Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 809582, 13 pages
http://dx.doi.org/10.1155/2015/809582

Research Article

Detecting Traffic Anomalies in Urban Areas Using Taxi GPS Data

Weiming Kuang, Shi An, and Huifu Jiang

School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China

Correspondence should be addressed to Huifu Jiang; jianghuifu1987@outlook.com

Received 21 November 2014; Revised 26 January 2015; Accepted 22 April 2015

Academic Editor: Chi-Chun Lo

Large-scale GPS data contain hidden information and provide us with the opportunity to discover knowledge that may be useful for transportation systems using advanced data mining techniques. In major metropolitan cities, many taxicabs are equipped with GPS devices. Because taxies operate continuously for nearly 24 hours per day, they can be used as reliable sensors for the perceived traffic state. In this paper, the entire city was divided into subregions by roads, and taxi GPS data were transformed into traffic flow data to build a traffic flow matrix. In addition, a highly efficient anomaly detection method was proposed based on wavelet transform and PCA (principal component analysis) for detecting anomalous traffic events in urban regions. The traffic anomaly is considered to occur in a subregion when the values of the corresponding indicators deviate significantly from the expected values. This method was evaluated using a GPS dataset that was generated by more than 15,000 taxies over a period of half a year in Harbin, China. The results show that this detection method is effective and efficient.

1. Introduction

Traffic anomalies widely exist in urban traffic networks and negatively effect traffic efficiency, travel time, and air pollution [1]. The traffic flow in a road network is abnormal when traffic accidents, traffic congestion, and large gatherings and events, such as construction, occur [2]. Thus, the detection of traffic anomalies is important for traffic management and has become important in transportation research [3]. Fortunately, most taxies in cities in China are equipped with GPS devices [2]. Because taxies can use road networks widely over long periods, their trajectories can reflect the traffic condition in the road network [4]. In other words, taxies can be observed as "flowing detectors" in the urban road network. Thus, the difficulty of collecting data is reduced so that people can improve the detection of anomalies with a large volume of data.

Several data mining methods have been proposed to achieve the goal of detecting anomalies by using GPS data. Most previous studies can be divided into two categories: (1) studies on taxi GPS trajectory anomalies and (2) studies on traffic anomalies. In the first category, most studies focus on how to observe a small number of drivers with travelling trajectories that are different from the popular choices of other drivers [5]. Some of these studies can be used to detect fraudulent taxi driving behavior to monitor the behavior of taxi drivers [6–8]. Others have paid more attention to hijacked taxi driving behavior, which can protect taxi drivers and passengers from assaultive injury [9]. With the development of vehicle navigation technology, new interest in trajectory anomaly research has occurred, which can be integrated with navigation to provide dynamic routes for drivers or travelers [10–13]. In addition, this research can provide accurate real-time advisor routes compared with navigation based on static traffic information. The purpose of the second category is different from the above studies. In the second category, detection algorithms and optimization methods have been used to detect anomalies and piece them together to explore the root causes of anomalies [14, 15]. In addition, some other methods were proposed for monitoring large-area traffic [16, 17] and determining the defects of existing traffic planning [18]. The differences between these two categories include the following aspects. First, the comparison between trajectories

in the anomalous trajectory process always focuses on a small number of trajectories and the remaining normal trajectories at the same location during a certain period. Second, the detection of traffic anomalies is used to detect a large number of taxies with anomalous behaviors and detect potential events with time.

