# Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories

Yuqi Wang*, Jiannong Cao*, Wengen Li* and Tao Gu†
*Department of Computing, Hong Kong Polytechnic University
{csyqwang, csjcao, cswgli}@comp.polyu.edu.hk
†School of Computer Science and IT, RMIT University
tao.gu@rmit.edu.au

***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 study 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. We found some important patterns that lead to a high/low congestion correlation, and they can facilitate building various transportation applications. The proposed techniques in our framework are general, and can be applied to other pairwise correlation analysis.**

***Index Terms*—Traffic congestion; Congestion correlation; GPS trajectories; Classification;**

## I. 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, traffic flow propagation [5], etc. 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 to 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, how congestions propagate from one road to another, etc. Furthermore, it can also facilitate building various applications including road planning, traffic condition prediction, impact analysis of congestion, etc. As such, both governments and individuals can be beneficial. 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 travelling 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 and 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 study traffic congestion correlation from a classification perspective using real world datasets. Our contributions are three folds:

- We propose a novel framework to study 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 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.
- 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 related work. Section 3 gives an overview

978-1-5090-0898-8/16/$31.00 ©2016 IEEE

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.

## II. 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 travelling time correlation, and found that the heavier the congestion, the lower the correlation of travelling 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 travelling 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 on GPS trajectories, which can benefit various applications including traffic prediction, traffic light management, road planning, etc.

## III. Overview

Figure 1 presents the framework of our work. In this framework, we utilize GPS trajectory of taxis, road network and POI data to study the congestion correlation between road segments. We divide the framework into three phases.

1) Extract congestion information on each road segment from GPS and road network data, define and mine congestion correlation between each road segment pair.
2) Extract various topological features and POI features from road network and POI data, respectively, and generate training samples on road segment pairs.
3) 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, correlation definition, and etc.

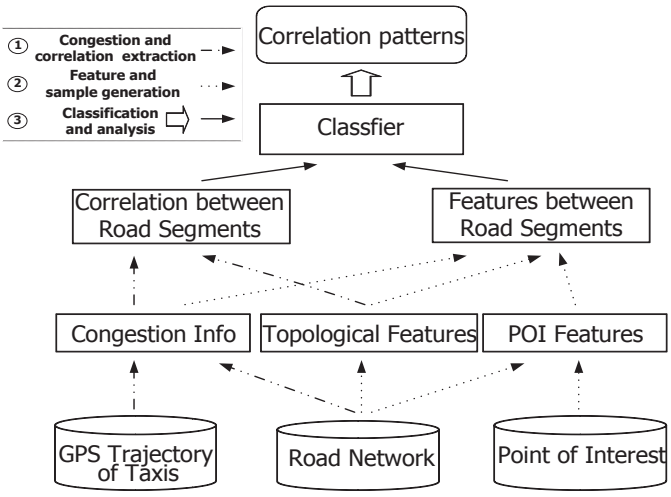



Fig. 1. Framework of congestion correlation mining

## IV. 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.

### A. Data sources

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

*Definition 1 (Road network):* A road network is modelled as a graph $G = (V, E)$, where $v_i \in V$ represents an intersection of road segments, $e_{i,j} \in E$ 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 $o_i$.

### B. 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.

*1) Traffic information acquisition:* 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.

*2) Congestion extraction:* To extract congestion information from traffic information, we divide a day into time slots, and obtain the traffic information $T_{rt}$ 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 $C_{rt}$, and

$$C_{rt} = \begin{cases} 1 & \text{if } T_{rt} \leq T_r * Ratio; \\ 0 & \text{otherwise.} \end{cases}$$

where $T_r$ is the average speed of all GPS points in road segment $r$ in all time, and $Ratio$ will be discussed in Section V.

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

$$\begin{array}{c} \text{Time Slot} \\ \text{Road ID} \end{array} \begin{bmatrix} 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 1 \end{bmatrix}$$

Fig. 2. Congestion matrix

*3) Correlation extraction:* To study how congestion occurs sequentially in terms of time, and consider the propagation rate of congestion in terms of space, as shown in figure 3, we define congestion correlation between road segments as follows.

