Pervasive and Mobile Computing 41 (2017) 470–483

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Fast track article

Exploring traffic congestion correlation from multiple data sources

Yuqi Wang [a,*], Jiannong Cao [a,*], Wengen Li [a], Tao Gu [b], Wenzhong Shi [c]

[a] Department of Computing, Hong Kong Polytechnic University, Hong Kong

[b] School of Computer Science and IT, RMIT University, Australia

[c] Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Available online 4 April 2017

Keywords:
Traffic congestion
Congestion correlation
Multiple data sources
Classification

ABSTRACT

Traffic congestion is a major concern in many cities around the world. Previous work mainly focuses on the prediction of congestion and analysis of traffic flows, while the congestion correlation between road segments has not been studied yet. In this paper, we propose a three-phase framework to explore the congestion correlation between road segments from multiple real world data. In the first phase, we extract congestion information on each road segment from GPS trajectories of over 10,000 taxis, define congestion correlation and propose a corresponding mining algorithm to find out all the existing correlations. In the second phase, we extract various features on each pair of road segments from road network and POI data. In the last phase, the results of the first two phases are input into several classifiers to predict congestion correlation. We further analyze the important features and evaluate the results of the trained classifiers through experiments. We found some important patterns that lead to a high/low congestion correlation, and they can facilitate building various transportation applications. In addition, we found that traffic congestion correlation has obvious directionality and transmissibility. The proposed techniques in our framework are general, and can be applied to other pairwise correlation analysis.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid process of urbanization, traffic congestion becomes an increasingly serious problem in more and more cities around the world. Understanding, alleviating, and further tackling traffic congestion have received urgent attentions from governments and their citizens. Much research work has been conducted to study congestion from different aspects, including traffic congestion prediction [1], traffic condition estimation [2], impact [3] and correlation [4] of traffic congestion and traffic flow propagation [5]. They provide many useful insights on traffic congestions, which may facilitate the building of many practical applications.

However, the existing work typically assumes or ignores correlations [6], leaving the impact of correlated patterns on traffic congestion largely unknown. Analyzing and uncovering the correlated patterns in traffic congestion can reveal the insights of congestion such as what factors are correlated in congestion and how congestions propagate from one road to another. Furthermore, it can also facilitate building various applications including road planning, traffic condition prediction,

* Corresponding authors.
E-mail addresses: csyqwang@comp.polyu.edu.hk (Y. Wang), csjcao@comp.polyu.edu.hk (J. Cao), cswgli@comp.polyu.edu.hk (W. Li), tao.gu@rmit.edu.au (T. Gu), lswzshi@polyu.edu.hk (W. Shi).

http://dx.doi.org/10.1016/j.pmcj.2017.03.015
1574-1192/© 2017 Elsevier B.V. All rights reserved.

impact analysis of congestion and etc. As such, both governments and individuals can benefit. For example, when a person is stuck in traffic congestion, the information about nearby congestion correlated road segments (i.e., these roads are likely to be congested as well) will be very useful since s/he can better estimate the traveling time, or possibly choose to bypass those roads to avoid congestion. Besides, with the information of congestion correlation between road segments acquired, governments are able to make better decisions on traffic light management, road planning, etc.

To fill the gap of existing work on congestion correlation analysis, we utilize multiple real world data to predict whether a road segment is correlated with another one in terms of congestion, where we can uncover some correlated congestion patterns from features on road segments. Thanks to the wide deployment of GPS devices and the widely available road and Point Of Interest (POI) information, we are able to obtain congestion information and features on road segments easily. To analyze the correlation between road segments, we apply a mining algorithm to find out all the existing correlations, and extract features on each road segment pair. We then build learning models based on classifiers to infer the correlated road segments from data. The models also help to identify some important features and correlated patterns.

To the best of our knowledge, we are the first to explore traffic congestion correlation from a classification perspective using real world datasets. In sum, our contributions are four folds:

- We propose a novel framework to explore traffic congestion correlation between road segments. The framework utilizes multiple sources of data to mine and analyze congestion correlation. In addition, the framework is general, and can be applied to other pairwise correlation analysis problems as well.
- We analyze the congestion correlation patterns based on our formulation, and found traffic congestion correlation has obvious directionality and transmissibility.
- We focus on congestion analysis of two peak periods during a day, train two corresponding models on several well-known classifiers to predict congestion correlation, and compare the results of different models on different feature sets.
- We predict congestion correlation and found some important patterns, such as congestions are very likely to propagate between trunk roads during the evening peak hours, which can facilitate the decision making for both individuals and governments.

The rest of this paper is organized as follows. In Section 2, we summarize the related work. Section 3 gives an overview of the proposed framework. Section 4 details each phase of the framework. Section 5 shows the experimental and analysis results. We conclude the paper in Section 6.

2. Related work

This section surveys the related work on traffic congestion prediction, traffic condition estimation, impact and correlation of congestion and traffic propagation. In [7], Yang formulated congestion prediction as a binary classification problem and applied feature selection techniques to reduce the dimensionality of data, yet still maintained the comparable accuracy. In [8], Min et al. proposed an approach based on the multivariate spatial–temporal autoregressive model to incorporate spatial and temporal characteristics for real-time traffic prediction, and found that congestion can change the traffic flow patterns. Gajewski et al. proposed a Bayesian-based approach in [9] to estimate link traveling time correlation, and found that the heavier the congestion, the lower the correlation of traveling time between links. In [6], Rachtan et al. argued that correlation patterns among the traffic variables are largely unknown, while most of the work ignores congestion correlation or assumes correlation exits. Jenelius et al. estimated traveling time based on low frequency GPS data in [2], and demonstrated that there is significant correlation between segments and showed the feasibility of using low frequency GPS data for monitoring the performance of transport system. In [10,5], the authors studied the traffic flow propagation by simulation. In [11,12], the authors reviewed several approaches on traffic density estimation, detection and avoidance. In [4, 13,3,14], the authors studied the impact and correlation on weather, accident, employment, safety, respectively.

Different from the above work, we focus on congestion correlation between road segments, which can benefit various applications including traffic prediction, traffic light management, road planning and etc.

3. Overview

Fig. 1 presents the framework of our work. In this framework, we utilize three sources of data i.e. GPS trajectory of taxis, road network and POI data to explore the congestion correlation between road segments. We divide the framework into following three phases.

(1) Congestion and correlation extraction: Extract congestion information on each road segment from GPS trajectories and road network data, define and mine congestion correlation between each road segment pair.
(2) Feature and sample generation: Extract various topological features and POI features from road network and POI data, respectively, and generate training samples on road segment pairs.
(3) Classification and analysis: Input the results of the first two phases into several classifiers to predict congestion correlation, and analyze the evaluation results for pattern discovery.

