC STATE RECOGNITION

1) INTEGRATED CLUSTERING ALGORITHM

The FCM algorithm can be used to calculate the membership of each sample to all clusters. Nevertheless, the FCM algorithm has some obvious disadvantages which can be concluded as follows:

- The computational amount of FCM algorithm is very large, resulting in low efficiency.
- Since the initial clustering center is randomly specified, the objective function may converge to a local minimum value by gradient method.

To overcome the drawbacks of the FCM algorithm, it is extended by integrating the k-means++ algorithm. In detail, the clustering results of the k-means++ algorithm is used as the initial clustering centers of the FCM algorithm.

Both k-means++ and FCM are classical clustering algorithms which are not repeated in this study. One can refer [47] and [24] for details.

2) PARAMETER SETS OF THE CLUSTERING ALGORITHM

In general, several essential parameters should be specified for the FCM algorithm, such as the number of clusters, the fuzzy weight coefficient, termination condition of the algorithm, and so on.

a: OPTIMIZATION OF CLUSTERING NUMBER

So far, there is no confirmed conclusion about the classification of traffic states. In practice, traffic flow is generally divided into 3 levels, 4 levels, 5 levels, and so on.

In this study, sum of squared errors (SSE) is employed as the performance index of the clustering results. It is expressed as,

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{24}$$

where C_i denotes cluster i; p denotes the sample of C_i; and m_i is the center of C_i. In general, the SSE can be regarded as the clustering error. The smaller the SSE, the better the clustering results.

b: THE FUZZY WEIGHT COEFFICIENT m

The fuzzy weight coefficient m is an important parameter of FCM algorithm, which may seriously impact on the convergence. According to [24] and [48], the optimal value of m is 2 in most cases, which is applied in this study as well.

c: THE TERMINATION CONDITION

A stopping tolerance $\epsilon > 0$ should be predetermined for FCM algorithm. In this study, a sufficient small $\epsilon = 10^{-6}$ is specified to ensure the accuracy of the FCM algorithm.

3) PERFORMANCE INDEX OF CLUSTERING RESULTS

FCM algorithm is an unsupervised clustering algorithm. It attempts to obtain clustering results with high intra-class aggregation and low inter-class coupling. Here, the Xie-Beni index [49] is used as the performance index of the clustering. It is expressed as,

$$v_{XB}(U, V, X) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2 \|x_k - v_i\|^2}{n\left(\min_{i \neq j} \left\|v_i - v_j\right\|^2\right)} = \frac{\delta/n}{sep} \tag{25}$$

where X is the data sample set; V is the cluster centroid; U is a positive definite matrix; $sep = \min_{i \neq j} \left\|v_i - v_j\right\|^2$ denotes the degree of separation between clusters; $\delta = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2 \|x_k - v_i\|^2$ is total variation; n denotes the number of samples; and δ/n represents the intra-class compactness.

In terms of Equation 25, the smaller the $v_{XB}(U, V, X)$, the better the clustering results.

B. DATA PREPARATION

In this study, the grid traffic states are evaluated in terms of both the speed and flux, which have been extensively used in traffic flow theory to represent a specific traffic state.

The prepared data in Section IV is employed with each grid of $1km \times lkm$ and time interval of 15 minutes. In addition, the data within [22:00, 6:00] is omitted due to the extraordinary low traffic during the night, and traffic congestion may not appear. To this end, the time range is selected as [6:00, 22:00], with totally 64 time intervals during a day. The data of 29 days (from 2 June 2017 to 30 June 2017) is used in the following clustering, and the number of samples is 167040. Each data sample contains two parameters, i.e., speed and flux, which can be represented as $X_i = [x_{i1}, x_{i2}], i = 1, 2, \ldots, n$. Here, n denotes the number of samples.