*Definition 6 (Congestion correlation between road segments):* A congestion correlation from road segment $a$ to 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 $d$

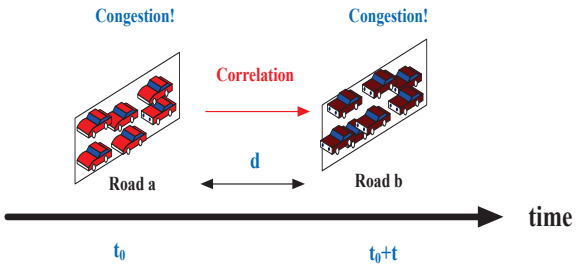



Fig. 3. Congestion correlation

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 segment $i$ and $k$ from $t_{start}$ to $t_{end}$.

**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;
```

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 $R_{i.}$. 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 distance in all pairs of road segments to make sure the distance requirement in congestion correlation is also satisfied. The time complexity of the proposed algorithm is O($n^2$m), 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.

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

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

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

### C. 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 for each road segment pair to generate training samples.

*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].

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

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).

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

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.

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

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.

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

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, 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 $i$-th POI category of a given road segment, we have the following formula:

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

where $n_i$ is the number of POIs in $i$-th category and $N$ is the total number of POIs in 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 $i$-th category.

The extracted features are summarized in Table I.

TABLE I
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 |

*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 difference between segments, and then add them to the features for each pair. We also add network distance and Pearson similarity of POI TF-IDF value distributions between road segments into features for each pair.

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

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

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 the importance level of it 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 II.

TABLE II
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 |

### D. 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.

*1) Classification:* After finishing the first two phases, we have all the congestion correlation between road segments, and all features on each pair of road segments. We now combine these two parts to generate training samples 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), then evaluate the performance of the built models using classic metrics.

*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. 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.

## V. EXPERIMENTS

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

### A. 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 data set 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 data set is comprised of 220, 137 POIs which cover 21 categories.

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

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

### B. Data filtering

From the view of road segments, each road segment has a set of GPS records with time stamps. The goal of this work is to study the congestion correlation between road segments. Therefore, it is very important to obtain the traffic information on road segments as accurate as possible.

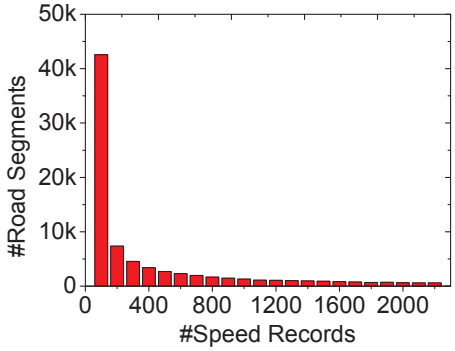