This research belongs to the traffic anomaly detection; some relevant works are those researching anomaly detection with GPS data [14, 19, 20], and some others use social media data as the source of mobility data to detect anomalies [21, 22]. Most of these methods can be grouped into four categories: distance-based, cluster-based, classification-based, and statistics-based categories [23, 24]. In this paper, the research focuses on taxi GPS data and the detection method can be classified as statistics-based. According to an analysis of the existing literatures, most studies have only considered traffic volume, velocity, and other visualized parameters and have not considered the spatial information hidden in the traffic flow [25]. Moreover, most existing methods are simple methods based on single detection methods [17, 23–25] or modified versions of traditional outlier detection methods [14]. These methods can easily detect long-term anomalies but lose many short-term anomalies which can continue for a short period; thus, the focus of this study is to improve the sensitivity of detection methods. Some methods for detecting anomalies in computer networks or financial time series use the wavelet transform method to improve the performance of detecting rapid anomalous changes [26, 27]. This idea can be introduced into this research to achieve the same goal because the road network is similar to the computer network. Next, a traffic anomalies detection method was proposed, which can be distinguished in two ways. First, this method combines the wavelet transform method and PCA to detect traffic anomalies due to low or high rates of change in traffic flow. Therefore, this method can more effectively detect traffic anomalies than other detection methods that only use PCA [14]. Further, this method can provide information regarding the spatial distribution of traffic flows. The advantage of this method is identifying the roots while detecting the anomalies, which reduces the blindness of traffic guidance.

The organizational structure of this paper is organized as follows. In Section 2, the GPS data transformation and the anomalies detecting method are described in detail. In Section 3, case study is conducted based on taxi GPS data of Harbin and the effectiveness and performance of the proposed method are analyzed at the same time. Finally, in Section 4, the conclusions from this research are summarized.

2. Material and Methods

Traffic anomalies always occur in regions with large traffic volume or high road network densities and deviate due to changes in external conditions when compared with the performance of normal traffic. Many factors can result in traffic anomalies, including traffic accidents, special traffic controls, large gatherings, demonstrations, and natural disasters [1]. These causes may lead to a wide range of traffic changes and further produce anomalous traffic flow patterns. Furthermore, traffic anomaly levels can be serious because of traffic flow propagation.

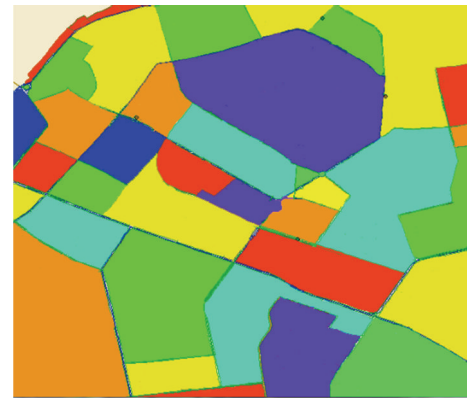


Figure 1: Network-based urban area segmentation.

2.1. Road Network Traffic and Traffic Flow Matrix

2.1.1. Road Network Traffic. In the taxi GPS data, each taxi trajectory consists of a sequence of points with ID number, latitude, longitude, vehicle state (passenger/empty/no-service), and timestamp information. Taxi drivers need to stop their vehicles to pick up or drop off passengers (referred to as a vehicle state transition); thus, each trajectory can be divided into several end-to-end subtrajectories that are defined as "trip" in this paper. Because three types of vehicle state are used, the trips can be considered as "passenger" trips, "empty" trips, and "no-service" trips.

Although three types of vehicle state are used, the "no-service" GPS points will be merged to one point in the map-matching process, which can be ignored in this research. Only two classes of the trips were investigated: one is the "passenger" trip and the other is the "empty" trip. Each trip represents the behavioral characteristics of traveling from an origin point O to a destination point D. However, any two trips will not have the same origin point or destination point (spatial dimension) in real life. Consequently, road network traffic is hidden among different trips, and it is difficult to detect traffic anomalies. Therefore, the transport network was simplified and a novel network traffic model was proposed for in-depth analysis and reducing complexity. Urban areas were segmented into subregions by road networks [28]. As demonstrated in Figure 1, each subregion is surrounded by a certain level of road, and any two adjacent subregions do not overlap in space. This model can provide more natural and semantic segmentation of urban spaces. Next, a traffic model was constructed based on urban segmentation. In this model, the vehicles mobility in the subregion was ignored, and all subregions were abstracted into nodes. The road network was modeled as a directed graph $G = (N, L)$, where N is a set of nodes (subregions) and L is a set of links that connect two adjacent subregions. A link can represent the mobility of

Table 1: Virtual OD nodes pairs.