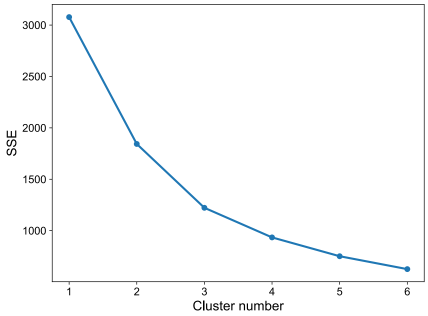


FIGURE 5. Variation of SSE with cluster number.

TABLE 2. Xie-beni coefficients of different clustering algorithms.

Performance index	Original FCM	Extended FCM
Xie-beni coefficient	0.49	0.23

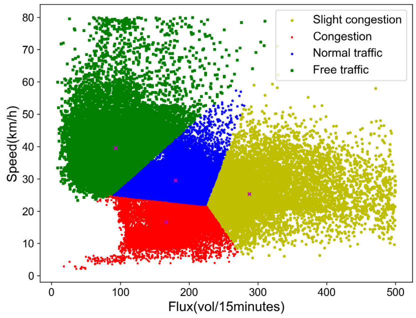


FIGURE 6. Traffic state classification based on the extended FCM.

The pre-processed data should be normalized to avoid the differences in magnitude. Here, the Z-score approach is applied, which is formulated as,

$$z = \frac{x - \mu}{\sigma} \tag{26}$$

where z is the normalized value; μ and σ are the mean and standard deviation of the sample data, respectively.

C. CLUSTERING RESULTS OF TRAFFIC STATE

Figure 5 shows variation of SSE with cluster number. One can see that SSE begins to vary slowly as the cluster number is larger than 4. Since the purpose is to obtain possible simple clustering results with sufficient compact samples for each cluster, the cluster number is set to be 4 in this study.

Table 2 shows the Xie-beni coefficients of the original FCM algorithm and the extended FCM algorithm. As can be seen, the extended FCM algorithm has smaller Xie-beni coefficient indicating the advantages of the extended FCM algorithm.

Figure 6 shows the clustering results with corresponding cluster centers in Table 3. In accordance with the speed and flux of cluster centers, the traffic states are named as free traffic, normal traffic, slight congestion and congestion, respectively.

TABLE 3. Clustering center for various traffic states.

Traffic state	Speed (km/h)	Flux (vol/15 minutes)
Free traffic	39.6	93
Normal traffic	29.53	191
Slight congestion	25.3	287
Congestion	16.64	167

TABLE 4. Several examples of the clustering results.

Sample	Speed (km/h)	Flux (vol / 15minutes)	Membership degree[†]	Traffic state
1	37.63	58	[0.88, 0.06, 0.02, 0.04]	Free
2	38.85	74	[0.96, 0.02, 0, 0.02]	Free
12	29	211	[0.08, 0.77, 0.04, 0.11]	Normal
19	17.17	202	[0.12, 0.03, 0.1, 0.75]	Congestion
302	20.95	419	[0.17, 0.18, 0.56, 0.09]	Slight congestion

[†] The membership is represented as [Free, Normal, Slight congestion, Congestion].

- Free traffic corresponds to the green points with high speed and low flux.
- Normal traffic corresponds to the blue points. In comparison with free traffic, the speed becomes lower and the flux is higher.
- Slight congestion corresponds to the yellow points. Traffic condition becomes worse, and stop-and-go traffic appears. Nevertheless, the flux is still high.
- Congestion corresponds to the red points with very low speed and flux. Traffic congestion is very serious.

Table 4 shows several examples of the clustering results. The traffic state of each sample is determined according to the membership. The degree of membership indicates the probability for the sample belonging to a cluster. Accordingly, each sample can be labeled by a traffic state with the largest degree of membership.

VI. CASE STUDY AND RESULTS

In this section, the IntraT-ST-Apriori and InterT-ST-Apriori algorithms will be applied on road network of downtown in Tianjin, China, to investigate the association and evolution characteristics of regional congestion.