We design the framework in a way that it is general enough to be used for other pairwise correlation analysis problems by changing the specific data sources and implementing techniques such as feature extraction and correlation definition.

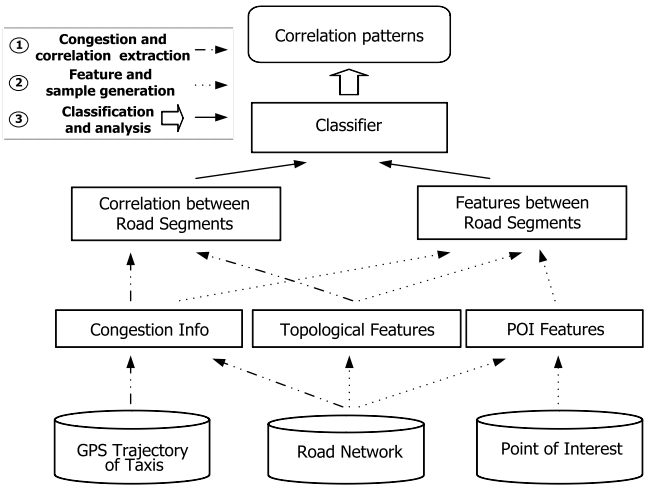


Fig. 1. Framework of congestion correlation mining.

4. Methodology

In this section, we describe the proposed framework in details. Specifically, we first present the three data sources we use, and then show how each phase of the framework works.

4.1. Data sources

Traffic congestion usually results from multiple factors. Intuitively, the underlying transportation infrastructure and human mobility are the major ones. Therefore, in this work, we exploit three data sources, i.e., road network, GPS trajectories of taxis and POIs to cover these major factors. Concretely, road network describes the spatial topology of the transportation infrastructure; GPS trajectory of taxis contain the traffic information related to human mobility; and POIs implicitly convey some information about the mobility of people whose daily activities are relevant to them. We formalize these information as follows.

Definition 1 (*Road Network*)**.** A road network is modeled as a directed graph $G = (V, E)$, where $v_i \in V$ represents an intersection of road segments, and $e_{i,j} \in E$ represents the direct road segment from v_i to v_j.

Definition 2 (*GPS Point*)**.** A GPS point, *gp* is denoted by a quadruple, i.e., $gp = (TaxiID, t, s, l)$, where *TaxiID* is the identifier of the taxi, t is the time at which this GPS point is sampled, s is the speed of the taxi, and l is the spatial location consisting of longitude and latitude.

Definition 3 (*GPS Trajectory*)**.** A GPS trajectory, *tr*, is consisted of a sequence of GPS points, i.e., $tr = (gp_1, gp_2, \ldots, gp_n)$, where n is the length of *tr* and $gp_i.t \leq gp_j.t$ if $i \leq j$.

Definition 4 (*Point of Interest, POI*)**.** A POI, o_i, is denoted by $o_i = (ID, Cate, Lng, Lat)$, where *ID* is the identifier of o_i, *Cate* is the category of o_i, and *Lng* and *Lat* is the longitude and latitude, respectively, of the spatial location of o_i.

4.2. Congestion and correlation extraction

In this phase, we first extract the congestion information from the GPS trajectories of taxis on each road segment. After that, with a definition of congestion correlation between road segments, we propose a mining algorithm to find out all the existing congestion correlation from data.

4.2.1. Congestion extraction

To extract the congestion information, we need to first obtain the traffic information on each road segment. According to the definition of GPS trajectories, a GPS trajectory is a sequence of discrete spatial points. Thus, we need to map-match each GPS trajectory to the underlying road segments. In this work, we leverage the map-matching technique in [15]. Meanwhile, considering the time-consuming characteristic of map-matching operation, a spatial index R^*-tree [16] is built on all road segments to accelerate the process of map-matching. After map-matching, each road segment is associated with a set of GPS points capturing the traffic information there.

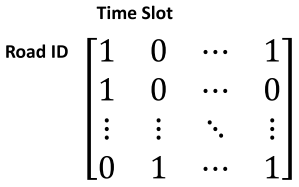


Fig. 2. Congestion matrix.

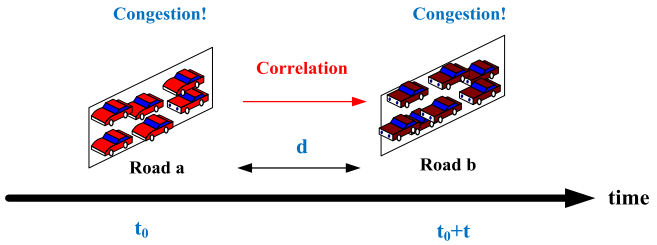


Fig. 3. Congestion correlation.

To extract congestion information from traffic information, we divide a day into time slots, and obtain the traffic information T_r^t on road segment r in a specific time slot t, using the average speed of all GPS points on road segment r in time slot t as the proxy. Then we have the definition of congestion as follows.

Definition 5 (*Congestion*)**.** A congestion on road segment r in a specific time slot t is denoted by C_r^t, and

$$C_r^t = \begin{cases} 1 & \text{if } T_r^t \leq T_r * Ratio; \\ 0 & \text{otherwise} \end{cases}$$

where T_r is the average speed of all GPS points in road segment r in all time, and *Ratio* is a parameter to set the speed threshold of congestion, and its settings will be discussed in Section 5.

We store the congestion information of a day in a congestion matrix as illustrated in Fig. 2, where each row represents a road segment and each column represents a time slot.

4.2.2. Correlation extraction

To study how congestion occurs sequentially in terms of time, and consider the propagation rate of congestions in terms of space, as shown in Fig. 3, we define congestion correlation between two road segments as follows.

Definition 6 (*Congestion Correlation between Two Road Segments*)**.** A congestion correlation from road segment a to road segment b, i.e. $Cor(a, b)$, occurs if the following requirements are satisfied:

(1) a congestion occurs on road a at time t_0
(2) from time t_0 to $t_0 + t$, a congestion occurs on road b
(3) a and b are within a certain distance d.

We propose Algorithm 1 to mine all congestion correlations in a designated time period, i.e., from t_{start} to t_{end}. The correlations are stored in a square matrix R, where R_{ik} stores the occurrence count of congestion correlation between road segments i and k from t_{start} to t_{end}.

In Algorithm 1, at each time slot j, for each congested road segment i, we retrieve all the congested road segments in next t time slots, and increase the occurrence count of correlation stored in $R_{i\cdot}$. We use a vector cv to store the retrieved congested road segments, so that the retrieving process only executes once in each time slot, thus improving the efficiency of the algorithm. Then, we also check the distances of all pairs of road segments to make sure that the distance requirement is also satisfied. The time complexity of the proposed algorithm is $O(n^2m)$, where n is number of road segments and m is the number of time slots from t_{start} to t_{end}.

