


RESEARCH ARTICLE

Detection of clusters in traffic networks based on spatio-temporal flow modeling

Yan Shi^{1,2}  | Min Deng¹  | Jianya Gong² | Chang-Tien Lu³ |
Xuexi Yang¹ | Huimin Liu¹

¹Department of Geo-informatics, Central South University, Changsha, China

²State Key Laboratory of Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan, China

³Department of Computer Science, Virginia Tech, Falls Church, Virginia, USA

Correspondence

Xuexi Yang, Department of Geo-Informatics, Central South University, No. 932 South Lushan Road, Changsha, Hunan 410083, China
Email: studyang@sina.cn

Funding information

National Key Research and Development Foundation of China, Grant/Award Numbers: 2017YFB0503601 and 2017YFB0503700; China Postdoctoral Science Foundation, Grant/Award Number: 2017M610486; and National Natural Science Foundation of China, Grant/Award Numbers: 41471385, 41601424, 41730105 and 41771492

Abstract

Spatio-temporal clustering is a highly active research topic and a challenging issue in spatio-temporal data mining. Many spatio-temporal clustering methods have been designed for geo-referenced time series. Under some special circumstances, such as monitoring traffic flow on roads, existing methods cannot handle the temporally dynamic and spatially heterogeneous correlations among road segments when detecting clusters. Therefore, this article develops a spatio-temporal flow-based approach to detect clusters in traffic networks. First, a spatio-temporal flow process is modeled by combining network topology relations with real-time traffic status. On this basis, spatio-temporal neighborhoods are captured by considering traffic time-series similarity in spatio-temporal flows. Spatio-temporal clusters are further formed by successive connection of spatio-temporal neighbors. Experiments on traffic time series of central London's road network on both weekdays and weekends are performed to demonstrate the effectiveness and practicality of the proposed method.

1 | INTRODUCTION

With the rapid development of earth observation technology and the wide usage of sensors, enormous amounts of spatio-temporal data have been collected in recent years. How to adequately and rigorously discover latent and significant patterns and knowledge from massive spatio-temporal data has become a challenge (Han, Kamber, & Tung, 2001). Spatio-temporal clustering is an important technology of data mining and is primarily aimed at extracting a series of clusters from spatio-temporal data to ensure that objects in the same cluster are both similar to

each other and distinct from those in other clusters (Miller & Han, 2009; Shekhar, Vatsavai, & Celik, 2009). At present, spatio-temporal clustering has been widely applied in climate change detection (Birant & Kut, 2007; Deng, Liu, Wang, & Shi, 2013; Wu, Zurita-Milla, & Kraak, 2015; Wu, Zurita-Milla, Verdiguier, & Kraak, 2017), earthquake outbreak detection (Pei, Zhou, Zhu, Li, & Qin, 2010; Liu, Deng, Bi, & Yang, 2014), epidemic analysis (Delmelle, Dony, Casas, Jia, & Tang, 2014; Kulldorff, Heffernan, Hartman, Assunção, & Mostashari, 2005), crime hotspot detection (Nakaya & Yano, 2010; Shiode & Shiode, 2013), socio-economic analysis (Hagenauer & Helbich, 2013), and traffic analysis (Cheng & Anbaroglu, 2010; Feng, Wang, & Chen, 2014; Xie & Yan, 2013).

In terms of different applications, existing spatio-temporal clustering methods are mostly designed for five types of spatio-temporal data, namely spatio-temporal events, geo-referenced variables, geo-referenced time series, moving objects, and trajectories (Kisilevich, Mansmann, & Nanni, 2010). This study focuses on clustering geo-referenced time series, which record time series with respect to measured non-spatial attribute values at fixed locations represented by point/line/area entities. The clusters in geo-referenced time series should be formed

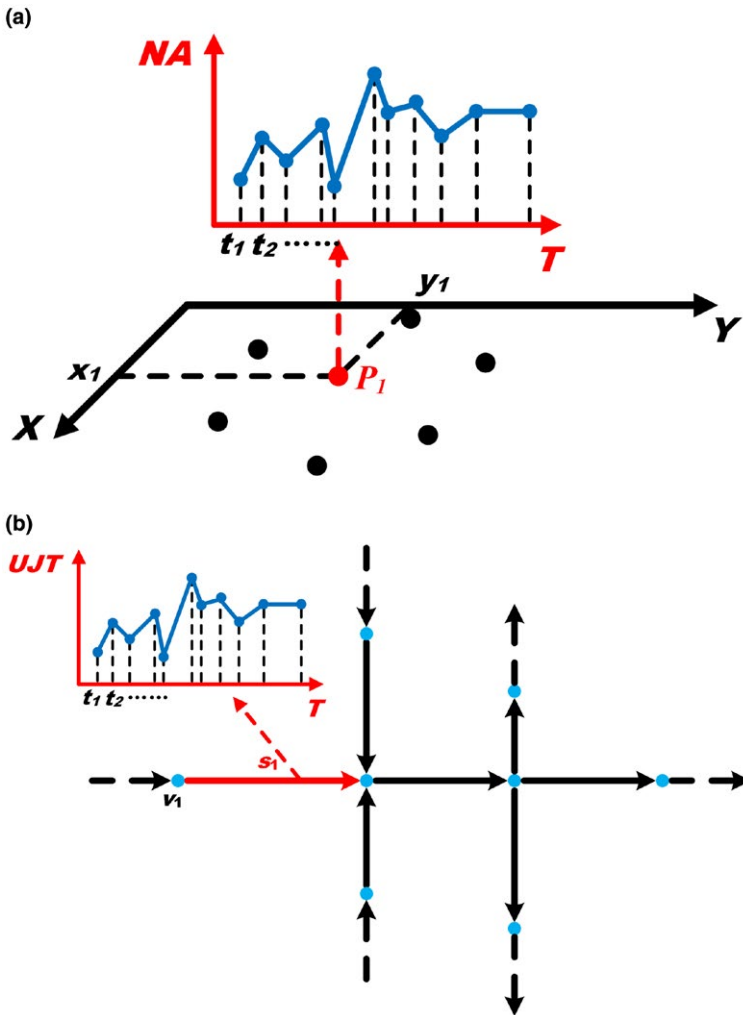


FIGURE 1 Examples of geo-referenced time series: (a) geo-referenced time series by point objects, where the coordinate pair (x_1, y_1) indicates the spatial location of P_1 and “NA” represents the non-spatial attribute; (b) traffic time series on a road network, where the arrows indicate the driving direction of vehicles, “ v_1 ” and “ s_1 ” denote a vertex and a segment of the road network, respectively, and “UJT” represents the unit journey time of vehicles (Shi et al., 2018)

by considering both spatio-temporal proximity and non-spatial attribute similarity. Figure 1a gives an example of geo-referenced time series by a point entity P_1 . Under some special circumstances in capturing geo-referenced time series, such as monitoring traffic flow on roads, non-spatial attribute values (e.g. average velocity of traffic flows) recorded by each road segment are derived from the flow of vehicles from upstream to downstream sensors (Mehboob et al., 2015; Ren, Zhang, Zhang, Wang, & Feng, 2018). Taking the traffic time series shown in Figure 1b as an example, vehicles keep driving on roads with directionality, so traffic status in the upstream can interact with that in the downstream. For example, in Figure 1b, if the vehicles on s_1 maintain adequate speed at t_1 , they can arrive at s_4 within t_1 . Conversely, in the case of traffic congestion at t_1 , these vehicles may remain on s_1 at t_2 . Thus, the correlations among road segments are temporally dynamic and spatially heterogeneous on account of various real-time traffic status and different segment lengths, suggesting that the spatio-temporal neighborhood range of each road segment will vary with changing traffic conditions (Kang, Shekhar, Wennen, & Novak, 2008; Min, Hu, & Zhang, 2010; Cheng, Wang, Haworth, Heydecker, & Chow, 2014). Specifically, the range will be larger in smooth traffic conditions and smaller in congested traffic conditions. However, existing clustering methods designed for geo-referenced time series mostly construct fixed spatio-temporal neighborhoods by defining a spatial coverage and time window. In light of this situation, this study aims to handle the temporally dynamic and spatially heterogeneous correlations in traffic time series based on spatio-temporal flow modeling for accurate and effective detection of clusters. The major contributions are as follows:

- Modeling spatio-temporal flows to construct temporally dynamic and spatially heterogeneous neighborhoods.
- Performing clustering by integrating spatio-temporal flows to accurately present the traffic running rules.
- Visualizing spatio-temporal clustering results meticulously to facilitate further analysis.

The remainder of this article is organized as follows. Section 2 reviews related work on geo-referenced time-series clustering and presents the proposed strategy of cluster detection. In Section 3, the proposed method is fully elaborated. In Section 4, extensive experiments on real-life data are performed and analyzed to demonstrate the effectiveness and practicability of the proposed method. Section 5 summarizes the most interesting findings and presents directions for future research.

2 | RELATED WORK

In this section, a systematic review will first be presented on the methods of cluster detection in geo-referenced time series and in traffic time series on road networks. On this basis, the performance of existing methods on traffic time series will be critically analyzed.

2.1 | Detection of clusters in geo-referenced time series

Existing cluster detection methods specifically designed for geo-referenced time series can be roughly divided into partition-based, density-based, and model-based clustering. Partition-based methods aim to directly aggregate objects based on their similarities with respect to observed non-spatial attribute values at different time stamps. For example, Zhang, Huang, Shekhar, and Kumar (2003) considered correlation analysis of geo-referenced time series and proposed a filter-and-refine approach to find pairs of potentially interacting spatial locations. Wu et al. (2015) employed the Bregman block average co-clustering algorithm with I-divergence to reveal cluster patterns from spatial and temporal dimensions. On this basis, Wu et al. (2017) developed a Bregman cuboid average triclustering algorithm with I-divergence to extract triclusters from spatial, temporal, and any third dimensions. Furthermore, the K-means algorithm was adopted to partition these triclusters into spatio-temporal cluster patterns.

Density-based methods extract spatio-temporal clusters by measuring the local density of each object. For example, the ST-DBSCAN algorithm proposed by Birant and Kut (2007) simultaneously considered spatial distance, temporal adjacency, and non-spatial attribute similarity when defining spatio-temporal density. On this basis, the notions of density-reachable and density-connected are proposed to define density-based clusters. Deng et al. (2013) constructed spatio-temporal neighborhoods using constrained Delaunay triangulation and spatio-temporal autocorrelation. A density-based clustering process was further performed to discover spatio-temporal clusters by considering non-attribute attribute similarity between spatio-temporal adjacent objects.