Origin virtual node	Destination virtual node			
	VN_1	VN_2	VN_3	VN_4
VN_1	(VN_1, VN_1)	(VN_1, VN_2)	(VN_1, VN_3)	(VN_1, VN_4)
VN_2	(VN_2, VN_1)	(VN_2, VN_2)	(VN_2, VN_3)	(VN_2, VN_4)
VN_3	(VN_3, VN_1)	(VN_3, VN_2)	(VN_3, VN_3)	(VN_3, VN_4)
VN_4	(VN_4, VN_1)	(VN_4, VN_2)	(VN_4, VN_3)	(VN_4, VN_4)

vehicles between two adjacent subregions. Meanwhile, "trip" and "path" must be redefined based on this new model.

Definition 1 (trip). A trip, tr, is a time sequence consisting of subregions with timestamp and can be transformed into a time sequence of nodes that can represent subregions in the model (i.e., $\mathrm{tr} : \langle N_1, t_1\rangle \rightarrow \langle N_2, t_2\rangle \rightarrow \cdots \rightarrow \langle N_n, t_n\rangle$).

Definition 2 (path). A path, P, is a sequence of nodes without temporal information (i.e., $\mathrm{tr} : N_1 \rightarrow N_2 \rightarrow \cdots \rightarrow N_n$). A path can represent the common spatial trajectory of some trips that have the same node sequences when the timestamp is ignored.

Definition 3 (trajectory). A trajectory T is a sequence of connected trips (i.e., $T = \mathrm{tr}_1 \rightarrow \mathrm{tr}_2 \rightarrow \cdots \rightarrow \mathrm{tr}_n$), where $\mathrm{tr}_{(k+1)} \cdot s = \mathrm{tr}_k \cdot e$ $(1 \leq k < n)$, $\mathrm{tr}_{(k+1)} \cdot s$ is the start node of $\mathrm{tr}_{(k+1)}$, and $\mathrm{tr}_k \cdot e$ is the end node of tr_k.

This road network traffic model can represent the spatial mobility characteristics of flows from the origin to destination nodes. Thus, they not only flow within different nodes and links in the road network but also tell us how traffic flows from origin nodes to destination nodes. The road network traffic is used to obtain the sizes of the OD traffic flows. All of the traffic in the network will flow from origin nodes and across some different intermediate nodes and links before reaching the destination nodes. This method is useful because all of the network topology information can be expressed, as shown in Figure 2. In the logical topology layer, each node can be observed as an origin/destination node, and the link between two nodes represents the traffic flow from the origin node to the destination node. However, when the logical topology layer is mapped to the physical topology layer, each path of the logical topology layer is divided into several different sequences of links, as defined in Definition 2. This method can help us extract the traffic information from traffic flow data. However, in this research, the aim is not only to detect which OD nodes pairs have anomalous traffic but also to identify which trips between the OD nodes pairs are anomalous. Further, two concepts called "virtual node" and "virtual OD nodes pair" are defined as follows.

Definition 4 (virtual node). Virtual node is an imaginary node. Each node in this road network has at least one virtual node, and the virtual nodes have the same spatial-temporal characteristics, as shown in Figure 2.

Definition 5 (virtual OD nodes pair). The virtual OD nodes pair is composed of virtual nodes, with each virtual OD node pair possessing traffic flow across a unique path. Only the origin/destination nodes of the path can be represented by the virtual node, and the intermediate nodes must be real. Virtual OD node pairs can help us build different paths between the same OD node pairs (i.e., $P = VN_1 \rightarrow N_2 \rightarrow \cdots \rightarrow N_{k-1} \rightarrow VN_k$, $k = 1, 2, \ldots$, where P is a path and VN_1 and VN_k are origin virtual node and destination virtual node, resp.). As shown in Figure 2, there are four virtual OD node pair paths (virtual node 3 $\rightarrow$ virtual node 1). The number of a virtual OD nodes pair is equal to the number of the path that connects the OD nodes.