To further refine congestion correlation, we have the following definition about confidence of correlation.

Definition 7 (*Correlation Confidence*)**.** Correlation confidence from road segment a to road segment b, i.e., CC_{ab} indicates the confidence level of the congestion correlation and is computed as below:

$$CC_{ab} = \frac{\text{occurrence count of } Cor(a, b)}{\text{No. of congestions at } a}.$$

Algorithm 1 Congestion Correlation Mining

Input: the congestion matrix C, time threshold t, distance threshold d, start time slot t_{start} and end time slot t_{end};
Output: the correlation matrix R;

```
1: R = 0; Create a vector cv of size C.rowNumber;
2: for j = t_start to t_end do
3:    cv = 0;
4:    isFound = false;
5:    for i = 1 to C.rowNumber do
6:       if C[i][j] == 1 then
7:          if isFound == false then
8:             for k = 1 to C.rowNumber do
9:                for t = j+1 to j+t do
10:                  if C[k][t] == 1 then
11:                     cv[k] = 1;
12:                     break;
13:            isFound = true;
14:         for k = 1 to C.rowNumber do
15:            R[i][k] = R[i][k] + cv[k];
16: for i = 1 to C.rowNumber do
17:    for k = 1 to C.colNumber do
18:       if Dist(i, k) > d then
19:          R[i][k] = 0;
20: return R;
```

With the correlation confidence, an analogy to the confidence in Association Analysis [17], we are able to identify some false positive and true positive correlations, and use them to conduct analysis more accurately in later phases.

We also perform some analysis on the found congestion correlations, which will be further discussed in Section 5.4.1 later.

4.3. Feature and sample generation

In this phase, we first extract various features on each road segment from road network and POI data, and then fuse the features of each road segment pair to generate training samples.

4.3.1. Feature extraction

To extract features on each road segment from road network data, we consider not only their traditional features, including length, type, and degree, but also some advanced features, including betweenness and closeness. It is straightforward to extract those traditional features. Therefore, we will only detail how to extract the advanced features as follows.

In graph theory, betweenness is used to measure the importance of nodes in terms of the number of shortest paths passing them. The intuition is that a node is more important if more shortest paths go through it. The betweenness of a node v_i is computed with the following formula [18].

$$B(v_i) = \frac{1}{(N-1)(N-2)} \sum_{v_j, v_k \in V \wedge i \neq j \neq k} \frac{n_{jk}(v_i)}{n_{jk}} \tag{1}$$

where n_{jk} is the total number of shortest paths between nodes v_j and v_k, $n_{jk}(v_i)$ is the number of shortest paths between nodes v_j and v_k that pass node v_i.

Similarly, we compute the betweenness of a road segment, e_{i_1,i_2} as below (cf. Definition 1).

$$B(e_{i_1,i_2}) = \frac{1}{(N-1)(N-2)} \sum_{v_j, v_k \in V} \frac{n_{jk}(e_{i_1,i_2})}{n_{jk}} \tag{2}$$

where n_{jk} is the total number of shortest paths between nodes v_j and v_k, $n_{jk}(e_{i_1,i_2})$ is the number of shortest paths between nodes v_j and v_k that pass edge e_{i_1,i_2}.

According to [18], closeness centrality is used to measure the centrality of a node, v_i, in the network and is computed as below.

$$C(v_i) = \frac{N-1}{\sum_{j \in V \wedge j \neq i} netDis(v_i, v_j)} \tag{3}$$

Table 1
Extracted features on a road segment.

Features	Description
Length	The length of each road segment
Degree	The degree of each road segment
Type	Type of road segments, e.g., motorway and trunk
$B(e_{i,j})$	The betweenness of the road segment $e_{i,j}$
$C(e_{i,j})$	The closeness of the road segment $e_{i,j}$
#POIs	The total number of POIs
#CatPOIs	The number of POIs in each category
POI-TF-IDF	The tf-idf value of each POI category

where $netDis(v_i, v_j)$ is the network distance between nodes v_i and v_j.

To compute the closeness of a road segment, e_{i_1,i_2}, we change the formula above to the following form.

$$C(e_{i_1,i_2}) = \frac{N-1}{\sum_{e \in E \wedge e \neq e_{i_1,i_2}} netDis(e, e_{i_1,i_2})} \tag{4}$$

where $netDis(e, e_{i_1,i_2})$ is the network distance between edges e and e_{i_1,i_2} (cf. Eq. (6)).

To extract features from POI data on each road segment, we consider the total number of POIs, the number of POIs in each category, and the Term Frequency-Inverse Document Frequency (TF-IDF) value of each POI category. Specifically, we treat road segments as documents and POI categories as terms, and TF-IDF value indicates the importance of POI categories on road segments. Similar to [19], to compute TF-IDF value of the ith POI category of a given road segment, we have the following formula:

$$\text{TF-IDF}_i = \frac{n_i}{N} \times \log \frac{R}{\|\{r| \text{ the } i\text{th POI category} \in r\}\|} \tag{5}$$

where n_i is the number of POIs in the ith category and N is the total number of POIs on the given road segment. The first term calculates POI frequency in the given road segments, and the second term calculates the inverse segment frequency by taking the logarithm of a quotient, resulting from the number of road segments R divided by the number of segments which have POIs in ith category.

The extracted features are summarized in Table 1.

4.3.2. Sample generation

To generate training samples, considering all features extracted on a road segment, we need to fuse the features of each road segment pair, and generate features for each pair.

For length, degree, betweenness, closeness and total number of POI, we calculate their differences between two road segments, and then add them to the features for each pair of road segments. We also add network distance and Pearson similarity of POI TF-IDF value distributions between two road segments into features for each pair.

Network distance between road segments e_{i_1,i_2} and e_{j_1,j_2} is computed based on the underlying road network (cf. Definition 1), i.e.,

$$netDis(e_{i_1,i_2}, e_{j_1,j_2}) = \min_{i \in \{i_1,i_2\}, j \in \{j_1,j_2\}} \{netDis(v_i, v_j)\} \tag{6}$$

where $netDis(v_i, v_j)$ is the length of the shortest path between nodes v_i and v_j. To accelerate the computation of network distance, we index road network G with CH (Contraction Hierarchy) [20] which organizes G in a hierarchy structure.

For each distinct ordered combination of two road types in a pair of road segments, we create a binary indicator variable to represent the existence of it between road segments. For example, a road segment type is 'trunk' and that of the other is 'primary', then the corresponding indicator variable that represents the existence of the ordered combination 'trunk → primary' is set to 1, and all other indicator variables of this ordered pair are set to 0. The idea of this design is to see how congestion correlation varies from one road type to another. Slightly different, for each distinct ordered combination of two POI categories, we create a variable to represent its importance level by calculating the product of TF-IDF values of the two categories on each pair of road segments. The idea of this design is to see how congestion correlation varies from one POI category to another.