Model-based methods perform spatio-temporal clustering by building models based on statistics or machine learning theory. For example, Kulldorff et al. (2005) employed spatio-temporal scan statistic methods to detect outbreaks of epidemics in space and time. In recent years, some machine learning methods—such as the self-organizing feature map (SOM), which is a kind of artificial neural network—have been adopted to efficiently analyze spatio-temporal aggregation patterns from geo-referenced time series (Feng et al., 2014; Hagenauer & Helbich, 2013).

2.2 | Detection of clusters in traffic time series on road networks

In addition to the above discussed geo-referenced time-series clustering research, specific studies have been performed on traffic time series, which can be captured by video surveillance, on road networks (Mehboob et al., 2015; Ren et al., 2018). For example, Chen, Zhang, Hu, and Yao (2006) modeled the traffic time series in all links as a matrix $m \times n$, where m and n represented the number of time intervals and links, respectively, and implemented a SOM-based partitioning to reveal temporal distribution patterns of traffic flows at a certain spatial region. Ntoutsis, Mitsou, and Marketos (2008) proposed a hierarchical strategy to cluster sensors by means of shape-based distance between traffic time series, structure-based distance between sensors, and value-based distance between traffic time series successively. Hu, Luo, Yan, and Shi (2011) measured spatial neighborhood relationships between road segments by a “shortest path” analysis and employed dynamic time warping to measure the similarity between pairs of traffic time series. The two similarity measurements were then combined to partition the road network based on fuzzy clustering. Zhou, Lin, and Xi (2013) introduced an agglomerative hierarchical clustering method to partition the entire road network by integrating the length of road segments, the number of lanes, the traffic time series, and the queue length to measure the similarity between adjacent road segments. In addition, Anbaroglu, Heydecker, and Cheng (2014) aimed at clustering non-recurrent congestion events, which were the percentile-based or space-time scan statistics-based extraction of episodes respecting long link journey time with high confidence that occurred on spatially adjacent links and at the same time interval.

2.3 | A critical analysis of existing methods

Based on the systematic review of related work, the performance of existing spatio-temporal clustering methods designed for geo-referenced time series can be summarized as follows.

They do not consider the varying interactions of distinct spatial entities at different time stamps when constructing spatio-temporal neighborhoods. However, in traffic time series on road networks, there are directionality and dynamic characteristics of traffic flows. Specifically, taking the simulated data in Figure 1b as an example, the driving directions determine the direction relationship between adjacent road segments (e.g. $s_1 \rightarrow s_4$). The average velocity of vehicles on a road segment can directly reflect the real-time traffic status. The spatio-temporal influence extents (i.e. spatio-temporal neighborhoods) of each road segment will change continuously on account of various real-time traffic conditions (i.e. larger in smooth traffic conditions and smaller in congested traffic conditions) (Cheng et al., 2014; Min et al., 2010). Existing traffic time-series clustering methods were primarily designed to partition a road network into a set of subregions, instead of revealing clusters regarding traffic flow variables in both space and time. Cheng and Anbaroglu (2010) could discover clusters by measuring traffic flow

variable similarity between spatio-temporal objects, but could not construct temporally dynamic spatio-temporal neighborhoods in the clustering process.

As a matter of fact, without the consideration of vehicles flowing from upstream to downstream, it is difficult for existing spatio-temporal clustering methods to accurately unveil actual traffic mobility patterns. In our previous work, the directionality and dynamic nature of traffic flows have been considered in spatio-temporal anomaly detection (Shi, Deng, Yang, & Gong, 2018). Regarding traffic flow similarities, the aim of anomaly detection is to find those highly dissimilar spatio-temporal objects in the data. Using the spatio-temporal flow modeling proposed in this work, this study develops a clustering approach for traffic time series in order to group objects with high similarities. Through clustering analysis, this study further focuses on discovering and revealing city traffic flow regularities in both spatial and temporal dimensions from a global perspective.

3 | THE SPATIO-TEMPORAL FLOW-BASED CLUSTER DETECTION METHOD

This study constructs a framework to detect spatio-temporal flow-based clusters. The framework contains two parts, including spatio-temporal flow modeling and spatio-temporal cluster detection, as illustrated in Figure 2.

Spatio-temporal flow modeling. Existing clustering methods for geo-referenced time series mostly determined a fixed spatio-temporal coverage for each spatio-temporal object by defining a cylinder with a spatial radius and time window (Birant & Kut, 2007; Deng et al., 2013). With respect to traffic time series on road networks, as explained in Section 2.3, the spatio-temporal neighborhoods of each spatio-temporal object can vary with the continuous change in traffic status. In this case, the topological relationships between road segments are first

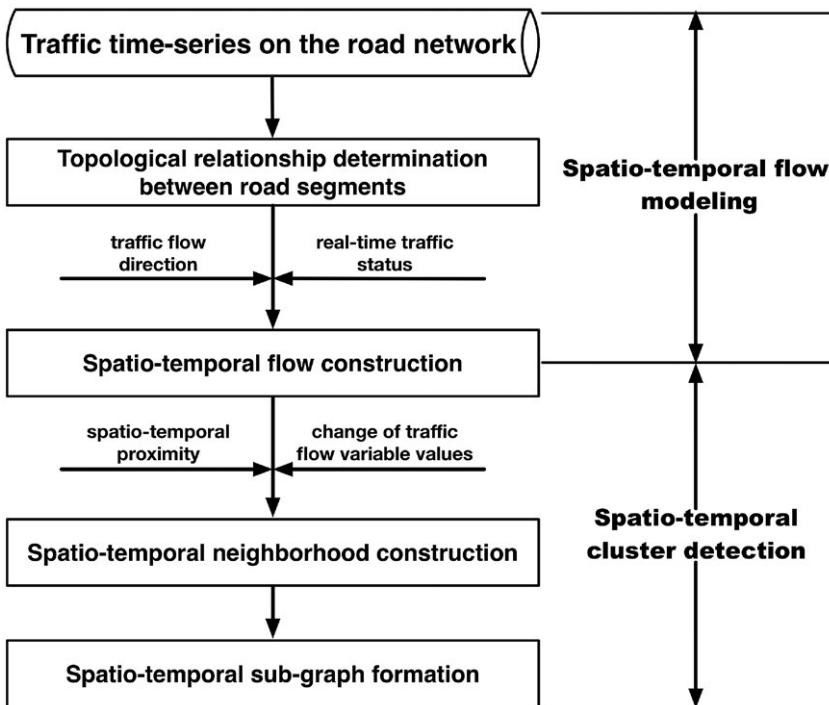


FIGURE 2 The framework of spatio-temporal flow-based cluster detection in traffic time series on a road network

determined. On this basis, spatio-temporal flows are further constructed by combining the traffic flow direction and the real-time traffic status.

Spatio-temporal cluster detection. Clustering aims to group highly similar objects into the same cluster, while dissimilar objects belong to different clusters. Spatio-temporal clustering for geo-referenced time series should ensure that objects in the same cluster are similar in spatial, temporal and non-spatial attribute dimensions (Shekhar et al., 2009). In this case, both spatio-temporal proximity and traffic flow variable similarity should be considered when constructing spatio-temporal neighborhoods. Likewise, the rates of consecutive change in spatio-temporal flow (i.e. the slope of change proposed in Cheng & Anbaroglu, 2010) should also be small to homogenize neighborhoods. In addition, the accumulation of gradual changes in the flow will make neighborhoods heterogeneous, so the "spatio-temporal connected" proposal in Deng et al. (2013) is employed to eliminate the inconsistency compared with the connected upstream flow. Considering the above issues, the spatio-temporal neighborhoods are constructed for all spatio-temporal objects and further form subgraphs to indicate spatio-temporal clusters.

With the full consideration of the above issues, the proposed spatio-temporal flow-based clustering method can be described in detail by taking traffic time series as the example. Section 3.1 introduces basic definitions regarding traffic time series on road networks. Sections 3.2 and 3.3 address the process of spatio-temporal flow modeling and spatio-temporal cluster detection in succession. The implementation of the proposed method is described in Section 3.4.

3.1 | Traffic time series on road networks

This study focuses on investigating the macroscopic spatio-temporal characteristics of traffic flows on road networks, instead of the individual behavior of each vehicle that can be captured by GPS trajectories. Based on this precondition, traffic time series, recording the measured traffic flow variables (i.e. unit journey time and average velocity) at a certain time resolution, on a road network, will be selected as the analyzed object. Using the simulated dataset in Figure 3 as an example, some basic definitions are introduced.

Definition 1: Vertices and road segments. On a road network, the intersections between two roads constitute a series of vertices, denoted v_i (e.g. v_1, v_2, \dots, v_8) in Figure 3a. These nodes partition a network into a series of road segments s_i (e.g. s_1, s_2, \dots, s_7) in Figure 3a. Each segment with a certain length is oriented to reflect the driving direction of vehicles and can be represented by a ternary array $s_i = (v_{start_i}, v_{end_i}, len_i)$. This means that vehicles on s_i with length len_i flow from the start vertex v_{start_i} to the end vertex v_{end_i} , denoted $v_{start_i} \rightarrow v_{end_i}$. In Figure 3a, $s_1 = (v_1, v_2, 1000 \text{ m})$, $s_2 = (v_4, v_2, 800 \text{ m})$, ..., $s_7 = (v_5, v_8, 550 \text{ m})$.

Definition 2: Spatio-temporal cells. For each road segment s_i , the recorded traffic time series regarding traffic flow variable tfv can be decomposed into discrete spatio-temporal cells at each time interval t_j , denoted $stc_{s_i, t_j} = (s_i, t_j, tfv_{s_i, t_j})$. Figures 3b and c simulate the journey time and average velocity of vehicles on each road segment in Figure 3a with time interval 300 s. For example, the spatio-temporal cell $(s_1, t_1, 200 \text{ s})$ signifies that vehicles require 200 s to pass through s_1 during t_1 . Correspondingly, $(s_1, t_1, 5.0 \text{ m/s})$ signifies that the average driving velocity of vehicles on s_1 is 5.0 m/s during t_1 .

3.2 | Modeling of spatio-temporal flows

To address the directionality and dynamic nature of traffic flows, this section describes a process of spatio-temporal flow modeling. First, the directionality of each segment and the topological connectivity among different segments together form the matrix of spatial neighbors.