Briefly, based on traffic data of a workday, strong association rules will be extracted based on the proposed algorithms. Furthermore, together with the map of Tianjin, we attempt to investigate the association and dynamic evolution of regional congestion in Tianjin.

A. DATA PREPARATION

As discussed in Section IV, traffic data of workday is divided into data samples with time interval of 15 minutes. Here, the night period is neglected because there are very few vehicles on roads, and the selected time period is 7:00-22:00,

TABLE 5. Spatiotemporal congestion transaction set.

TID	Period	Grid Index
1	1	[E,9], [F,10], ..., [N,9]
2	1	[E,9], [F,12], ..., [M,13]
...	...	...
30	2	[E,11], [E,12], ..., [N,13]
...	...	...

TABLE 6. Transaction bool matrix.

TID	[E, 9]	[E, 10]	...	[F, 9]	[F, 10]	...	[N, 9]	...	[N, 17]
1	1	1	...	0	1	...	1	...	1
2	0	1	...	0	1	...	1	...	0
...	...	...	...	...	...	...	...	...	...
n-1	1	0	...	1	0	...	1	...	0
n	1	1	...	0	0	...	0	...	1

TABLE 7. Frequent itemsets of congestion items for morning and evening rush hours, respectively.

Time period	2−intemset	3−intemset	4−intemset
Morning Peak	226	251	98
Evening Peak	117	46	9

in total 15 hours, containing 60 time intervals. The morning and evening rush hours are 7:00-9:00 and 16:00-19:00, respectively. In generally, traffic congestion is serious during rush hours.

With the method in Section V, the congested grids can be obtained for each time interval. Then, the spatiotemporal congestion transaction set can be derived. Examples are revealed in Table 5

By scanning the congested transaction set, the transaction Bool matrix of time interval i can be obtained, as shown in Table 6.

B. MINING REGIONAL CONGESTION ASSOCIATION RULES

Based on the prepared data, the Intra-ST-Apriori algorithm is applied to study the congestion association among the grids. Here, the minimum support is set to be 0.5, and the minimum confidence is 0.7.

Regarding to the morning and evening rush hours, the frequent 2-item set, frequent 3-item set and frequent 4-item set can be obtained, as shown in Table 7. Obviously, the size of item set for morning rush hours is larger than that of the evening rush hours, indicating more serious congestion for morning rush hours. In addition, the results demonstrate significant association of regional traffic congestion during morning rush hours due to the over saturated traffic flow. As a consequence, it should pay more attention on traffic control and management during morning rush hours.

TABLE 8. Strong association rules for morning rush hours.

No.	Time period	Rules	Confidence
1	7:00-7:15	$J12 \Rightarrow H13$	0.83
2	7:00-7:15	$J13 \Rightarrow H13$	1
...	...	...	...
1232	8:45-9:00	$K10 \Rightarrow (H12, I9, J13)$	0.7

TABLE 9. Strong association rules for evening rush hours.

No.	Time period	Rules	Confidence
1	16:00-16:15	$I13 \Rightarrow K11$	0.76
...	...	...	...
112	18:00-18:15	$(J11, H13) \Rightarrow J10$	0.86
...	...	...	...
379	18:45-19:00	$H12 \Rightarrow (J11, I9)$	0.82

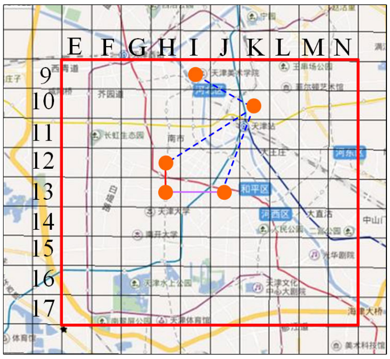


FIGURE 7. Congestion associations among different grids during the morning peak of workdays.