Fig. 4. The distribution of the number of speed records on each road segment per day

According to [19], more than 12 percent of traffic flow in Beijing is occupied by taxi trips, 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 minutes time slots, resulting in 144 time slots one day. According to Figure 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 3,004 road segments which have enough traffic information to support our further analysis. The remaining roads are plotted with red color in Figure 5(a). Figure 5(b) illustrates the real traffic in Beijing at 6pm, where red color represents busy traffic. Obviously, the remaining roads in Figure 5(a) cover most of the roads that have busy traffic in Figure 5(b). Therefore, it is reasonable for us to conduct analysis on remaining roads since our goal is to study the congestion correlation between road segments.

### C. 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 Figure 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 congested roads and the proportion, which corresponds to morning peak and evening peak in a day.

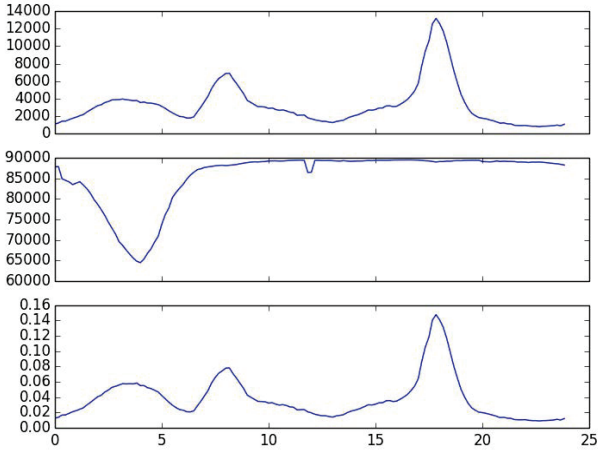


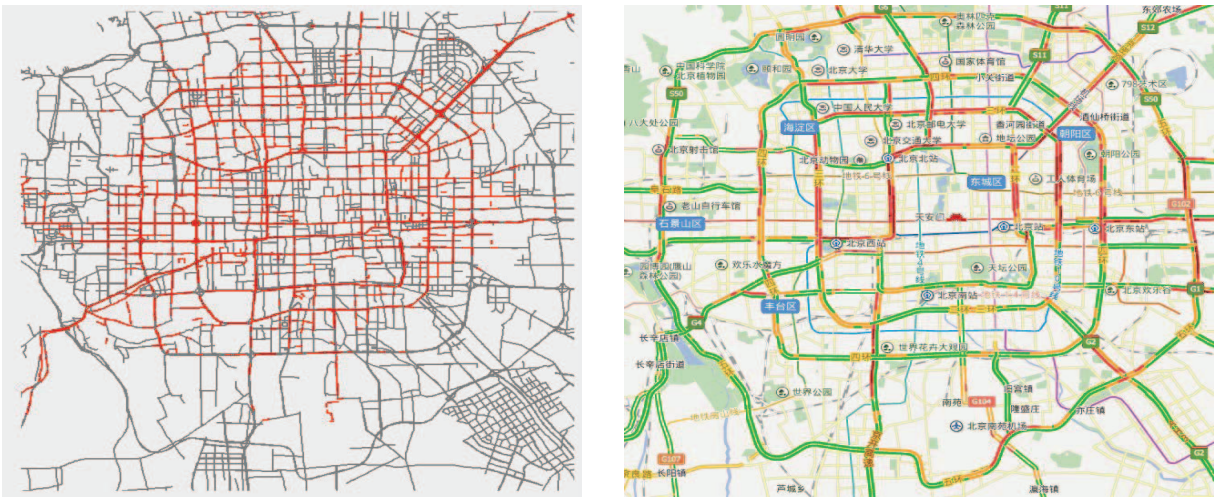

Fig. 6. The number of congested roads, the number of roads with GPS records, and the proportion of congested roads

Besides, during late night, the number of roads with GPS records dramatically falls, which is probably because there are much fewer taxis travelling during this period. Since our goal is to study 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 7:30-9:00, and the evening peak is from 17:30 - 19:00.

Recall Definition 6, in the experiments, we set $t$ = 2, which is 20 minutes; $d$ = 5 km, since the average speed of all GPS records in congested roads is about 16 km/h, and in 20 minutes 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, 33909 positive samples are collected, and 386875 negative samples are collected. After RUS, a balanced morning peak samples are generated with a total of 67915 samples. In the evening peak samples, 53968 positive samples are collected, and 495808 negative samples are collected. After RUS, a balanced evening peak samples are generated with a total of 108435 samples. For each sample, we initially generate 618 features as described in Table II. 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 and train the two peak models on four well-known classifers: 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.



(a) Remaining roads (red) (b) Real traffic in Beijing at 6pm

Fig. 5. The remaining road segments after filtering and the real traffic in Beijing at 6pm.

TABLE III
10-FOLD CV RESULTS ON DIFFERENT CLASSIFIERS OF TWO MODELS

| Metircs / Classfiers | Morning Peak | | Evening Peak | |
|---|---|---|---|---|
| | 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)** |
| Support Vector Machine | 0.639(0.027) | 0.446(0.055) | 0.692(0.011) | 0.633(0.032) |

TABLE IV
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 |

### *D. Results and analysis*

We evaluate the trained models using average precision and recall. The 10-fold cross validation results are shown in Table III, 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.

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 IV. In addition, Table V 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 IV).

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 of congestion in both morning and evening peaks. On the other hand, 'motorway→motorway' and 'catering→catering' are more important in the morning

TABLE V
GENERATED RULES

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

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.

## VI. CONCLUSION

In this paper, we outline a three-phase framework to study the congestion correlation between road segments from multiple sources of data. 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. Notably, the proposed framework is general and can be applied to other pairwise correlation analysis.