Definition 3: Spatial neighbors. Given a road network, if two road segments s_i and s_j satisfy the condition $v_{end_i} = v_{start_j}$, then s_j is defined as the first-order spatial neighbor of s_i , denoted $SN^1(s_i) = \{s_j | v_{end_i} = v_{start_j}\}$. Further, $SN^1(s_i)$ and the first-order spatial neighbors of $SN^1(s_i)$ constitute the second-order spatial neighbors of s_i , denoted $SN^2(s_i) = SN^1(s_i) \cup \{s_k | s_k \in SN^1(s_j) \text{ and } s_j \in SN^1(s_i)\}$. In Figure 3a, $SN^1(s_1) = \{s_4\}$ and $SN^2(s_1) = \{s_4, s_5, s_6, s_7\}$. The n th-order

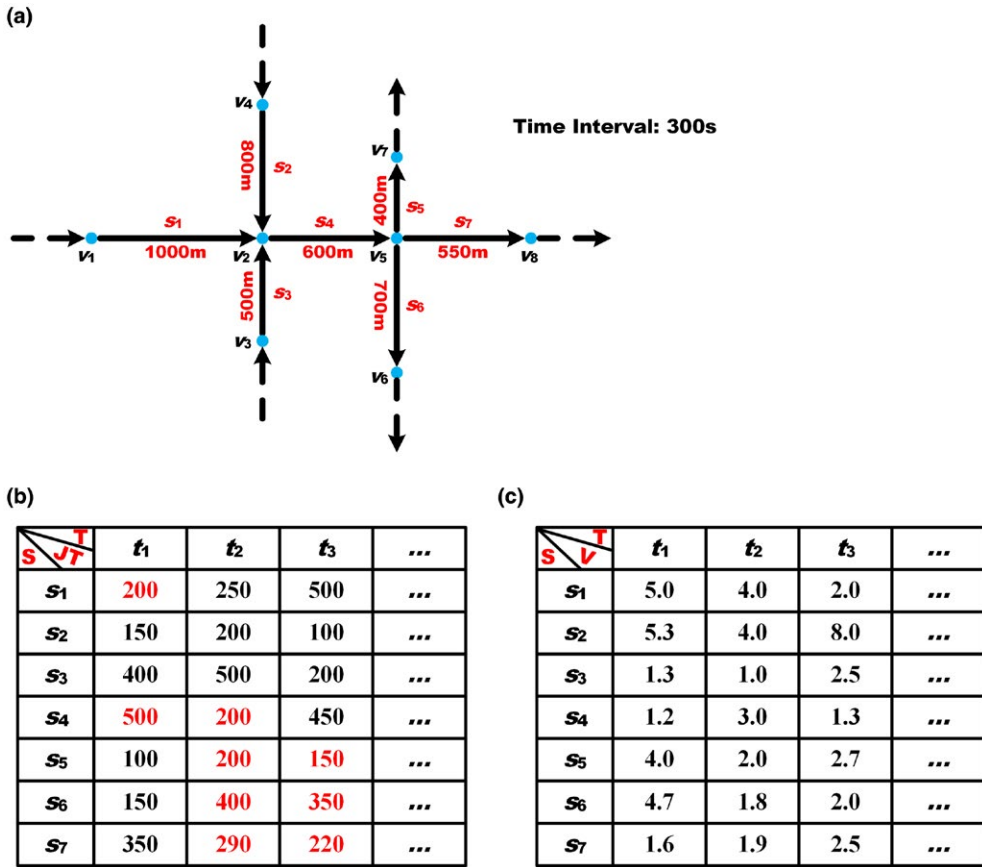


FIGURE 3 A group of simulated traffic time series on a road network: (a) the topological structure of the road network; (b) the time series about the journey time JT (s) of vehicles on each segment; and (c) the time series about the average driving velocity V (m/s) of vehicles on each segment

spatial neighbors can be obtained in this manner. A variable w_{ij} can be defined to indicate whether s_j belongs to the spatial neighbor of s_i . Specifically, if s_j is in $SN^n(s_i)$, then w_{ij} is set as 1; else, w_{ij} is set as 0. Then, the spatial neighbor matrix of this road network can be built. Figures 4a and b illustrate the first- and second-order spatial neighbor matrix for all road segments in Figure 3a, respectively.

Definition 4: Spatio-temporal flows. Starting from any spatio-temporal cell $stc_{s_i-t_j} = (s_i, t_j, ujt_{s_i-t_j})$, where $ujt_{s_i-t_j}$ represents the average unit journey time of vehicles passing through s_i at t_j , a process of spatio-temporal flow can be modeled based on the recorded real-time traffic status of spatio-temporal cells. Specifically, assuming that it costs vehicles the journey time of JT to arrive at $stc_{s_m-t_n} = (s_m, t_n, ujt_{s_m-t_n})$ from $stc_{s_i-t_j}$, Dis_{s_m} is the distance that has been passed through on s_m and T is the time interval of the traffic time-series, the updated Dis'_{s_m} and JT' can be calculated as:

$$Dis'_{s_m} = \begin{cases} Dis_{s_m} + [(k+1) * T - JT] / ujt_{s_m-t_{n+1}}, & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m-t_{n+1}} \geq (k+1) * T - JT \\ 0, & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m-t_{n+1}} < (k+1) * T - JT \end{cases} \quad (1)$$

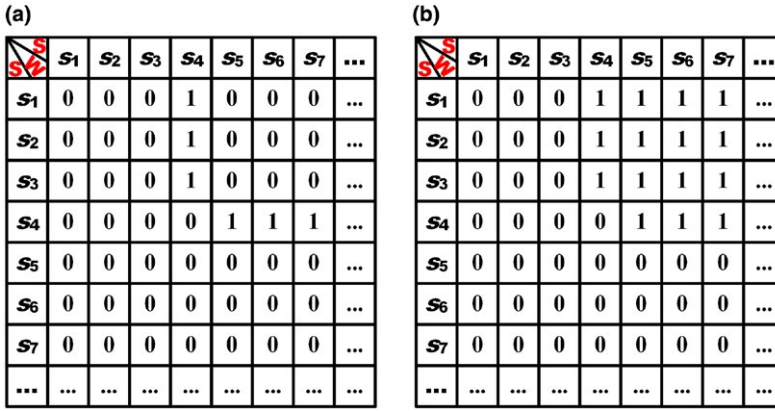


FIGURE 4 The spatial neighbor matrix of the road network, where S denotes road segments and W indicates whether two road segments are spatial neighbors or not: (a) the 1st-order spatial neighbor matrix; and (b) the 2nd-order spatial neighbor matrix

$$JT = \begin{cases} (k+1) * T, & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m, t_{n+1}} \geq (k+1) * T - JT \\ k * T + (len_m - Dis_{s_m}) * ujt_{s_m, t_{n+1}}, & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m, t_{n+1}} < (k+1) * T - JT \end{cases} \quad (2)$$

where k is the number of time intervals that have elapsed. Then, the next arrived spatio-temporal cell after stc_{s_m, t_n} can be determined as:

$$Arrival(stc_{s_m, t_n}) = \begin{cases} stc_{s_m, t_{n+1}}, & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m, t_{n+1}} \geq (k+1) * T - JT \\ stc_{SN}^1(s_m, t_n), & \text{if } (len_m - Dis_{s_m}) * ujt_{s_m, t_{n+1}} < (k+1) * T - JT \end{cases} \quad (3)$$

Intuitively, focusing on the simulated dataset in Figure 3, if setting stc_{s_1, t_1} as the starting cell, as shown in Figure 5a, the vehicles will arrive at stc_{s_4, t_1} after 200 s and 100 s remains in t_1 . After traveling 120 m (1.2 m/s * 100 s), the vehicles will be located at stc_{s_4, t_2} with 480 m remaining in s_4 . Using this analogy, Figures 5b and c show the flowing of vehicles in t_2 and t_3 , respectively, and the spatio-temporal flow starting from stc_{s_1, t_1} , denoted $STF(stc_{s_1, t_1})$, can be modeled as:

$$STF(stc_{s_1, t_1}) = stc_{s_1, t_1} \xrightarrow[200s]{1000m} stc_{s_4, t_1} \xrightarrow[100s]{120m} stc_{s_4, t_2} \xrightarrow[160s]{480m} \begin{cases} stc_{s_5, t_2} \xrightarrow[140s]{280m} stc_{s_5, t_3} \xrightarrow[44.4s]{120m} \dots \\ stc_{s_6, t_2} \xrightarrow[140s]{252m} stc_{s_6, t_3} \xrightarrow[224s]{448m} \dots \\ stc_{s_7, t_2} \xrightarrow[140s]{266m} stc_{s_7, t_3} \xrightarrow[113.6s]{284m} \dots \end{cases} \quad (4)$$

Figure 5 describes the moving process of vehicles and the modeled spatio-temporal flow, respectively, starting from stc_{s_1, t_1} .

3.3 | Detection of spatio-temporal clusters

Based on the modeled spatio-temporal flows, two steps are performed to detect spatio-temporal clusters, including spatio-temporal neighborhood construction and spatio-temporal clustering. The two steps will be elaborated in the following subsections.

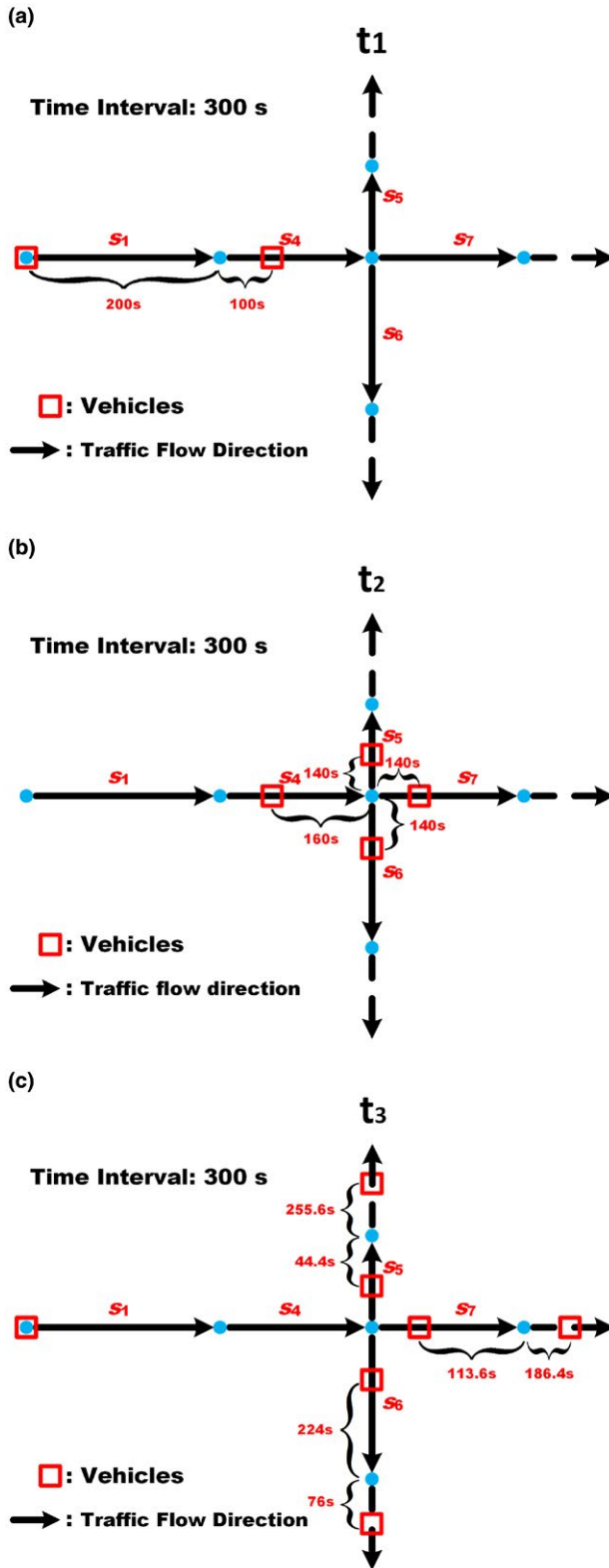


FIGURE 5 The modeling of spatio-temporal flow starting from stc_{s_1,t_1} : (a) the flow of vehicles in t_1 ; (b) the flow of vehicles in t_2 ; and (c) the flow of vehicles in t_3