Next, virtual OD node pairs were built according to the logical topology layer, as shown in Table 1. Based on the information shown in Table 1, one node can connect with multiple nodes and those multiple nodes can have the same destination node. Previously, the network traffic feature was formulated and the traffic model can hold the spatial correlation of traffic flows, the network wide traffic is a time sequence model, and the time and frequency properties of the traffic can be held well. In the next step, a transform domain analysis was conducted for the road network traffic to detect traffic flow anomalies.

2.1.2. Index Building. An index structure was created for anomaly detection process. Each OD node pair can have several paths that can connect the OD nodes (virtual OD nodes). However, the research goal is to determine which paths of the OD node pairs are anomalous. Thus, an index structure was built, which is an offline index structure between the path and links that can connect the nodes/virtual nodes. For example, in Figure 3(a), the points represent the nodes/virtual nodes, the solid directed lines represent the links, and the dashed lines represent the paths between the OD nodes pairs. This index method is offline but can be updated to be online when new data are received, as shown in Figure 3(b).

2.1.3. Traffic Flow Matrix. The traffic anomalies detecting method based on multiscale PCA (MSPCA) in this paper uses the traffic flows matrix as a data source. Thus, the related definitions of the traffic matrix are presented as follows.

Definition 6 (traffic flow matrix). A traffic flow matrix is the traffic demand of all the virtual OD nodes pairs in a road

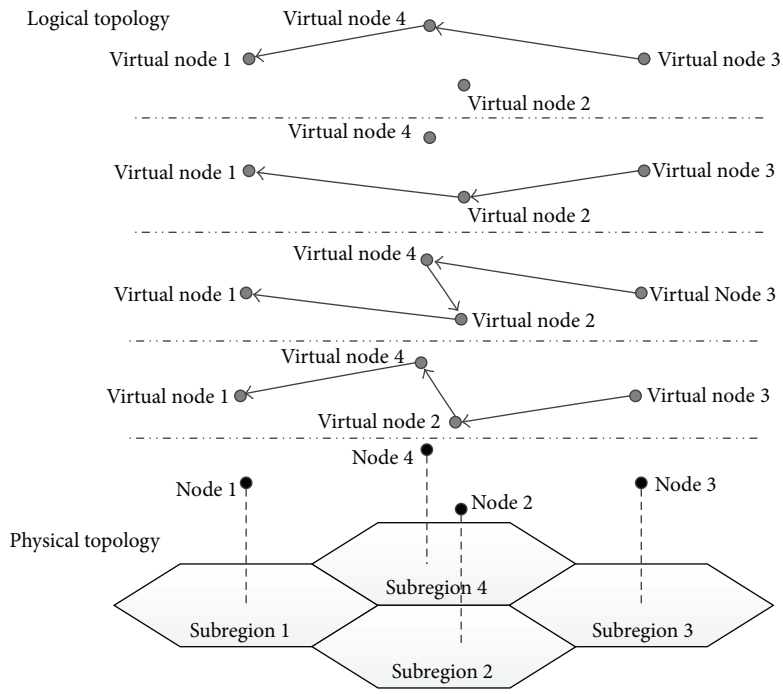
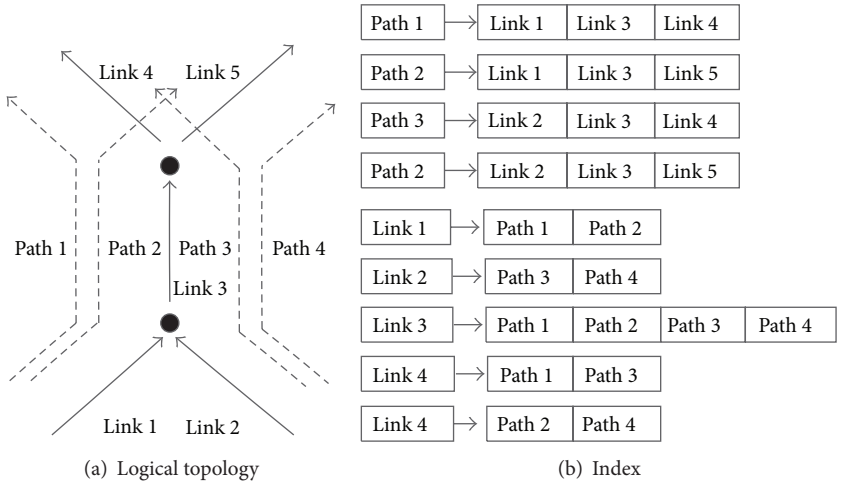