Finally, we apply Min–Max scaling [21] to scale all the features for each pair of road segments into the range of [0, 1], which not only enhances the performance of the trained models, but also facilitates the process of analysis on feature importance later, since the trained models are not biased towards the features simply due to their large numeric range.

The features for each road segment pair are summarized in Table 2.

4.4. Classification and analysis

In this phase, we input the results of the first two phases into several classifiers to predict congestion correlation, and analyze the evaluation results for pattern discovery.

Table 2
Features for each road segment pair.

Features	Description
Diff-Len	The difference of length
Diff-Degree	The difference of degree
Diff-B	The difference of betweenness
Diff-C	The difference of closeness
Diff-POI	The difference of the total number of POIs
netDis	The network distance
SimPOIs	Pearson similarity of POI TF-IDF value distributions
OrderedComb-types	The binary indicator variable for ordered combination of road types
OrderedComb-POI	The variable for ordered combination of POI categories

4.4.1. Classification

After the first two phases, we have all congestion correlations and extracted features for each pair of road segments. We now combine these two parts to generate training samples and build models for binary classification.

For any given pair of road segments, the models will predict whether there exists high congestion correlation between them. To refine and enhance the knowledge models learn from data, we set a threshold of Correlation confidence (cf. Definition 7) for positive class and negative class, respectively. Thus, we only keep those pairs of road segments, whose correlation confidence is higher than the threshold for positive class, and treat them as positive training samples; or lower than the threshold for negative class and higher than 0, and treat them as negative training samples.

Usually the classes of training samples are highly imbalanced, i.e., the samples in uncorrelated class are much more than those in the correlated class, which will impair the performance of classifiers. Therefore, we apply random majority undersampling (RUS) [22] to generate a balanced training samples.

Finally, we input the balanced training samples into well-known classifiers including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM), and then evaluate the performance of the built models using classic metrics.

4.4.2. Analysis

After the evaluation of the models, we analyze the built models for pattern discovery.

Feature importance indicates how important a feature is for the prediction of classifiers, which can help to identify important features and patterns during the analysis process. We employ different feature importance measures for different classifiers based on how those classifiers are built. For Decision Tree and Random Forest, we use Gini importances [23]. For Logistic Regression and Support Vector Machine, we consider the absolute values of feature coefficients as the measure of feature importance. Besides, we also generate some decision rules from Decision tree for better understanding of the analysis results.

With different training samples, we can build different models on different classifiers. The comparison of evaluation results, identified features and patterns among different models on different classifiers can also provide useful insights on congestion correlation between road segments.

5. Experiments

In this section, we present the details of datasets, experiment settings and results.

5.1. Datasets

In the experiments, we use three datasets, i.e., road network, POIs, and the GPS trajectories of taxis. All these three datasets are for Beijing, China and their details are elaborated as below.

The road network data is extracted from OpenStreetMap (OSM),[1] an open source online map. In Beijing road network, we have 109, 029 edges and 105, 030 nodes, with 13 categories of road types.

POI dataset contains all kinds of physical objects in spatial space such as shops, schools, banks, and restaurants. Though we can also download POIs from OSM, the number of POIs there is quite small. To collect enough POIs, we obtain the POI data from a data sharing web site called DataTang.[2] This POI dataset is comprised of 220, 137 POIs, which is divided into 21 categories.

We collect a large set of GPS trajectories of over 10, 000 taxis in Beijing for 30 consecutive days in 2012.

1 https://www.openstreetmap.org.
2 http://www.datatang.com/.

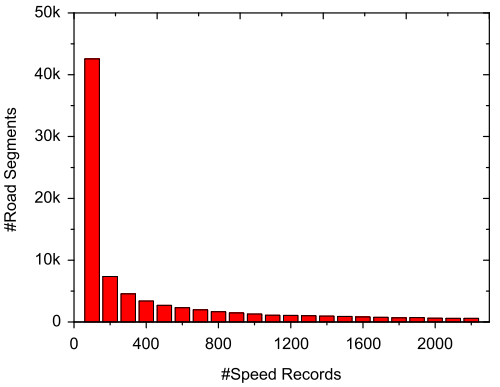
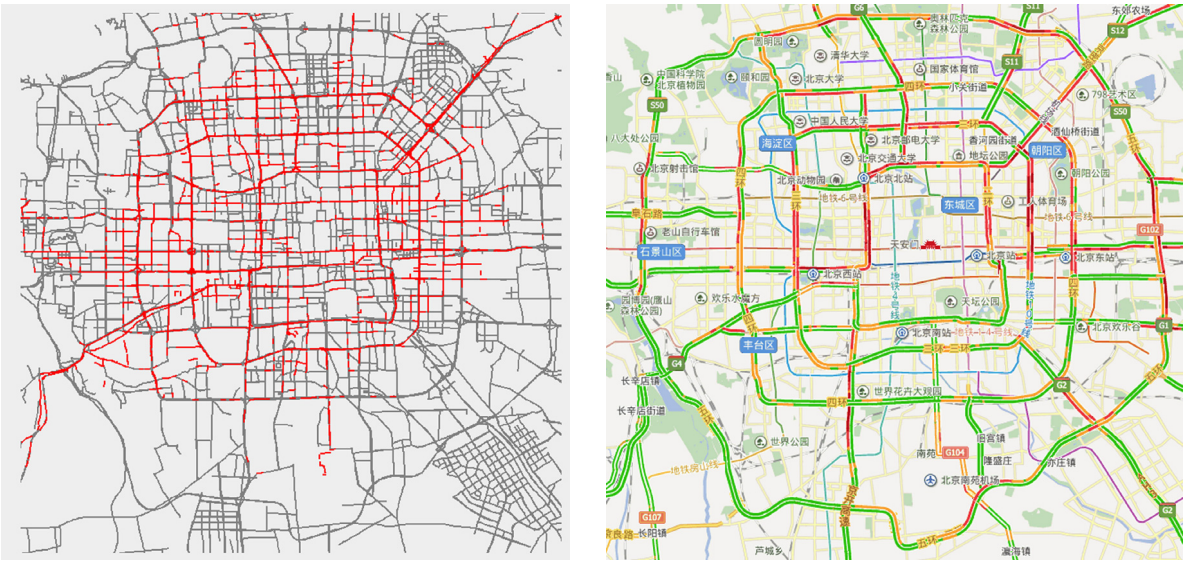


Fig. 4. The distribution of the number of speed records on each road segment per day.