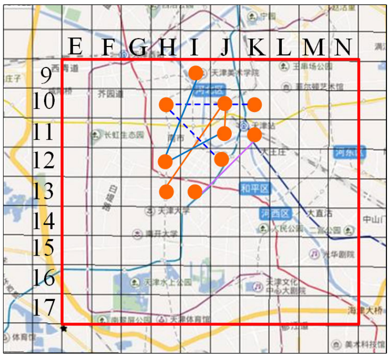


FIGURE 8. Congestion associations among different grids during the evening peak of workdays.

Furthermore, the strong association rules can be mined based the IntraT-ST-Apriori algorithm. Examples are shown in Tables 8 and 9 (as well in Figures 7 and 8) for morning and evening rush hours, respectively. Specifically, several rules are selected and analyzed to intuitively reveal the association characteristics of regional congestion.

- The strong association rule "$J12 \Rightarrow H13$" (i.e., "*Northeast of Yingkou Road Metro Station* $\Rightarrow$ *Tianjin University*") can be obtained during time interval

TABLE 10. Spatiotenporal association rules for morning rush hours.

No.	Time period	item 1	item 2	Confidence
1	7:00-7:30	$L14$	$K15$	1
2	7:15-7:45	$I13$	$H12$	0.71
3	7:15-7:45	$I14$	$H13$	0.79
...	...	...	...	...
16	8:30-9:00	$G12$	$H13$	0.66

of 7:00-7:15. The support is 0.5; the confidence is 0.83; the lift is 1.2; and *Kulc* is 0.575. The lift of the rule indicates a positive association. However, the rule is invalidated because *Kulc* is close to 0.5, which means that the antecedent and the consequent of the rule are irrelevant. Therefore, during 7:00-7:15, one cannot infer the traffic state of grid $H13$ based on that of grid $J12$. Vice versa.

- For the rule $(H13, J13) \Rightarrow H12$ during 7:30-7:45, the support, confidence, lift *Kulc* and *IR* are 0.5, 1, 1.3, 0.19, and 0.16, respectively. Accordingly, there exists equilibrium positive association between the antecedent and the consequent. It can thus be inferred that when the grids of both $H13$ and $J13$ are congested, grid $H12$ will also be congested. The probability is 50% for congestion appearing in all the three areas.
- During 18:00-18:15, there is a rule $(J11, H13) \Rightarrow J10$, i.e., "(*Italian Style Street*, *Tianjin University*) ⇒ *Lion Forest Street*". The support, confidence, lift *Kulc* and *IR* are 0.5, 0.86, 1.42, 0.77, and 0.44, respectively. Likewise, there exists equilibrium positive association between the antecedent and the consequent. If both the grids $J11$ and $H13$ are congested, grid $J10$ will be congested with the probability of 86%.

Based on the spatiotemporal association rules of regional congestion, once a region is congested, the traffic states of correlated regions can be predicted. This is valuable which can significantly support the collaborative control of congestion regions.

C. MINING REGIONAL CONGESTION PROPAGATIONS

The InterT-ST-Apriori algorithm is applied on urban road network of Tianjin to further explore the trajectory of regional congestion propagation. Here, the sliding time-window is set as 2, and the time interval λ is 1. The minimum support and minimum confidence are 0.5 and 0.6, respectively.

By applying the InterT-ST-Apriori algorithm, 16 strong association rules are obtained for morning rush hours, as shown in Table 10; and 13 strong association rules are obtained for evening rush hours, as shown in Table 11. Obviously, there are more strong association rules for morning rush hours, indicating that congestion propagation is more serious during morning rush hours.

TABLE 11. Spatiotenporal association rules for evening rush hours.

No.	Time period	item 1	item 2	Confidence
1	16:00-16:30	$H11$	$G12$	0.66
2	16:30-17:00	$I10$	$H10$	0.66
3	18:45-17:15	$I15$	$H14$	0.5
...	...	...	...	...
13	17:30-18:00	$J16$	$I17$	0.58

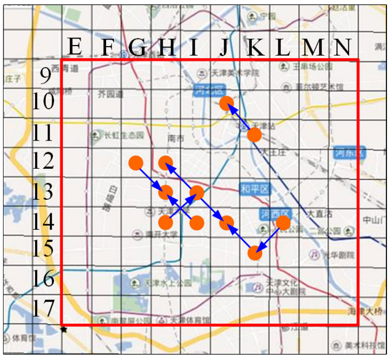


FIGURE 9. Congestion propagation trajectories during the morning peak of workdays.