3.3.1 | Spatio-temporal neighborhood construction

For each spatio-temporal cell, its spatio-temporal neighborhoods should be the members of spatio-temporal flow starting from this cell and a consecutive spatio-temporal chain must be formed by considering non-spatial attribute similarity. The non-spatial attribute similarity can be measured using the differences of attribute values between spatio-temporal adjacent cells, which may be roughly categorized into absolute and relative differences (Cheng & Anbaroglu, 2010). In the process of clustering, gradual changes regarding attribute values should be taken into account as well (Deng et al., 2013). Taking Figure 6 as an example, if the thresholds for absolute and relative differences are respectively set at 5 and 3, specific cases in spatio-temporal neighborhood construction can be described as:

- (i) **Sudden changes regarding absolute differences.** In Figure 6a, the absolute difference between G and H is 18 and significantly larger than the absolute difference threshold. Hence, the connectedness between G and H (i.e. $G \rightarrow H$) should be broken.
- (ii) **Sudden changes regarding relative differences.** In Figure 6b, the relative difference of $G \rightarrow H$ equals 4, which is smaller than the absolute difference threshold. However, considering the upstream object of G (i.e. F), the relative difference between G and H can be calculated as $(G \rightarrow H)/(F \rightarrow G) = 4$. This value is larger than the relative difference threshold, so the object H cannot be connected to G.
- (iii) **Gradual changes in the flow.** In Figure 6c, the absolute difference between any two adjacent objects is equivalent to 1. For $J \rightarrow K$, the difference between K and the average value of $A \rightarrow \dots \rightarrow J$ is 5.5, which is larger than the absolute difference threshold. In other words, the gradual changes lead the object K to be separated from the flow $A \rightarrow \dots \rightarrow J$.

Motivated by the above issues, the notions of spatio-temporal reachable and spatio-temporal connected will first be introduced.

Definition 5: Spatio-temporal reachability. Given any cell stc_q in $STF(stc_{s_i}, t_i)$, the absolute and relative differences between stc_q and its upstream adjacent cell stc_p regarding traffic flow variable values can be respectively expressed as:

$$A_Diff(stc_p, stc_q) = \begin{cases} |ujt_{stc_q} - ujt_{stc_p}|, & \text{if } |ujt_{stc_q} - ujt_{stc_p}| > \varepsilon 1 \\ \varepsilon 1, & \text{if } |ujt_{stc_q} - ujt_{stc_p}| \leq \varepsilon 1 \end{cases}, \text{ where } stc_q \neq stc_{s_i}, t_i \quad (5)$$

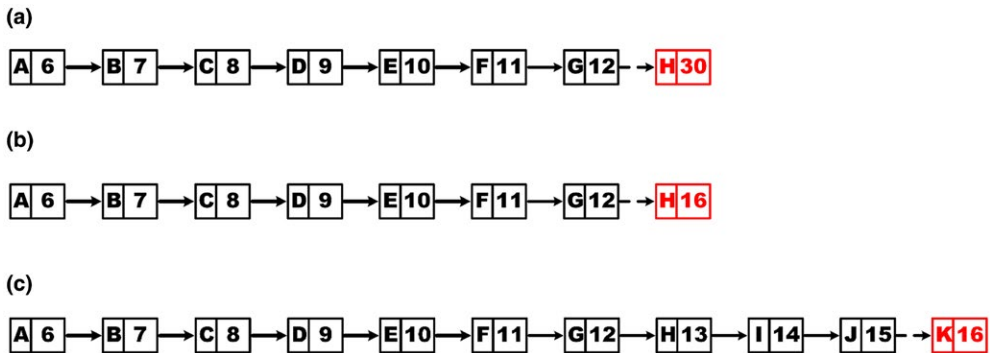


FIGURE 6 Three cases in the process of spatio-temporal neighborhood construction: (a) sudden changes regarding absolute differences; (b) sudden changes regarding relative differences; and (c) gradual changes in the flow

$$R_Diff(stc_p, stc_q) = \begin{cases} \frac{|ujt_{stc_q} - ujt_{stc_p}|}{A_Diff_{stc_q \rightarrow stc_p}}, & \text{if } stc_p \neq stc_{s_i, t_j} \\ \frac{|ujt_{stc_q} - ujt_{stc_p}|}{\epsilon 1}, & \text{if } stc_p = stc_{s_i, t_j} \end{cases} \quad (6)$$

where stc_o is the upstream adjacent cell of stc_p ; $\epsilon 1$ is a threshold that can determine whether the absolute difference between stc_q and stc_p is small enough to be ignored. Further, if $A_Diff(stc_p, stc_q) \leq \epsilon 2$ and $R_Diff(stc_p, stc_q) \leq \epsilon 3$, where $\epsilon 2$ and $\epsilon 3$ are another two given thresholds, then it is defined that there exists spatio-temporal reachability from stc_p to stc_q .

Definition 6: Spatio-temporal connectivity. Let $STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)$ represent the spatio-temporal flow from stc_{s_i, t_j} to stc_p . The absolute and relative differences between stc_q and $STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)$ can be calculated respectively as:

$$A_Diff[STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p), stc_q] = \begin{cases} |ujt_{stc_q} - ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}|, & \text{if } |ujt_{stc_q} - ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}| > \epsilon 1 \\ \epsilon 1, & \text{if } |ujt_{stc_q} - ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}| \leq \epsilon 1 \end{cases} \quad (7)$$

where $stc_q \neq stc_{s_i, t_j}$

$$R_Diff[STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p), stc_q] = \begin{cases} \frac{|ujt_{stc_q} - ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}|}{A_Diff[STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p), stc_q]}, & \text{if } stc_q \neq stc_{s_i, t_j} \\ \frac{|ujt_{stc_q} - ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}|}{\epsilon 1}, & \text{if } stc_q = stc_{s_i, t_j} \end{cases} \quad (8)$$

where $ujt_{STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)}$ represents the average unit journey time of cells in $STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p)$. If $A_Diff[STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p), stc_q] \leq \epsilon 2$ and $R_Diff[STF(stc_{s_i, t_j} \rightarrow \dots \rightarrow stc_p), stc_q] \leq \epsilon 3$, then the cell stc_q is defined to be spatio-temporal connected to stc_{s_i, t_j} .

Definition 7: Spatio-temporal neighborhoods. For any cell in $STF(stc_{s_i, t_j})$, only if both its upstream adjacent cell is spatio-temporal reachable from it and it is spatio-temporal connected to stc_{s_i, t_j} will this spatio-temporal flow continue; else, the flow will stop and the cell is considered a broken point. In $STF(stc_{s_i, t_j})$, all upstream cells of this broken point will constitute the spatio-temporal neighborhoods of stc_{s_i, t_j} , denoted $STN(stc_{s_i, t_j})$.

3.3.2 | Spatio-temporal clusters detection

Based on the constructed spatio-temporal neighborhoods, spatio-temporal clusters can further be extracted by constructing a graph.

Definition 8: Spatio-temporal clusters. In the spatio-temporal neighborhoods of any cell, all members connected to their adjacent ones constitute a chain and all chains can further form a graph jointly. In this graph, each connected subgraph will constitute a spatio-temporal cluster. Taking Figure 7 as an example, the constructed spatio-temporal neighborhood relationships among distinct cells are represented by arrows. On this basis, a series of separated subgraphs can be formed by the spatio-temporal neighborhoods

(i.e. $stc_{s_1, t_2} \rightarrow stc_{s_4, t_2} \rightarrow \begin{cases} stc_{s_5, t_2} \rightarrow stc_{s_5, t_3} \rightarrow \dots \\ stc_{s_7, t_2} \rightarrow stc_{s_7, t_3} \rightarrow \dots \end{cases}$ and $stc_{s_6, t_2} \rightarrow stc_{s_6, t_3}$). Each subgraph is finally defined as a spatio-temporal cluster STC_i , as indicated in Figure 7.

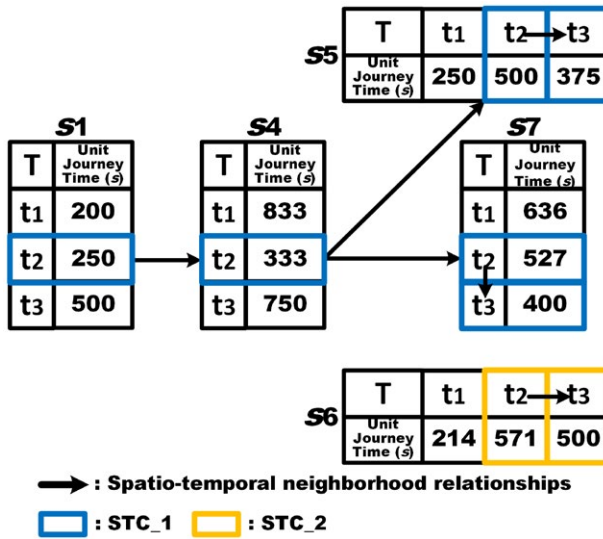


FIGURE 7 An example of spatio-temporal clustering. Here, s_1 , s_4 , s_5 , s_6 , and s_7 are consistent with the corresponding road segments in Figure 3; the constructed spatio-temporal neighborhood relationships among distinct cells are represented by arrows; STC denotes spatio-temporal clusters

3.4 | The implementation of the proposed method

Based on the aforementioned definitions, the pseudo-code of the proposed method can be elaborated as in Algorithm 1.