(a) Remaining roads (red).

(b) Real traffic in Beijing at 6 pm.

Fig. 5. The remaining road segments after filtering and the real traffic in Beijing at 6 pm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2. Data filtering

After map-matching the GPS trajectories to the underlying road network, each road segment has a set of GPS records with time stamps. Considering that the goal of this work is to explore the congestion correlation between road segments, it is very important to obtain the traffic information on road segments as accurate as possible.

According to [19], more than 12% of traffic flow in Beijing is occupied by taxi trips. Therefore, it is reasonable for us to use the speeds of GPS records of taxis to approximate the real traffic congestion information. However, though we have over 10,000 taxis, the number of speed samples of some road segments are still very small, which makes it difficult to capture the real traffic on these road segments. In the experiments, we divide a day into 10 min time slots, resulting in 144 time slots one day.

As shown in Fig. 4, many road segments have speed records less than 100 per day, meaning that there is no traffic information in some time slots for many road segments. To alleviate the impact of data sparsity, we remove those segments that have less than 500 speed samples in a whole day. Finally, we get 3004 road segments which have enough traffic information to support our further analysis. The remaining roads are plotted with red color in Fig. 5(a). Fig. 5(b) illustrates the real traffic[3] in Beijing at 6 pm, where red color represents busy traffic. Obviously, the remaining roads in Fig. 5(a) cover most of the roads that have busy traffic in Fig. 5(b). Therefore, it is reasonable for us to conduct analysis on remaining roads since our goal is to explore the congestion correlation between road segments.

[3] http://map.baidu.com.

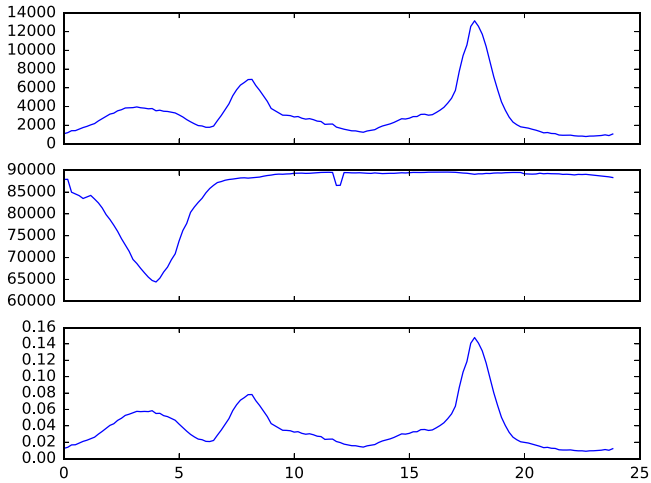


Fig. 6. The number of congested roads, the number of roads with GPS records, and the proportion of congested roads.

5.3. Settings

In the experiments, we set the ratio in Definition 5 to 0.5, which is similar to [7], and compute the average speed on a road segment by all GPS records on the segment over 30 days.

As illustrated in Fig. 6, there are three sub-figures representing respectively the number of congested roads, the number of roads with GPS records and the proportion of congested roads from 0:00 to 23:59 over 30 days. We can see two peaks of the number of congested roads and the proportion, which corresponds to morning peak and evening peak in a day. Besides, during late night, the number of roads with GPS records dramatically falls, which is probably because there are much fewer taxis traveling during this period. Since our goal is to explore congestion correlation between road segments, to ensure accurate traffic information extraction and enough congested roads for analysis, we focus on morning peak and evening peak. Specifically, we generate two sets of training samples from these two peaks in 30 days, respectively. The morning peak is from 7:30 to 9:00, and the evening peak is from 17:30 to 19:00.

Recall Definition 6, in the experiments, we set the time threshold $t = 2$, which is 20 min; $d = 5$ km, since the average speed of all GPS records in congested roads is about 16 km/h, and in 20 min the congestion can propagate at most around 5 km, thus reducing the false congestion correlation to some extent.

For the two sets of training samples, we set the threshold of correlation confidence for positive sample to 0.6, and the threshold for negative sample to 0.4. In the morning peak samples, 33,909 positive samples are collected, and 38,6875 negative samples are collected. After RUS, a balanced morning peak samples are generated with a total of 67,915 samples. In the evening peak samples, 53,968 positive samples are collected, and 495,808 negative samples are collected. After RUS, a balanced evening peak samples are generated with a total of 108,435 samples. For each sample, we initially generate 618 features as described in Table 2. Then we discard Diff-Len and Diff-Degree, since they hardly contribute to the performance of models during the experiments, and end up with 616 features for each sample.

We input the two sets of training samples with selected features, and train the two peak models on four well-known classifiers: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) [24] to predict congestion correlation. Then the average precision and recall computed by 10-fold cross validation are applied to evaluate the performance of the trained models. The precision and recall are defined as follows.

$$\text{Precision} = \frac{\text{No. of predicted true correlations}}{\text{No. of all predicted correlations}} \tag{7}$$

$$\text{Recall} = \frac{\text{No. of predicted true correlations}}{\text{No. of all true correlations}}. \tag{8}$$

We summarize the experiment settings as shown in Table 3.

5.4. Results and analysis

5.4.1. Congestion correlation and transmissibility

After the data filtering, we show and analyze the congestion correlation heatmap of road segments here. As shown in Fig. 7, the axes represent the numbers of road segments. Congestion correlation ranges from 0 to 1, and the denser the color, the higher congestion correlation are between two road segments. We can see that points with dense color are sparse overall, but we can also notice a clear diagonal line in the heatmap, as well as some vertical and horizontal bars.

Table 3
Experimental settings.

Ratio	0.5
t	20 min
d	5 km
Positive correlation confidence threshold	0.6
Negative correlation confidence threshold	0.4
Number of features	616
Morning peak	679,151 samples
Evening peak	108,435 samples
Classifiers	DT, RF, LR, SVM