Figure 2: The road network model used for detecting network traffic anomalies.



(a) Logical topology (b) Index

Figure 3: Example of the index.

network. The traffic flow matrix can be further classified as an NtN (node-to-node) traffic flow matrix.

Definition 7 (NtN traffic flow matrix). If the network has n nodes and the traffic flow of any path can be measured constantly over a certain time interval, then the measured value can be created as a $T \times w$ matrix to represent a time sequence of the measured traffic flow. Here, T is the number of measured cycles and w is the number of traffic flow measurements; thus, $w = n \times n$. Row t is a vector of traffic flow value, which is measured in the t cycle and can be represented by x_t. The column j is the time sequence of the traffic flow value of j virtual OD node pairs. In addition, x_{tj} represents the traffic flow of the j virtual OD node pairs during the t cycle:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,w-1} & x_{1w} \\ x_{21} & x_{22} & \cdots & x_{2,w-1} & x_{2,w} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{T-1,1} & x_{T-1,2} & \cdots & x_{T-1,w-1} & x_{T-1,w} \\ x_{T,1} & x_{T,2} & \cdots & x_{T,w-1} & x_{T,w} \end{bmatrix}. \quad (1)$$

2.2. Traffic Anomaly Detection Method

2.2.1. Traffic Anomaly Detection Process. The detection of traffic anomalies from a wide traffic network can be obtained by developing a method that can determine anomalous subregions in a network to provide effective information for transportation researchers and managers for improving transportation planning and dealing with emergencies. Generally, this problem can be described by considering how to capture the anomalous subregions whose characteristic values significantly deviate from normal values. To achieve this goal, a novel computing process was designed, as shown in Figure 4. In this process, the physical topology layer is transformed according to the structure of the real network. Then, the logical topology layer can be derived and the OD nodes pairs and virtual OD nodes pairs are established simultaneously. Furthermore, the traffic of the paths between the virtual OD nodes pairs is extracted with logical topology information while using the wavelet transform method and PCA to prove the spatial and temporal relationships. Based on the multiscale modeling ability of the wavelet transform and the dimensionality reduction ability of PCA, the network traffic anomalies detection method can be constructed based on multiscale PCA with Shewhart and EWMA control chart residual analyses. Finally, a judgment method is proposed for detecting the anomalous location.

2.2.2. Traffic Anomalies Detecting Method Based on MSPCA. In this section, the space-time relativity of the traffic flow matrix was used to model the ability of the wavelet transform and the dimensionality reduction of PCA to transform the traffic flow of the traffic flow matrix. Next, anomalies were detected using two types of residual flow analysis. The time complexity analysis will be discussed at the end of this section.

Normal traffic flow modeling can be met by using the MSPCA, which can combine the abilities of wavelet transform to extract deterministic characteristics with the ability of PCA to extract the common patterns of multiple variables. Normal traffic flow modeling based on MSPCA can be divided into the four following steps.