Algorithm 1: [Detection of spatio-temporal clusters]

Input: A road network RN , traffic time series TTS , thresholds ϵ_1 , ϵ_2 , and ϵ_3

Output: Spatio-temporal clusters $STCs$

Procedure:

BEGIN

 For each road segment s_i on the road network RN

 Do{

$s_i.SN = \text{Get_SN}(RN)$; // get spatial neighborhoods of s_i

 }

 End For

 For each spatio-temporal cell stc_{s_i, t_j} in the traffic time series TTS

$stc_{s_i', t_j'} = stc_{s_i, t_j}$;

 Do{

$stc_k = \text{Get_STF}(stc_{s_i, t_j}, s_i'.SN, TTS, \epsilon_1)$; // construct spatio-temporal flow $STF(stc_{s_i, t_j})$

 If $A_Diff[STF(stc_{s_i, t_j} \rightarrow \dots), stc_k] \leq \epsilon_2$ && $R_Diff_{stc_{s_i, t_j} \rightarrow stc_k} \leq \epsilon_3$ && ...

$A_Diff[STF(stc_{s_i, t_j} \rightarrow \dots), stc_k] \leq \epsilon_2$ && $R_Diff[STF(stc_{s_i, t_j} \rightarrow \dots), stc_k] \leq \epsilon_3$

 Then

$stc_{s_i, t_j}.STN.add(stc_k)$; // get spatio-temporal neighborhoods of stc_{s_i, t_j}

$stc_{s_i', t_j'} = stc_k$;

 Else

 Exit Do

 End If

 }

End For

For each spatio-temporal cell stc_{s_i, t_j} in the traffic time series TTS

 Do{

$STCs = \text{Sub_Graph}(stc_{s_i, t_j}.STN, TTS)$; // Get spatio-temporal clusters

 }

End For

END

The most time-consuming part is the construction of spatio-temporal neighborhoods considering spatio-temporal flows. In this part, the time complexity derives primarily from determining the next arrived spatio-temporal cell and determining whether this cell can be placed into the spatio-temporal neighborhood of the starting cell. Let N and M denote the number of spatio-temporal cells and the average volume of spatio-temporal neighborhoods, respectively. The time complexity of this process is approximately $O(N * M)$, where $N \gg M$. As a result, it is available for the proposed method to perform on large traffic datasets.

4 | EXPERIMENTAL COMPARISONS AND ANALYSIS

This section evaluates the effectiveness and practicality of the proposed method with experiments performed on real-life datasets. Section 4.1 elaborates the utilized datasets. Sections 4.2 and 4.3 elaborate the comparisons and analysis of the results by performing experiments on two groups of datasets. A discussion of the experimental results is given in Section 4.4. The proposed method was implemented using the MATLAB language. All the experiments were conducted on a computer with the Mac OS operating system, one 2.9 GHz Intel Core i5 CPU, and 8.0 GB RAM.

4.1 | Real-life datasets

The experimental datasets were captured by the London Congestion Analysis Project (LCAP) network, which is a system of automatic number plate recognition cameras designed to collect the journey time of vehicles on the road network of London. Specifically, the cameras were installed on the vertices of the road network and utilized to read the number plates of passing vehicles. The time that vehicles take to pass two adjacent cameras was recorded as the journey time on the corresponding road segment. On this basis, the average journey time of vehicles on each road segment every 300 s was calculated to form the traffic time series. As very few vehicles moved on the roads in the night, unreliable data was unavoidably collected during this time period. As a result, only those traffic time series recorded during the daytime period (i.e. 7:00–20:00 in this study) are adopted to perform the experiments. For each day, 12 record items were captured each hour and 156 data items were recorded in total. The road network in central London includes 22 road segments and the traffic time series were recorded from January 5th, 2009 to March 5th, 2009.

Considering the heterogeneity of road segment length, the time series regarding unit journey time on each road segment (i.e. the total journey time of vehicles divided by the length of the corresponding road segment) was utilized in our experiments. Figure 8 presents the road network structure of central London, reproduced from Cheng, Haworth, and Wang (2012). It indicates the real spatial distribution of the road network and describes the information regarding the vertices constituting each segment, the length of each segment, and the topological and directional relationships between segments. To facilitate the representation and analysis, the road network was simplified using simple systematic signs to denote each segment [i.e. (*Road1*, *Road2*, ..., *Road22*)].

To detect spatio-temporal clusters hidden in traffic flows on workdays and weekends, the traffic time series recorded on two days (i.e. January 5th, 2009—Monday and January 10th, 2009—Saturday) were employed in our experiments.

4.2 | Case study I: Spatio-temporal cluster detection in traffic time series on January 5th, 2009

To demonstrate the effectiveness and practicability of the proposed method, spatio-temporal clusters were first detected from the traffic data on January 5th, 2009. For comparison, a classical spatio-temporal clustering

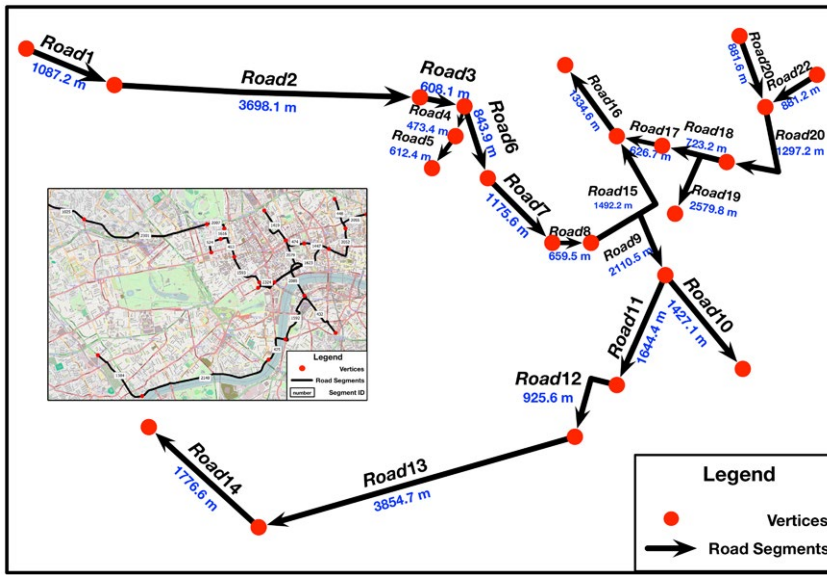


FIGURE 8 The distribution of the road network in the center of London, including the spatial distribution in the real world and detailed information regarding the road network (Cheng et al., 2012)

algorithm ST-DBSCAN (Birant & Kut, 2007), derived from the commonly used density-based clustering method DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), was also applied to this dataset.

4.2.1 | The testing of thresholds

For any cell stc_{si-t_j} , the members in its spatio-temporal neighborhoods $STN(stc_{si-t_j})$ may be broken points with respect to other cells' spatio-temporal neighborhoods. These spatio-temporal neighborhoods are defined as being adjacent to $STN(stc_{si-t_j})$. The non-spatial attribute differences between $STN(stc_{si-t_j})$ and its adjacent spatio-temporal neighborhoods, abbreviated as inter-neighborhood differences, can be used to quantize the independence of $STN(stc_{si-t_j})$. In addition, the non-spatial attribute similarity between different members of $STN(stc_{si-t_j})$, abbreviated as intra-neighborhood similarities, can describe the homogeneity of $STN(stc_{si-t_j})$. According to the objective of spatio-temporal clustering, the ratio between intra-neighborhood similarities and inter-neighborhood differences is calculated as an index (denoted $STNs_QI$) to quantitatively evaluate the overall quality of the constructed spatio-temporal neighborhoods. A more detailed process of calculation is discussed in our previous work (Shi et al., 2018). A small $STCs_QI$ indicates small differences in inner spatio-temporal neighborhoods and large differences between adjacent spatio-temporal neighborhoods. That is, a smaller $STCs_QI$ corresponds to a result that is more acceptable. Based on this criterion, the selection of thresholds is elaborated as follows.

In the three thresholds, ϵ_1 and ϵ_3 are utilized to adaptively homogenize spatio-temporal neighborhoods, so they were tested first by ignoring ϵ_2 (i.e. $\epsilon_2 = +\infty$). By setting ϵ_3 to 2.0, 2.5, and 3.0, Figure 9a displays the testing results, where ϵ_1 changed in the range [1 s/km, 50 s/km] with an interval of 1 s/km. One can see that the $STCs_QI$ all have apparent local minimum values in the range [17 s/km, 25 s/km] for ϵ_1 . Different values of $STCs_QI$ were further calculated by setting ϵ_1 to 17 s/km, 18 s/km, ..., 25 s/km, respectively, with ϵ_3 changing in the range [1.1, 1.2, ..., 5.0]. The minimum value of $STCs_QI$ appeared in [$\epsilon_1 = 20$ s/km, $\epsilon_3 = 2.5$]. Thus, [$\epsilon_1 = 20$ s/km, $\epsilon_3 = 2.5$] was adopted to further test ϵ_2 . Let ϵ_2 change from 1 s/km to 250 s/km with an interval of 1 s/km to capture a series of $STCs_QI$ values, as shown in Figure 9b. This reveals that $STCs_QI$ gradually tends to be stable when ϵ_2 attains 70 s/km. This means that a certain result can be captured with a combination of ϵ_1 and ϵ_3 , except for those extremely