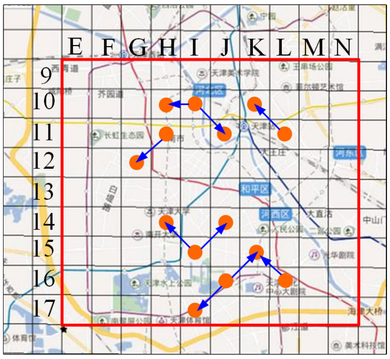


FIGURE 10. Congestion propagation trajectories during the evening peak of workdays.

Mapping the strong association rules to the map, the trajectories of congestion propagations can be obtained. Several typical examples are plotted in Figures 9 and 10 for morning and evening rush hours, respectively. Specifically, two typical rules are selected and explained in detail as follows:

- Rule $L14 \Rightarrow K15$ occurs during 7:00-7:30 with confidence of 1. This indicates that if congestion appears in grid $L14$ during 7:00-7:15, congestion would propagate to grid $K15$ during 7:15-7:30. To further validate the results, we checked the data in June, and found that the rule $L14 \Rightarrow K15$ existed for all the workdays of the month. In terms of the results, it is expected that traffic congestion of grid $K15$ can be alleviated once the congestion in grid $L14$ disappears.
- During 16:00-16:30, rule $H11 \Rightarrow G12$ is obtained with confidence of 0.71, which indicates that if congestion appears in grid $H11$ during 16:00-16:15, grid $G12$ may be congested during 16:15-16:30 with probability of 71%. It can thus be concluded that congestion in $G12$ is not entirely due to the internal factors, while it is also

affected by traffic state in grid $H11$. Therefore, it is suggested to solve traffic congestion problem in grid $H11$ in advance. Then, the congestion in grid $G12$ would be alleviated as well.

In brief, based on the proposed InterT-ST-Apriori algorithm, the spatiotemporal trajectories of regional congestion can be obtained. To this end, control strategies can be applied to the upstream grids to prevent the propagation of congestion.

VII. CONCLUSION

To explore the underlying inherent characteristics of regional congestion associations of a road network, this study develops a data-driven approach based on the Apriori algorithm.

To recognize regional traffic state based on real data, an alternative clustering algorithm is developed by integrating the classical k-means++ and FCM algorithms, which can significantly improve the efficiency and overcome the drawbacks of the FCM algorithm. By introducing both spatio and temporal attributes, the IntraT-ST-Apriori algorithm is developed to address the static features of regional traffic congestion. Furthermore, The InterT-ST-Apriori algorithm is developed to capture the dynamic characteristics of regional traffic congestion. To validate the proposed algorithms, case studies are carried out based on empirical data of Tianjin, China. Based on IntraT-ST-Apriori algorithm, the associated congestion regions during the same time interval can be obtained. Furthermore, the spatiotemporal trajectories of regional congestion propagation can be obtained by applying the InterT-ST-Apriori algorithm. To this end, the congested regions can be obtained in advance, which would be helpful for the regional traffic management and control, and significantly relieve traffic congestion.

REFERENCES

[1] M. J. Lighthill and G. B. Whitham, “On kinematic waves II. A theory of traffic flow on long crowded roads,” *Proc. Roy. Soc. London A, Math. Phys. Sci.*, vol. 229, no. 1178, pp. 317–345, 1955.

[2] P. I. Richards, “Shock waves on the highway,” *Oper. Res.*, vol. 4, no. 1, pp. 42–51, Feb. 1956.