## VII. ACKNOWLEDGEMENTS

This work is financially supported in part by HK PolyU Project of Strategic Importance (1-ZE26) and HK RGC under GRF Grant 510413.

## REFERENCES

[1] Y. Ando, O. Masutani, H. Sasaki, H. Iwasaki, Y. Fukazawa, and S. Honiden, "Pheromone model: Application to traffic congestion prediction," in *Engineering Self-Organising Systems*, pp. 182–196, Springer, 2006.

[2] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research Part B: Methodological*, vol. 53, pp. 64–81, 2013.

[3] K. Hymel, "Does traffic congestion reduce employment growth?," *Journal of Urban Economics*, vol. 65, no. 2, pp. 127–135, 2009.

[4] L. S. Nookala, *Weather impact on traffic conditions and travel time prediction*. PhD thesis, University of Minnesota Duluth, 2006.

[5] J. Long, Z. Gao, H. Ren, and A. Lian, "Urban traffic congestion propagation and bottleneck identification," *Science in China Series F: Information Sciences*, vol. 51, no. 7, pp. 948–964, 2008.

[6] P. Rachtan, H. Huang, and S. Gao, "Spatio-temporal link speed correlations: An empirical study," *Transportation Research Record*, vol. 2390, pp. 34–43, 2013.

[7] S. Yang, "On feature selection for traffic congestion prediction," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 160–169, 2013.

[8] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[9] B. J. Gajewski and L. R. Rilett, "Estimating link travel time correlation: an application of bayesian smoothing splines," *Journal of Transportation and Statistics*, vol. 7, no. 2/3, pp. 53–70, 2004.

[10] T. Nagatani, "Propagation of jams in congested traffic flow," *Journal of the Physical Society of Japan*, vol. 65, no. 7, pp. 2333–2336, 1996.

[11] M. A. Joshi and D. Mishra, "Review of traffic density analysis techniques," *Image*, vol. 4, no. 7, 2015.

[12] P. P. Dubey and P. Borkar, "Review on techniques for traffic jam detection and congestion avoidance," in *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pp. 434–440, IEEE, 2015.

[13] C. Wang, M. A. Quddus, and S. G. Ison, "Impact of traffic congestion on road accidents: a spatial analysis of the m25 motorway in england," *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 798–808, 2009.

[14] J. Kononov, B. Bailey, and B. Allery, "Relationships between safety and both congestion and number of lanes on urban freeways," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2083, pp. 26–39, 2008.

[15] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G. Sun, "An interactive-voting based map matching algorithm," in *Mobile Data Management*, pp. 43–52, 2010.

[16] N. Beckmann, H. Kriegel, R. Schneider, and 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.*, pp. 322–331, 1990.

[17] P.-N. Tan and V. Kumar, "Chapter 6. association analysis: Basic concepts and algorithms," *Introduction to Data Mining. Addison-Wesley. ISBN*, vol. 321321367, 2005.

[18] P. Crucitti, V. Latora, and S. Porta, "Centrality measures in spatial networks of urban streets," *Physical Review E*, vol. 73, no. 3.

[19] J. Yuan, Y. Zheng, and 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*, pp. 186–194, ACM, 2012.

[20] R. Geisberger, P. Sanders, D. Schultes, and D. Delling, "Contraction hierarchies: Faster and simpler hierarchical routing in road networks," in *WEA*, pp. 319–333, 2008.

[21] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "Data mining: A preprocessing engine," *Journal of Computer Science*, vol. 2, no. 9, pp. 735–739, 2006.

[22] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th international conference on Machine learning*, pp. 935–942, ACM, 2007.

[23] L. Breiman, J. Friedman, C. J. Stone, and 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," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.