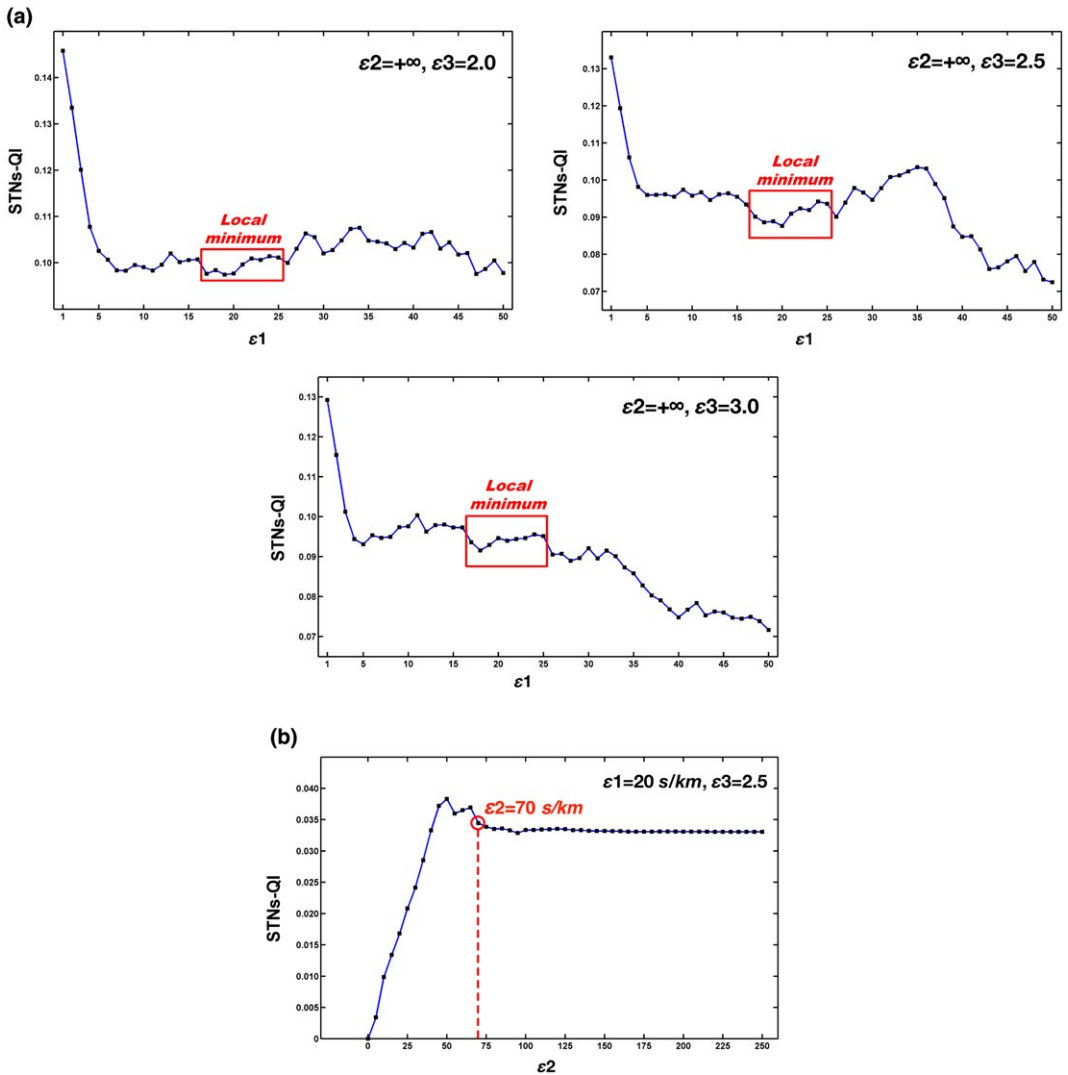


FIGURE 9 The testing of thresholds: (a) ϵ_1 ; and (b) ϵ_2

large distances that must be identified by ϵ_2 . As a result, the thresholds are set to $[\epsilon_1, \epsilon_2, \epsilon_3] = [20 \text{ s/km}, 70 \text{ s/km}, 2.5]$.

4.2.2 | Experimental results

Using the determined thresholds, the proposed method spent 1.20 s to obtain the spatio-temporal clustering results, as shown in Figure 10. Specifically, Figure 10a indicates the distribution of spatio-temporal clusters, denoted by different symbols, in the traffic time series by amplification. In these results, those small clusters containing relatively few cells can be regarded as outliers (He, Xu, & Deng, 2003). Generally, a numeric parameter should be assigned to distinguish large and small clusters. Here, clusters containing less than 10 cells were identified as outliers and discarded. One can see that there were three significantly large clusters, namely *Cluster 1*, *Cluster 2*, and *Cluster 3*. *Cluster 1* virtually occupied the entire time period; *Cluster 2* and *Cluster 3* were in the 7:00–10:00 and

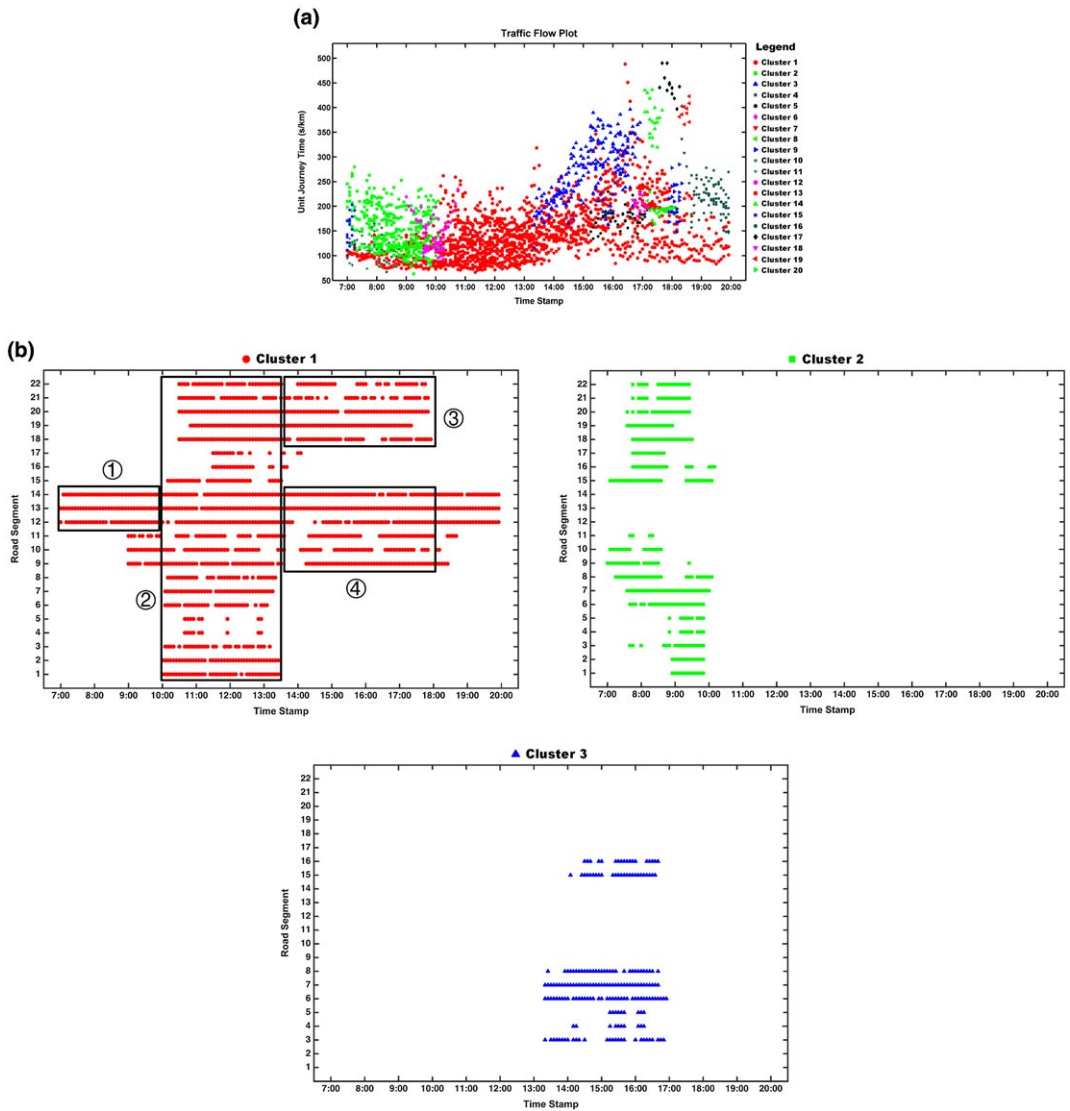


FIGURE 10 Spatio-temporal clusters obtained by the proposed method from the traffic time series on January 5th, 2009: (a) the distribution in the traffic time series, where cells in the same cluster are denoted by the same symbol; and (b) the spatio-temporal distribution of the top three large clusters, where each cluster is denoted using the symbol in accordance with that in Figure 10(a)

13:00–17:00 periods, respectively. Figure 10b exhibits the spatio-temporal distribution of *Cluster 1*, *Cluster 2*, and *Cluster 3*, respectively. The spatio-temporal flow in *Cluster 1* could basically be divided into four parts: ① Road 12 → Road 13 → Road 14 in 7:00–10:00; ② the entire road network in 10:00–13:30; ③ $\left. \begin{matrix} \text{Road22} \\ \text{Road21} \end{matrix} \right\} \rightarrow \text{Road20} \rightarrow \left\{ \begin{matrix} \text{Road19} \\ \text{Road18} \end{matrix} \right.$ in 13:30–18:00; and ④ $\text{Road9} \rightarrow \left\{ \begin{matrix} \text{Road10} \\ \text{Road11} \rightarrow \text{Road12} \rightarrow \text{Road13} \rightarrow \text{Road14} \end{matrix} \right.$ in 13:30–18:00. *Cluster 2* contained the majority of cells on all road segments, except Roads 12, 13, and 14 in

7:00–10:00. Further, the spatio-temporal flow $Road3 \rightarrow \begin{cases} Road4 \rightarrow Road5 \\ Road6 \rightarrow Road7 \rightarrow Road8 \rightarrow Road15 \rightarrow Road16 \end{cases}$ in 13:30–

17:00 was discovered in *Cluster 3*.

In summary, in 7:00–10:00, Roads 12, 13, and 14 had shorter unit journey time (approximately 100 s/km), which separated them from other segments. As the volume of vehicles gradually increased in the three segments and decreased in other segments, the two parts of segments met at approximately 10:00. Then, after 13:30, the spatio-temporal flow in *Cluster 3* was separated from *Cluster 1* owing to the longer unit journey time. In fact, the most significant cluster (i.e. *Cluster 1*) described the normal spatio-temporal distribution pattern of traffic time series on January 5th, 2009. Meanwhile, *Cluster 2* and *Cluster 3* revealed the main rush hours and congested road segments in the morning and afternoon, respectively. In addition, some small clusters (e.g. *Clusters 14, 17, and 19* in Figure 10a) also described traffic congestion; however, they represented anomalous situations as they included a very small number of cells.

For ST-DBSCAN, a heuristic strategy was proposed by Ester et al. (1996) to determine the necessary thresholds (i.e. *Eps* and *MinPts*). Specifically, *MinPts* is suggested to be set to $\ln(N)$ with respect to a database of size N . For each object, the distance to its *MinPts*-neighbor can be captured, denoted *MinPts*-distance. By sorting all the *MinPts*-distances in descending order, the distance corresponding to the first valley of the sorted curve is selected as the threshold *Eps*. Using the selected thresholds, 0.41 s were spent to obtain the spatio-temporal clustering results by ST-DBSCAN, as presented in Figure 11. For the top three large clusters, *Cluster 1* occupied the entire time period, whereas the cells in *Cluster 2* were located primarily in 15:00–20:00. The discreteness of the spatio-temporal clusters was further highlighted in Figure 11b. For example, only one spatio-temporal cell in Road 8 was assigned to *Cluster 1*, which formed a “neck” and weakened the connectivity of the cluster. Similarly, this kind of single cell existed in *Cluster 2* and *Cluster 3*.