[3] C. F. Daganzo, “The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory,” *Transp. Res. B, Methodol.*, vol. 28, no. 4, pp. 269–287, Aug. 1994.

[4] C. F. Daganzo, “The cell transmission model, part II: Network traffic,” *Transp. Res. B, Methodol.*, vol. 29, no. 2, pp. 79–93, Apr. 1995.

[5] K. Nagel and M. Schreckenberg, “A cellular automaton model for freeway traffic,” *J. Phys. I, France*, vol. 2, no. 12, pp. 2221–2229, Dec. 1992.

[6] O. Biham, A. A. Middleton, and D. Levine, “Self-organization and a dynamical transition in traffic-flow models,” *Phys. Rev. A*, vol. 46, no. 10, pp. R6124–R6127, Jul. 2002.

[7] J. Long, Z. Gao, H. Ren, and A. Lian, “Urban traffic congestion propagation and bottleneck identification,” *Sci. China F-Inf. Sci.*, vol. 51, no. 7, pp. 948–964, Jul. 2008.

[8] J. Long, Z. Gao, X. Zhao, A. Lian, and P. Orenstein, “Urban traffic jam simulation based on the cell transmission model,” *Netw. Spatial Econ.*, vol. 11, no. 1, pp. 43–64, Mar. 2011.

[9] B. J. Gajewski and L. R. Rilett, “Estimating link travel time correlation: An application of Bayesian smoothing splines,” *J. Transp. Statist.*, vol. 7, nos. 2–3, pp. 53–70, 2005.

[10] P. Rachtan, H. Huang, and S. Gao, “Spatio-temporal link speed correlations: An empirical study,” in *Proc. Transp. Res. Board 92th Annu. Meeting*, 2013.

[11] M. Wang, A. Ailamaki, and C. Faloutsos, “Capturing the spatio-temporal behavior of real traffic data,” *Perform. Eval.*, vol. 49, nos. 1–4, pp. 147–163, Sep. 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0166531602001086

[12] F. Zhang, X. Zhu, T. Hu, W. Guo, C. Chen, and L. Liu, “Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations,” *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 11, p. 201, Nov. 2016.

[13] Y. Wang, D. Zhang, Y. Liu, B. Dai, and L. H. Lee, “Enhancing transportation systems via deep learning: A survey,” *Transp. Res. C, Emerg. Technol.*, vol. 99, pp. 144–163, Feb. 2019, doi: 10.1016/j.trc.2018.12.004.

[14] H. Nguyen, W. Liu, and F. Chen, “Discovering congestion propagation patterns in spatio–temporal traffic data,” *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 169–180, Jun. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7511741/

[15] X. Deng, D. Zeng, and H. Shen, “Causation analysis model: Based on AHP and hybrid Apriori–Genetic algorithm,” *J. Intell. Fuzzy Syst.*, vol. 35, no. 1, pp. 767–778, Jul. 2018.

[16] X. Ma, H. Yu, Y. Wang, and Y. Wang, “Large–scale transportation network congestion evolution prediction using deep learning theory,” *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0119044.

[17] S. Goudarzi, M. Kama, M. Anisi, S. Soleymani, and F. Doctor, “Self–organizing traffic flow prediction with an optimized deep belief network for Internet of vehicles,” *Sensors*, vol. 18, no. 10, p. 3459, Oct. 2018.

[18] A. Ahmed, D. Ngoduy, and D. Watling, “Prediction of traveller information and route choice based on real-time estimated traffic state,” *Transportmetrica B, Transp. Dyn.*, vol. 4, no. 1, pp. 23–47, Jan. 2016, doi: 10.1080/21680566.2015.1052110.

[19] D. Chowdhury and A. Schadschneider, “Self-organization of traffic jams in cities: Effects of stochastic dynamics and signal periods,” *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 59, no. 2, pp. R1311–R1314, Jul. 2002.