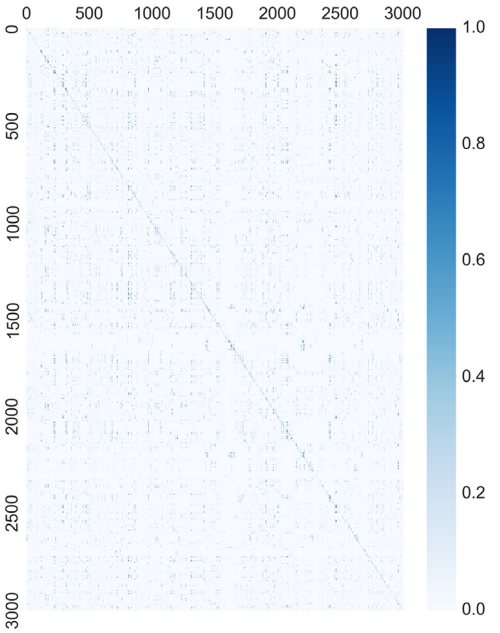


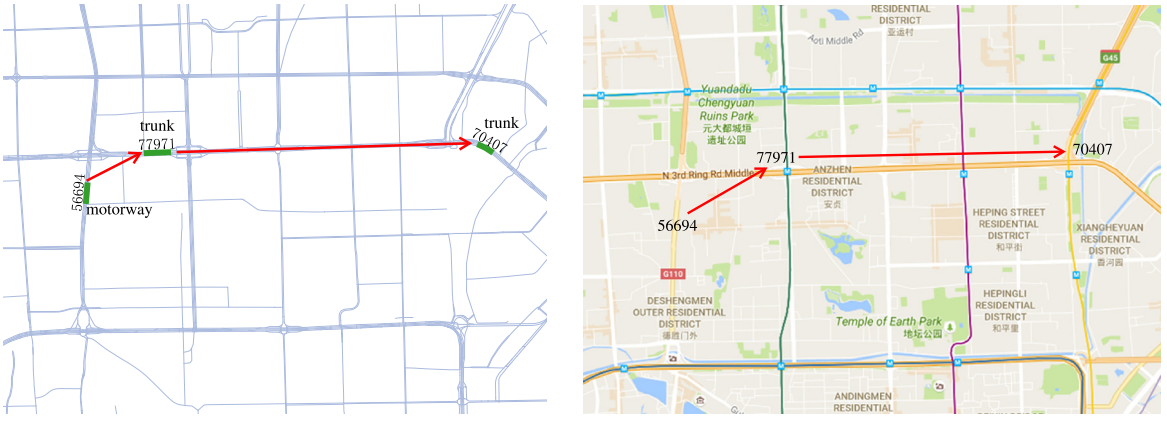
Fig. 7. Congestion correlation heatmap of roads. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The sparse points indicate that congestion correlations are normally not high, which may be due to two reasons. First, the congestions themselves only occur on some road segments, which make it hard to find correlations among most road segments. Second, the number of road segment pairs that can be considered as having high congestion correlations is quite small. Another observation of the diagonal line indicates that congestions on a road segment usually lasts longer than one time slot i.e. 10 min. Besides, the bars on the heatmap indicate that congestions on some road segments are more likely to propagate to multiple other road segments, which may be because they are hubs in the road network.

As discussed above, some traffic congestions are highly correlated. Moreover, we found that these correlations have some kind of transmissibility. In other words, one traffic congestion may leads to a series of congestions at different locations. For example, the congestion of road segment A results in the congestion of road segment B, and road segment B further leads to the congestion of road segment C. In some cases, such congestion transmission can cover a quite wide area.

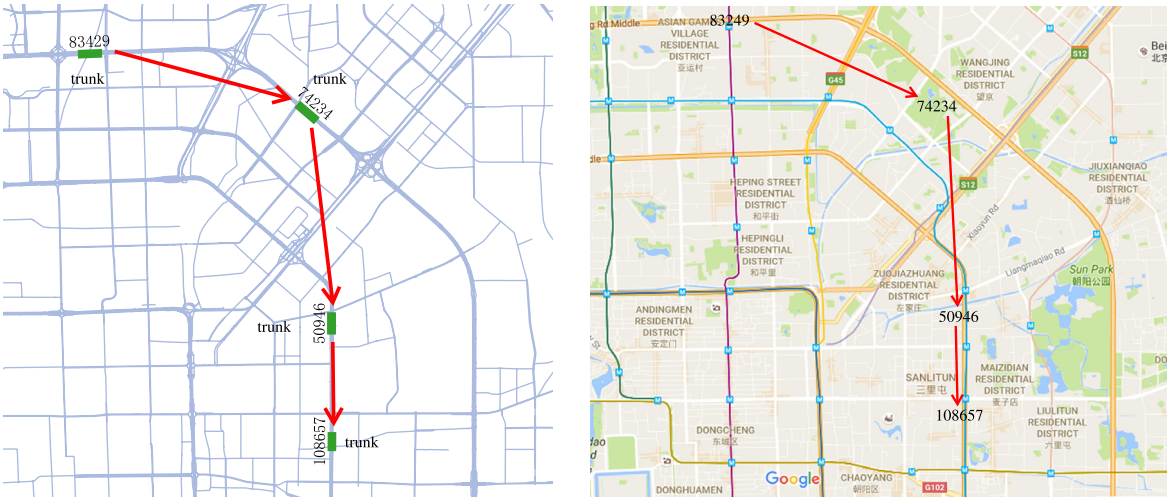
As illustrated in Fig. 8, traffic congestions propagate among road segments 56 694, 77 971 and 70 407 in the morning, where the numbers are road segment identifiers. According to Fig. 8, the traffic congestion at road segment 56 694 (an important motor way between the urban area and the suburbs of Beijing) will affect the traffic at road segment 77 971 (a trunk road on the third ring road next to a residential district). The congestion at road segment 77 971 will then results in the congestion at road segment 70 407 (a trunk road at the traffic hub of Northeastern Beijing).

Similarly, Fig. 9 shows the traffic congestion transmission among road segments 83 429, 74 234, 50 946 and 108 657 in the evening. All these road segments are trunk roads and the congestion transmits sequentially from road segment 83 429 to road segment 108 657. Particularly, road segments 83 429 and 74 234 are next to the Asian games village residential district and Wang Jing residential district, respectively, while both road segments 50 946 and 108 657 are within the Sanlitun commercial area.



(a) Road network. (b) Real map.

Fig. 8. Case study for the traffic congestion transmission in the morning.



(a) Road network. (b) Real map.

Fig. 9. Case study for the traffic congestion transmission in the evening.

Table 4
10-fold CV results on different classifiers of two models.