Furthermore, it is necessary to perform quantitative evaluation of the proposed clustering method by means of clustering validation measures, which include external and internal validation. In the case that class labels are not given, it is suitable to employ internal clustering validation measures. Related work has demonstrated that the S_Dbw index outperforms other indices by considering the combined impacts of monotonicity, noise, density, subclusters, and skewed distributions (Liu, Li, Xiong, Gao, & Wu, 2010). The S_Dbw index is constructed with a combination of inter-cluster separation and intra-cluster compactness (Halkidi & Vazirgiannis, 2001). Specifically, the inter-cluster separation compares the densities between each pair of cluster centers and their midpoints. The ratio of the average standard deviation of clusters to the standard deviation of the dataset is calculated to measure the intra-cluster compactness. A small S_Dbw value indicates a suitable clustering result. Considering the unit journey time of traffic flows, the calculated S_Dbw values of clusters detected by the proposed method and ST-DBSCAN were 0.1694 and 0.1891, respectively. This illustrates that the proposed method outperformed ST-DBSCAN.

4.3 | Case study II: Spatio-temporal cluster detection in traffic time series on January 10th, 2009

To further analyze the difference regarding spatio-temporal clusters between weekends and workdays, the proposed method was also executed on the traffic time series on January 10th, 2009.

Based on the testing strategy introduced in Section 4.2.1, the thresholds ϵ_1 , ϵ_2 , and ϵ_3 could respectively be determined as 28 s/km, 70 s/km, and 1.4. Then, the proposed method captured the spatio-temporal clusters, as visualized in Figure 12, after running for 1.13 s. Figure 12a indicates that the spatio-temporal clusters were more scattered compared with those of the traffic time series on January 5th, 2009. Figure 12b can summarize the spatio-temporal distribution of the top three large clusters as: *Cluster 1* included all the road segments except

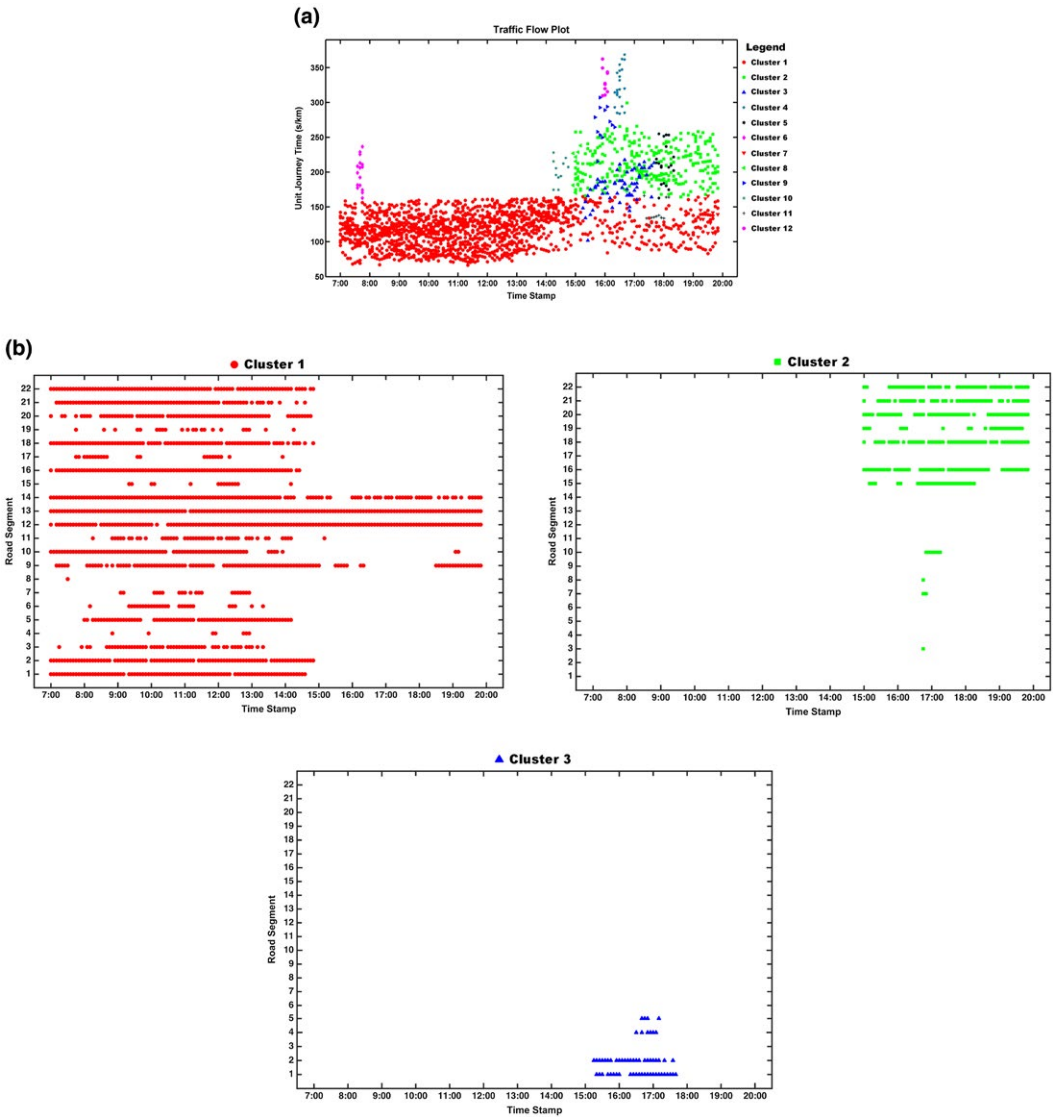


FIGURE 11 Spatio-temporal clusters obtained by ST-DBSCAN from the traffic time series on January 5th, 2009: (a) the distribution in the traffic time series, where cells in the same cluster are denoted by the same symbol; and (b) the spatio-temporal distribution of the top three large clusters, where each cluster is denoted using the symbol in accordance with that in Figure 11(a)

Road 1, Road 2, and Road 5 in 17:00–20:00; Cluster 2 was formed by Road9 → $\left\{ \begin{array}{l} \text{Road10} \\ \text{Road11} \rightarrow \text{Road12} \end{array} \right.$ in 7:00–12:00;

Cluster 3 could be divided into ① Road 12 → Road 13 → Road 14 in 12:30–16:00; and ②

Road9 → $\left\{ \begin{array}{l} \text{Road10} \\ \text{Road11} \rightarrow \text{Road12} \rightarrow \text{Road13} \rightarrow \text{Road14} \end{array} \right.$ in 16:00–18:00.

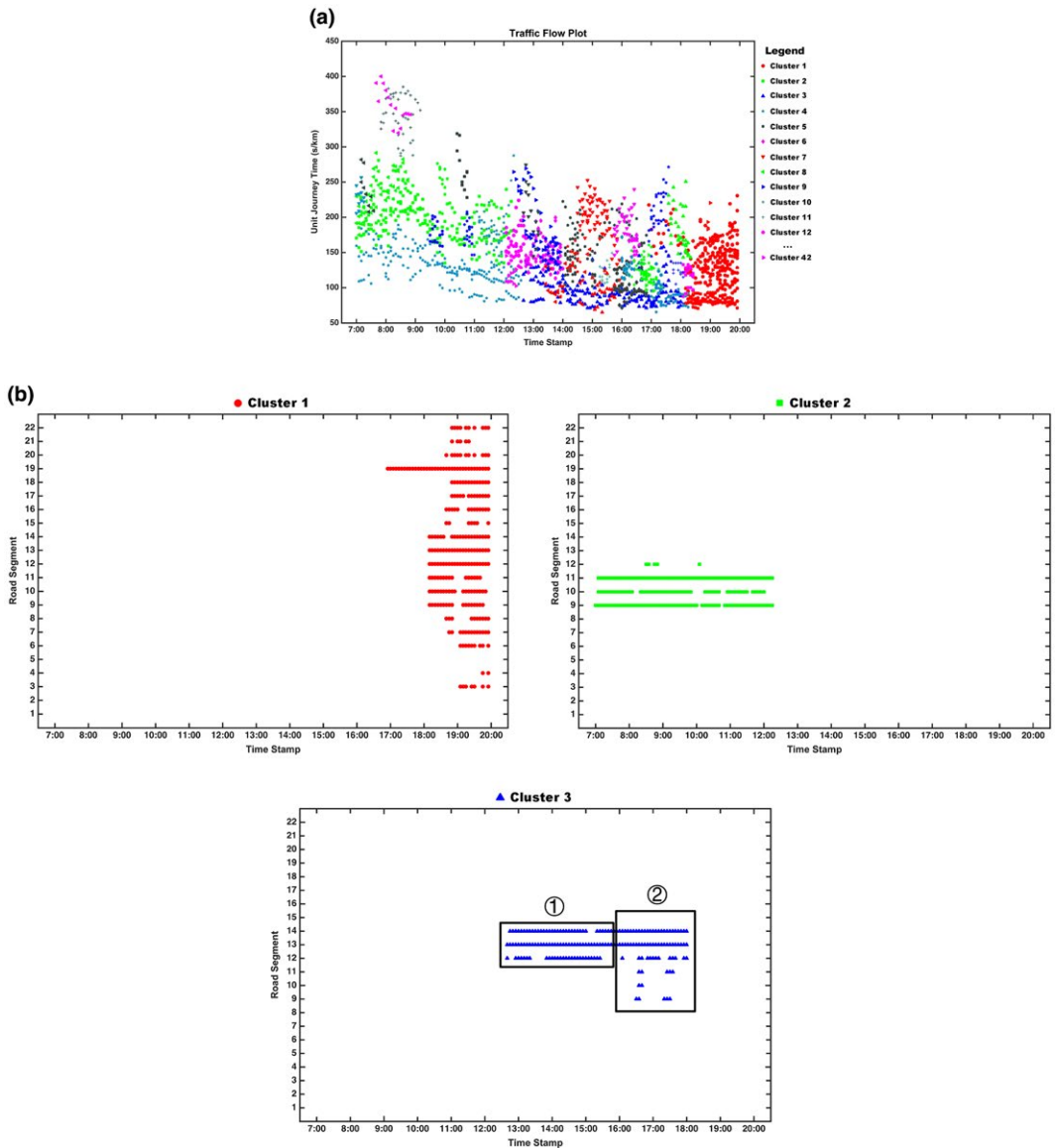


FIGURE 12 Spatio-temporal clusters obtained by the proposed method from the traffic time series on January 10th, 2009: (a) the distribution in the traffic time series, where cells in the same cluster are denoted by the same symbol; and (b) the spatio-temporal distribution of the top three large clusters, where each cluster is denoted using the symbol in accordance with that in Figure 12(a)

In summary, the unit journey time was small in the three clusters, approximately 100–300 s/km. They basically divided the day into three time periods: 7:00–12:00, 12:30–17:00, and 17:00–20:00. More importantly, *Cluster 1* in Figure 10 and *Cluster 3* in Figure 12 both highlighted the spatio-temporal influence on the traffic time series. Specifically, Roads 12, 13, and 14 were located downstream on the road network. Hence, during 12:30–16:00, the cells in these three segments were self-contained (i.e. Part ① of *Cluster 3*); then, they were gradually influenced by the upstream cells and merged with them to constitute Part ② of *Cluster 3*. This could reflect the dynamic nature of traffic flows on a road network.