[20] M. Li, Z.-J. Ding, R. Jiang, M.-B. Hu, and B.-H. Wang, “Traffic flow in a Manhattan-like urban system,” *J. Stat. Mech., Theory Exp.*, vol. 2011, no. 12, Dec. 2011, Art. no. P12001.

[21] J. Huang, M.-B. Hu, R. Jiang, and M. Li, “Effect of pre-signals in a Manhattan-like urban traffic network,” *Phys. A, Stat. Mech. Appl.*, vol. 503, pp. 71–85, Aug. 2018, doi: 10.1016/j.physa.2018.02.170.

[22] R. Jiang, J.-Y. Chen, Z.-J. Ding, D.-C. Ao, M.-B. Hu, Z.-Y. Gao, and B. Jia, “Network operation reliability in a Manhattan-like urban system with adaptive traffic lights,” *Transp. Res. C, Emerg. Technol.*, vol. 69, pp. 527–547, Aug. 2016, doi: 10.1016/j.trc.2016.01.006.

[23] M. Bermudez-Edo, P. Barnaghi, and K. Moessner, “Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation,” *Autom. Construct.*, vol. 88, pp. 87–100, Apr. 2018, doi: 10.1016/j.autcon.2017.12.036.

[24] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. Boston, MA, USA: Springer, 1981.

[25] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in *Proc. 20th Int. Conf. Very Large Data Bases*, 1994.

[26] A. A. Galib, M. Alam, N. Hossain, and R. M. Rahman, “Stock trading rule discovery based on temporal data mining,” in *Proc. 6th Int. Conf. Elect. Comput. Eng. (ICECE)*, Dec. 2010, pp. 566–569.

[27] P. Hu, M. Huang, P. Xu, W. Li, A. K. Usadi, and X. Zhu, “Finding nuggets in IP portfolios: Core patent mining through textual temporal analysis,” in *Proc. ACM Int. Conf.*, 2012.

[28] B. Liu, W. Hsu, Y. Ma, and B. Ma, “Integrating classification and association rule mining,” *Knowl. Discovery Data Mining*, vol. 98, pp. 80–86, Aug. 1998. [Online]. Available: http://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf%5Cn

[29] S. Qiao, N. Han, J. Wang, R.-H. Li, L. A. Gutierrez, and X. Wu, “Predicting long–term trajectories of connected vehicles via the prefix–projection technique,” *IEEE Trans. Intell. Transport. Syst.*, vol. 19, no. 7, pp. 2305–2315, Jul. 2018.

[30] J. Wang, Y. Mao, J. Li, Z. Xiong, and W.-X. Wang, “Predictability of road traffic and congestion in urban areas,” *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0121825.

[31] Z. Zhang, Q. He, H. Tong, J. Gou, and X. Li, “Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network,” *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 284–302, Oct. 2016, doi: 10.1016/j.trc.2016.08.006.

[32] H. Qi, M. Liu, L. Zhang, and D. Wang, “Tracing road network bottleneck by data driven approach,” *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0156089.

[33] K. Zhang, D. Sun, S. Shen, and Y. Zhu, "Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data," *J. Transp. Land Use*, vol. 10, no. 1, pp. 675–694, Aug. 2017.

[34] Z. Shan, Z. Pan, F. Li, H. Xu, and H. Xu, "Visual analytics of traffic congestion propagation path with large scale camera data," *Chin. J. Electron.*, vol. 27, no. 5, pp. 934–941, Sep. 2018.

[35] Z. Chen, Y. Yang, L. Huang, E. Wang, and D. Li, "Discovering urban traffic congestion propagation patterns with taxi trajectory data," *IEEE Access*, vol. 6, pp. 69481–69491, 2018.

[36] B. Anbaroglu, B. Heydecker, and T. Cheng, "Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks," *Transp. Res. C, Emerg. Technol.*, vol. 48, pp. 47–65, Nov. 2014, doi: 10.1016/j.trc.2014.08.002.