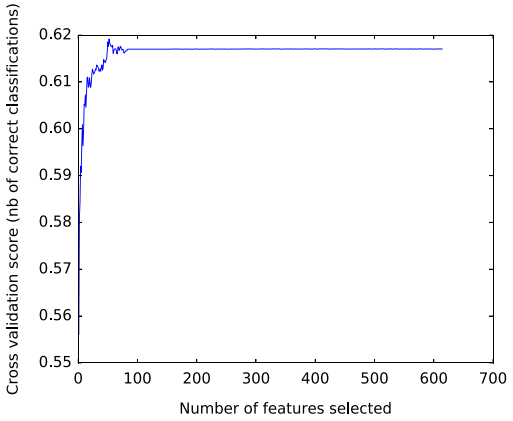
	Morning peak		Evening peak	
Classifiers	Metrics			
	Precision	Recall	Precision	Recall
Decision Tree	0.615(0.012)	**0.598**(**0.033**)	0.661(0.014)	0.642(0.053)
Random Forest	**0.693**(**0.020**)	0.550(0.029)	**0.742**(**0.013**)	0.627(0.031)
Logistic Regression	0.626(0.017)	0.559(0.048)	0.682(0.012)	**0.665**(**0.030**)
SVM	0.639(0.027)	0.446(0.055)	0.692(0.011)	0.633(0.032)

5.4.2. Model performance and feature selection

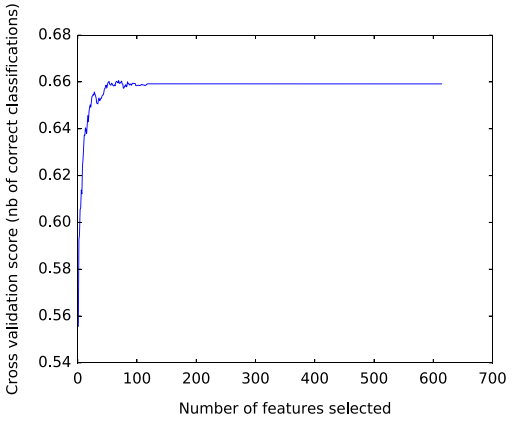
We evaluate the trained models using average precision and recall. The 10-fold cross validation results are shown in Table 4, where the number in the bracket is the standard deviation. Generally, the results are stable with satisfactory precision and recall, considering that we have not conducted a very fine parameter tuning for the best performance.

In terms of the two peak models, the evening peak models achieve better performance in both precision and recall than the morning peak models. In terms of precision, models trained on Random Forest achieve the best performance in both morning and evening peaks. In terms of recall, models trained on Decision Tree and Logistic Regression achieve the best performance, respectively in morning peak and evening peak.

Now we further perform Recursive Feature Elimination on the features. The basic idea is to iterate over all combinations of features on a designated model to find the optimal number of features with best performance. Here, we performed 3-fold

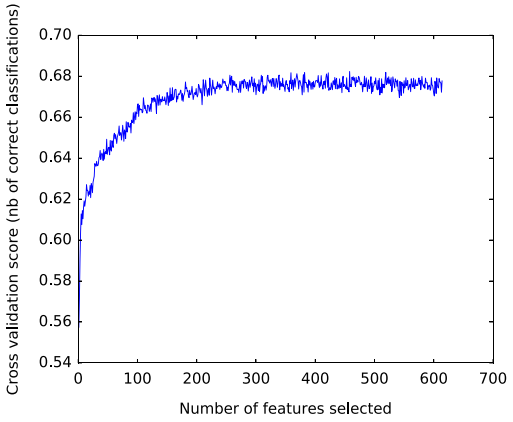


(a) Feature selection of Decision Tree in the morning peak.

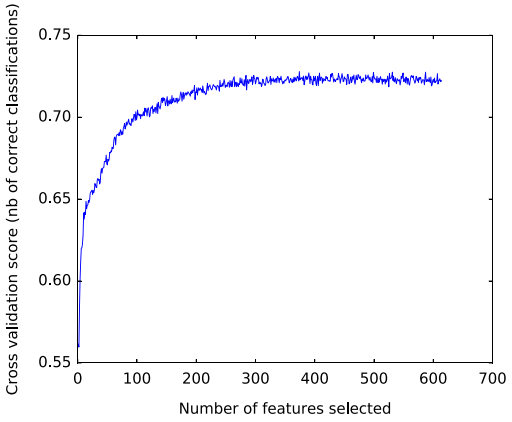


(b) Feature selection of Decision Tree in the evening peak.

Fig. 10. Feature selection of Decision Tree.



(a) Feature selection of Random Forest in the morning peak.



(b) Feature selection of Random Forest in the evening peak.

Fig. 11. Feature selection of Random Forest.

cross validation to measure the overall precision of Decision Tree and Random Forest with different number of features selected. As shown in Fig. 10, with the increasing number of selected feature, the overall precisions of Decision Tree keep increasing, though oscillating from time to time, until reach the optimal, and then go down a little bit and become stable; while in Fig. 11, the overall precisions of Random Forest, though show similar trends, keep increasing and oscillating a little slower, and reach the optimal on a larger number of features with higher value. The optimal number of features for Decision Tree is much smaller than that of Random Forest, and the precision is also lower. This may be because that Random Forest is capable of extracting more useful information from large feature sets for prediction, and making better prediction.

We also divide the feature sets into two categories. One includes features extracted from road network, the other includes features extracted from POI. We are trying to see how features from these two different categories contribute to the final prediction. As shown in Table 5, classifiers using Road network features achieve higher precision than using POI features, and again Random Forest achieve best precision overall. Besides, for some classifiers e.g. Decision Tree, solely using road network features achieve higher precision than using both two categories of features, while for others e.g. Random Forest, using both achieve best precision. The reason may be some classifiers are not able to extract useful features from a large number of features, when introducing more features, more noise are also introduced, resulting to a lower precision. While for others, they are able to extract useful information for prediction from a large number of features even with more noise, resulting to a higher precision. This is somehow in line with our analysis on optimal number of features previously.

5.4.3. Importance features and generated rules

We also compare the top 10 important features identified by two models on different classifiers, and list the commonly identified important features on two models in Table 6. In addition, Table 7 shows some rules generated by Decision Tree on the two models (note that all the features have been scaled into the range of [0, 1] as described in Section 4).

As we can see, Diff-B, Diff-C, Diff-POI, *netDis*, and 'trunk → trunk' are both commonly identified important features in the two models, meaning that they are important to predict whether a road segment is correlated with another one in terms

Table 5
10-fold CV results on different classifiers of two models based on two categories of features.

Classifiers	Morning peak		Evening peak	
	Features			
	RoadNetwork	POI	RoadNetwork	POI
Decision Tree	0.622(0.019)	0.586(0.010)	0.676(0.016)	0.591(0.010)
Random Forest	0.646(0.018)	0.613(0.017)	0.700(0.012)	0.623(0.016)
Logistic Regression	0.631(0.023)	0.556(0.020)	0.680(0.012)	0.550(0.011)
SVM	0.635(0.022)	0.592(0.016)	0.687(0.013)	0.602(0.015)

Table 6
Commonly identified important features.