4.4 | Discussion of experimental results

In Cheng et al. (2012), both the global and the local spatio-temporal autocorrelation structure of the road network of central London was analyzed, choosing the traffic time series for 33 consecutive Tuesdays. They divided the traffic time series on weekdays into three periods: 7:00–10:00 (AM peak), 10:00–16:00 (interpeak), and 16:00–19:00 (PM peak) by considering the different behaviors of traffic flows. Increasing unit journey time during the AM peak indicated the building of rush hour. Focusing on the clustering results on Monday (a weekday) by the proposed method, as shown in Figure 10, the rush hour in the AM peak could be characterized by *Cluster 2*. One can see that spatio-temporal cells in Roads 12–14 were separated from others by *Cluster 2* due to the shorter unit journey time. In essence, local stronger cross-correlations in the range of Roads 12–14 were extracted by Cheng et al. (2012). Then, the gradually decreasing unit journey time reflected that the vehicles on the roads returned to free-flow during the interpeak period. Correspondingly, *Cluster 1* in Figure 10 presented the evolving process of free-flowing traffic status and vehicles on all roads were under free-flow during 10:00–13:00. After 13:00, *Cluster 3* revealed that Roads 3–8 and Roads 15–16 became busy again. Consistently, Cheng et al. (2012) pointed out that the PM peak should begin earlier due to the stronger seasonal component in the interpeak period. The comparisons illustrate that consistent discoveries were obtained with Cheng et al. (2012)'s conclusions from spatio-temporal autocorrelation analysis, which can demonstrate the effectiveness and practicability of the proposed method.

Nevertheless, the results analyzed by Cheng et al. (2012) could not be repeated by ST-DBSCAN. ST-DBSCAN needs users to set a fixed spatio-temporal coverage, so it can only extract those static spatio-temporal regions with local high density. In the real world, traffic flows on road networks present a dynamic nature. The above discussion illustrates that the proposed method can capture this characteristic and provide more suitable results compared with ST-DBSCAN.

Moreover, the spatio-temporal clusters extracted by the proposed method can further elaborate the temporally dynamic and spatially heterogeneous spatio-temporal autocorrelation structure of traffic data on a road network obtained by Cheng et al. (2012). Different spatio-temporal distribution patterns of traffic flows were also discovered on weekends compared with weekdays in this study. This will facilitate the making of distinct traffic management plans for weekdays and weekends by traffic operators.

5 | CONCLUSIONS AND FUTURE WORK

In this article, a novel approach was developed for the accurate detection of clusters in traffic time series on road networks by handling the temporally dynamic and spatially heterogeneous correlations among road segments. Spatio-temporal flows are first modeled based on the topological relationships of the road network and the real-time traffic status. For each spatio-temporal cell, its spatio-temporal neighborhood was built by considering both spatio-temporal reachability and spatio-temporal connectivity in its spatio-temporal flow. Further, spatio-temporal clusters are detected by extracting connected graphs based on spatio-temporal neighbors. Compared with a classical spatio-temporal clustering method (ST-DBSCAN), the effectiveness and practicality of the proposed method were demonstrated by experimenting on the traffic time series on both weekdays and weekends on the road network of central London. For practical purposes, employing the proposed method to cluster traffic time series day by day on the whole network can help traffic operators discover the spatio-temporal characteristics of traffic flows on each day. Furthermore, by comparing the results of different days in a month or in a year, they can further discover periodical patterns of spatio-temporal clusters, which will be valuable in setting up optimum traffic management plans.

Our future work will be focused on three aspects. First, the selection of thresholds can to a large degree limit the practicability of the proposed method. As a result, it should be optimized by employing domain-related knowledge to guide the selection and testing of thresholds for different application demands in the future. Second, is

to construct a statistical clustering model by considering the statistical significance of thresholds involved in the proposed method. Third, is to improve the proposed method, adapting it so as to support the processing of real-time data streams for traffic monitoring.

ORCID

Yan Shi  <https://orcid.org/0000-0002-9136-9764>

Min Deng  <https://orcid.org/0000-0003-3305-9757>

REFERENCES

- Anbaroglu, B., Heydecker, B., & Cheng, T. (2014). Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, 48, 47–65.
- Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowledge Engineering*, 60, 208–221.
- Chen, Y., Zhang, Y., Hu, J., & Yao, D. (2006). Pattern discovering of regional traffic status with self-organizing maps. In *Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference* (pp. 647–652). Toronto, ONT: IEEE.
- Cheng, T., & Anbaroglu, B. (2010). Spatio-temporal clustering of road network data. In F. L. Wang, H. Deng, Y. Gao, & J. Lei (Eds.), *Artificial Intelligence and Computational Intelligence: AICI 2010* (Lecture Notes in Computer Science, Vol. 6319, pp. 116–123). Berlin, Germany: Springer.
- Cheng, T., Haworth, J., & Wang, J. (2012). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, 14, 389–413.
- Cheng, T., Wang, J., Haworth, J., Heydecker, B., & Chow, A. (2014). A dynamic spatial weight matrix and localized space-time autoregressive integrated moving average for network modeling. *Geographical Analysis*, 46, 75–97.
- Delmelle, E., Dony, C., Casas, I., Jia, M., & Tang, W. (2014). Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science*, 28, 1107–1127.
- Deng, M., Liu, Q., Wang, J., & Shi, Y. (2013). A general method of spatio-temporal clustering analysis. *Science China Information Sciences*, 56, 102315.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Portland, OR: AAAI.
- Feng, C.-C., Wang, Y.-C., & Chen, C.-Y. (2014). Combining Geo-SOM and hierarchical clustering to explore geospatial data. *Transactions in GIS*, 18, 125–146.
- Hagenauer, J., & Helbich, M. (2013). Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science*, 27, 2026–2042.
- Halkidi, M., & Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings of the 2001 IEEE International Conference on Data Mining* (pp. 187–194). San Jose, CA: IEEE.
- Han, J., Kamber, M., & Tung, A. K. H. (2001). *Geographic data mining and knowledge discovery*. London, UK: Taylor & Francis.
- He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24, 1641–1650.
- Hu, C., Luo, N., Yan, X., & Shi, W. (2011). Traffic flow data mining and evaluation based on fuzzy clustering techniques. *International Journal of Fuzzy Systems*, 13, 344–349.
- Kang, J. M., Shekhar, S., Wennen, C., & Novak, P. (2008). Discovering flow anomalies: A SWEET approach. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 851–856). Pisa, Italy: IEEE.
- Kisilevich, S., Mansmann, F., & Nanni, M. (2010). Spatio-temporal clustering. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed.). New York, NY: Springer.
- Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005). A space-time permutation scan statistics for disease outbreak detection. *PLoS Med*, 2, 216–224.
- Liu, Q., Deng, M., Bi, J., & Yang, W. (2014). A novel method for discovering spatio-temporal clusters of different sizes, shapes and densities in the presence of noise. *International Journal of Digital Earth*, 7, 138–157.
- Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining* (pp. 911–916). Sydney, Australia: IEEE.
- Mehboob, F., Abbas, M., Almotaery, R., Jiang, R., Al-Maadeed, S., & Bouridane, A. (2015). Traffic flow estimation from road surveillance. In *Proceedings of the 2015 IEEE International Symposium on Multimedia* (pp. 605–608). Miami, FL: IEEE.

- Miller, H., & Han, J. (2009). *Geographic data mining and knowledge discovery* (2nd ed.) Boca Raton, FL: CRC Press.
- Min, X., Hu, J., & Zhang, Z. (2010). Urban traffic network modeling and short-term traffic flow forecasting based on GSTARIMA model. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems* (pp. 1535–1540). Madeira Island, Portugal: IEEE.
- Nakaya, T., & Yano, K. (2010). Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14, 223–239.
- Ntoutsis, I., Mitsou, N., & Marketos, G. (2008). Traffic mining in a road-network: How does the traffic flow? *International Journal of Business Intelligence & Data Mining*, 3, 82–98.
- Pei, T., Zhou, C. H., Zhu, A. X., Li, B., & Qin, C. (2010). Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise. *International Journal of Geographical Information Science*, 24, 925–948.
- Ren, J., Zhang, C., Zhang, L., Wang, N., & Feng, Y. (2018). Automatic measurement of traffic state parameters based on computer vision for intelligent transportation surveillance. *International Journal of Pattern Recognition & Artificial Intelligence*, 32(4), 1855003.
- Shekhar, S., Vatsavai, R. R., & Celik, M. (2009). Spatial and spatio-temporal data mining: Recent advances. In H. Kargupta, J. Han, P. S. Yu, R. Motwani, & V. Kumar (Eds.), *Next generation of data mining* (pp. 549–584). Boca Raton, FL: CRC Press.
- Shi, Y., Deng, M., Yang, X., & Gong, J. (2018). Detecting anomalies in spatio-temporal flow data by constructing dynamic neighbourhoods. *Computers, Environment & Urban Systems*, 67, 80–96.
- Shiode, S., & Shiode, N. (2013). Network-based space-time search-window technique for hotspot detection of street-level crime incidents. *International Journal of Geographical Information Science*, 27, 866–882.
- Wu, X., Zurita-Milla, R., & Kraak, M. J. (2015). Co-clustering geo-referenced time series: Exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science*, 29, 624–642.
- Wu, X., Zurita-Milla, R., Verdiguier, E. I., & Kraak, M. J. (2017). Triclustering georeferenced time series for analyzing patterns of intra-annual variability in temperature. *Annals of the American Association of Geographers*, 108, 71–87.
- Xie, Z., & Yan, J. (2013). Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: An integrated approach. *Journal of Transport Geography*, 31, 64–71.
- Zhang, P., Huang, Y., Shekhar, S., & Kumar, V. (2003). Correlation analysis of spatial time series datasets: A filter-and-refine approach. In K. Shim, K.-Y. Wang, J. Jeon, & J. Srivastava (Eds.), *Advances in Knowledge Discovery and Data Mining: Proceedings of the 22nd Pacific-Asia Conference* (Lecture Notes in Computer Science, Vol. 2637, pp. 532–544). Berlin, Germany: Springer-Verlag.
- Zhou, Z., Lin, S., & Xi, Y. (2013). A fast network partition method for large-scale urban traffic networks. *Journal of Control Theory & Applications*, 11, 359–366.

How to cite this article: Shi Y, Deng M, Gong J, Lu C-T, Yang X, Liu H. Detection of clusters in traffic networks based on spatio-temporal flow modeling. *Transactions in GIS*. 2019;23:312–333. <https://doi.org/10.1111/tgis.12521>