[37] M. Saeedmanesh and N. Geroliminis, "Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks," *Transp. Res. B, Methodol.*, vol. 105, pp. 193–211, Jan. 2017.

[38] Y. Wang, J. Cao, W. Li, T. Gu, and W. Shi, "Exploring traffic congestion correlation from multiple data sources," *Pervasive Mobile Comput.*, vol. 41, pp. 470–483, Oct. 2017, doi: 10.1016/j.pmcj.2017.03.015.

[39] S. An, H. Yang, and J. Wang, "Revealing recurrent urban congestion evolution patterns with taxi trajectories," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 4, p. 128, 2018.

[40] W. Wei, H. Zhou, K. Shimada, S. Mabu, and K. Hirasawa, "Comparative association rules mining using genetic network programming (GNP) with attributes accumulation mechanism and its application to traffic systems," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2008, pp. 292–298.

[41] D. Shen, L. Zhang, J. Cao, and S. Wang, "Forecasting citywide traffic congestion based on social media," *Wireless Pervasive Commun.*, vol. 103, no. 1, pp. 1037–1057, Nov. 2018, doi: 10.1007/s11277-018-5495-x.

[42] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transp. Res. B, Methodol.*, vol. 42, no. 9, pp. 771–781, Nov. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0191261508000799

[43] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transp. Res. B, Methodol.*, vol. 42, no. 9, pp. 759–770, Nov. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0191261508000180

[44] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, "Discovering spatiotemporal causal interactions in traffic data streams," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1010–1018.

[45] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On detection of emerging anomalous traffic patterns using GPS data," *Data Knowl. Eng.*, vol. 87, pp. 357–373, Sep. 2013, doi: 10.1016/j.datak.2013.05.002.

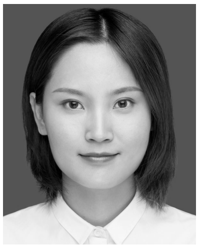
[46] Y. Guo, L. Yang, S. Hao, and J. Gao, "Dynamic identification of urban traffic congestion warning communities in heterogeneous networks," *Phys. A, Stat. Mech. Appl.*, vol. 522, pp. 98–111, May 2019, doi: 10.1016/j.physa.2019.01.139.

[47] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007.

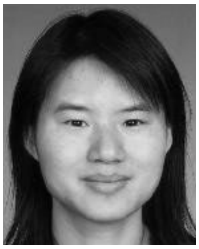
[48] N. Pal and J. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 370–379, Aug. 1995.

[49] X. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

DONG-FAN XIE received the B.S. degree in traffic engineering and the Ph.D. degree in traffic and transportation planning and management from Beijing Jiaotong University, Beijing, China, in 2005 and 2011, respectively. From May 2011 to May 2015, he worked as a Research Fellow with the School of Civil and Environmental Engineering, Nanyang Technological University, Singapore. From May 2012 to December 2016, he worked as a Lecturer with the School of Traffic and Transportation, Beijing Jiaotong University, where he is currently an Associate Professor. His current research interests include the traffic flow modeling, intelligent transportation systems, data-driven traffic system modeling, and network traffic dynamics.



MEI-HONG WANG received the B.S. degree in traffic and transportation from the Wuhan University of Science and Technology, Wuhan, China, in 2016, and the M.S. degree in traffic engineering from Beijing Jiaotong University, Beijing, China, in 2019. She is currently a Research Fellow with the Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, in July 2019. Her research interests include big data mining of traffic systems and network traffic flow dynamics.



XIAO-MEI ZHAO received the B.S. and M.S. degrees in control engineering from the Shaanxi University of Science and Technology of China, in 1997 and 2000, respectively, and the Ph.D. degree in control science from Zhejiang University, China, in 2003. From April 2003 to November 2014, she worked as a Lecturer and an Associate Professor with the School of Traffic and Transportation, Beijing Jiaotong University, where she has been a Professor, since 2014. Her research interests include traffic flow modeling, traffic control, and traffic dynamics.