	Features	Description
Morning peak	Diff-B	The difference of betweenness
	Diff-C	The difference of closeness
	Diff-POI	The difference of the total number of POIs
	SimPOIs	Pearson similarity of POI TF-IDF value distributions
	netDis	The network distance
	'trunk → trunk'	Binary indicator variable for the ordered combination 'trunk → trunk' of road types
	'motorway → motorway'	Binary indicator variable for the ordered combination 'motorway → motorway' of road types
	'catering → catering'	Variable for the ordered combination 'catering → catering' of POI categories
Evening peak	Diff-B	The difference of betweenness
	Diff-C	The difference of closeness
	Diff-POI	The difference of the total number of POIs
	netDis	The network distance
	'trunk → trunk'	Binary indicator variable for the ordered combination 'trunk → trunk' of road types
	'trunk → secondary'	Binary indicator variable for the ordered combination 'trunk → secondary' of road types
	'tertiary → secondary'	Binary indicator variable for the ordered combination 'tertiary → secondary' of road types

Table 7
Generated rules.

	Rules
Morning peak	If $0.4184 <$ Diff-POI ≤ 0.4454 and Diff-B > 0.4755 and Diff-C ≤ 0.4906, then uncorrelated If Diff-POI > 0.4947 and *netDis* ≤ 0.294 and 'motorway → motorway' $= 1$, then correlated
Evening peak	If Diff-POI ≤ 0.49 and Diff-B > 0.4938 and 'tertiary → secondary' $= 1$, then uncorrelated If $0.0038 <$ *netDis* ≤ 0.0877 and 'trunk → trunk' $= 1$, then correlated

of congestion in both morning and evening peaks. On the other hand, 'motorway → motorway' and 'catering → catering' are more important in the morning peak, and 'trunk → secondary' and 'tertiary → secondary' are more important in the evening peak. The results reveal the common and different patterns between morning and evening peaks.

From the generated rules, we can observe more different patterns in the morning and evening peaks. For example, in the morning peak, there exits high congestion correlation from one motorway to another if the POI numbers of them are quite different, meaning that congestions are more likely to propagate from a motorway with more POIs to another one with less POIs in the morning peak. On the other hand, there exits high congestion correlation from one trunk road to another in the evening peak, meaning that congestions are more likely to propagate between trunk roads in the evening peak.

6. Conclusion

In this paper, we outline a three-phase framework to explore the congestion correlation between road segments from multiple data sources. We first obtain congestion information on road segments from GPS data, give the definition of congestion correlation and design the mining algorithm. Then we extract topological and POI features on each road segment, and fuse them to generate the features of training samples for each pair of road segments. Finally, the congestion correlation and features on each pair of road segments are input to well-known classifiers including Decision Tree, Random Forest, Logistic Regression and Support Vector Machine. We train two models on different classifiers to predict congestion correlation, compare and analyze the performance and important features. The experiment results show stable and satisfactory performance as well as some important patterns of congestion correlation. In addition, we discuss the patterns of congestion correlations, and found obvious directionality and transmissibility. Meanwhile, we also analyze the impact of feature selection on the performance of models. Notably, the proposed framework is general and can be applied to other pairwise correlation analysis.

Acknowledgments

The work described in this paper was partially supported by the funding for Project of Strategic Importance provided by The Hong Kong Polytechnic University (Project Code: 1-ZE26), HK RGC under GRF Grant PolyU 5104/13E and NSFC key grant with Project No. 61332004.

References

[1] Y. Ando, O. Masutani, H. Sasaki, H. Iwasaki, Y. Fukazawa, S. Honiden, Pheromone model: Application to traffic congestion prediction, in: Engineering Self-Organising Systems, Springer, 2006, pp. 182–196.
[2] E. Jenelius, H.N. Koutsopoulos, Travel time estimation for urban road networks using low frequency probe vehicle data, Transp. Res. B 53 (2013) 64–81.
[3] K. Hymel, Does traffic congestion reduce employment growth? J. Urban Econ. 65 (2) (2009) 127–135.
[4] L.S. Nookala, Weather impact on traffic conditions and travel time prediction (Ph.D. thesis), University of Minnesota Duluth, 2006.
[5] J. Long, Z. Gao, H. Ren, A. Lian, Urban traffic congestion propagation and bottleneck identification, Sci. China Ser. F 51 (7) (2008) 948–964.
[6] P. Rachtan, H. Huang, S. Gao, Spatio-temporal link speed correlations: An empirical study, Transp. Res. Rec. 2390 (2013) 34–43.
[7] S. Yang, On feature selection for traffic congestion prediction, Transp. Res. C 26 (2013) 160–169.
[8] W. Min, L. Wynter, Real-time road traffic prediction with spatio-temporal correlations, Transp. Res. C 19 (4) (2011) 606–616.
[9] B.J. Gajewski, L.R. Rilett, Estimating link travel time correlation: an application of Bayesian smoothing splines, J. Transp. Stat. 7 (2/3) (2004) 53–70.
[10] T. Nagatani, Propagation of jams in congested traffic flow, J. Phys. Soc. Japan 65 (7) (1996) 2333–2336.
[11] M.A. Joshi, D. Mishra, Review of traffic density analysis techniques, Image 4 (7) (2015).
[12] P.P. Dubey, P. Borkar, Review on techniques for traffic jam detection and congestion avoidance, in: 2015 2nd International Conference on Electronics and Communication Systems, ICECS, IEEE, 2015, pp. 434–440.
[13] C. Wang, M.A. Quddus, S.G. Ison, Impact of traffic congestion on road accidents: a spatial analysis of the m25 motorway in england, Accid. Anal. Prev. 41 (4) (2009) 798–808.
[14] J. Kononov, B. Bailey, B. Allery, Relationships between safety and both congestion and number of lanes on urban freeways, Transp. Res. Rec.: J. Transp. Res. Board (2083) (2008) 26–39.
[15] J. Yuan, Y. Zheng, C. Zhang, X. Xie, G. Sun, An interactive-voting based map matching algorithm, in: Mobile Data Management, 2010, pp. 43–52.
[16] N. Beckmann, H. Kriegel, R. Schneider, B. Seeger, The r*-tree: An efficient and robust access method for points and rectangles, in: Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, May 23–25, 1990, 1990, pp. 322–331.
[17] P.-N. Tan, V. Kumar, (Chapter 6). Association analysis: Basic concepts and algorithms, Introduction to Data Mining. Addison-Wesley. ISBN 321321367.
[18] P. Crucitti, V. Latora, S. Porta, Centrality measures in spatial networks of urban streets, Phys. Rev. E 73 (3) (2006) 35–39.
[19] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and pois, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 186–194.
[20] R. Geisberger, P. Sanders, D. Schultes, D. Delling, Contraction hierarchies: Faster and simpler hierarchical routing in road networks, in: WEA, 2008, pp. 319–333.
[21] L. Al Shalabi, Z. Shaaban, B. Kasasbeh, Data mining: A preprocessing engine, J. Comput. Sci. 2 (9) (2006) 735–739.
[22] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 935–942.
[23] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.