Step 1. The first step is the wavelet decomposition of the traffic flow matrix. First, the traffic flow matrix, X, will undergo multiscale decomposition through an orthonormal wavelet transform [29]. Next, the wavelet coefficient matrix Z_L, Y_m $(m = 1, \ldots, L)$ can be obtained on every scale. Then the MAD method [30] is used to filter the wavelet coefficients. Finally, the following filtered wavelet coefficient matrix is obtained:

$$\overline{Z}_L, \overline{Y}_m \quad (m = 1, \ldots, L). \tag{2}$$

Step 2. The second step is principal component analysis and refactoring of the wavelet coefficient matrix. First, the wavelet coefficient matrix $\overline{Z}_L, \overline{Y}_m$ $(m = 1, \ldots, L)$ in every scale is analyzed using PCA. Next, the number of nodes is selected according to the scree plot method [31]. Finally, the wavelet coefficient matrix $\widehat{Z}_L, \widehat{Y}_m$ $(m = 1, \ldots, L)$ is reconstructed.

Step 3. The third step is reconstructing the traffic flow matrix using the invert wavelet transform W^Taccording to the wavelet coefficient matrix $\widehat{Z}_L, \widehat{Y}_m$ $(m = 1, \ldots, L)$ at all scales.

Step 4. The fourth step is principal component analysis and refactoring of the traffic flow matrix. This method is similar to that of Step 2, and the traffic flow matrix can be reconstructed, denoted by $\widehat{X}$.

After the normal traffic flow was modeled, several residual traffic flows were determined, including two components, noise and anomalous traffic. These flows mainly resulted from errors of the traffic flow model and traffic anomalies, respectively. The squared prediction error was used to analyze the residual traffic flows,

$$\mathrm{SPE}_i = \sum_{j=1}^{W} \left(x_{ij} - \widehat{x}_{ij}\right)^2, \tag{3}$$

where $\widehat{x}_{ij}$ is the element in the traffic flow matrix $\widehat{X}$ and W is the number of links in the network.

Then two types of control chart methods were used to analyze the residual traffic flows, Shewhart and EWMA [32]. The Shewhart control chart method can detect rapid changes in traffic flow, but its detection speed is slow for detecting anomalous traffic flows, which change slowly. However, the EWMA control chart method can detect anomalous traffic flows that have a long duration but change slowly.

Shewhart Control Chart Method. The Shewhart control chart method directly detects the time sequence of the squared prediction error and defines ξ_α^2 as the threshold for the squared prediction error at the $1 - \alpha$ confidence level. A statistical test known as the Q-statistic [31] is used to test the residual traffic flows, as follows:

$$\xi_\alpha^2 = \phi_1 \left[\frac{c_\alpha \sqrt{2\phi_2 h_0^2}}{\phi_1} + 1 + \frac{\phi_2 h_0 \left(h_0 - 1\right)}{\phi_1^2} \right]^{1/h_0}, \tag{4}$$

where $h_0 = 1 - 2\phi_1\phi_3/3\phi_2^2$, $\phi_i = \sum_{j=r+1}^{W} \lambda_j^i$, $i = 1, 2, 3$, λ_j is the variance, which can be obtained by projecting the traffic flow matrix to the jth principal component, c_α is the $1 - \alpha$ percentile in the standardized normal distribution, and r is the intrinsic dimensionality of the residual traffic flows data. If the value of the squared prediction error is not less than the threshold value ξ_α^2, an anomaly will appear.

According to the Q-statistic, the multivariate Gaussian distribution follows the assumption of derivation. The Q-statistic will display few changes, even when the distribution of the original data differs from the Gaussian distribution [31]. Thus, the Q-statistic can provide prospective results in practice without examining traffic flows data for adaption assumptions due to its robustness.

EWMA Control Chart Method. The EWMA control chart method can be used to predict the value of the next moment in the time sequence according to historical data. The predicted value of residual traffic flow at time t can be recorded

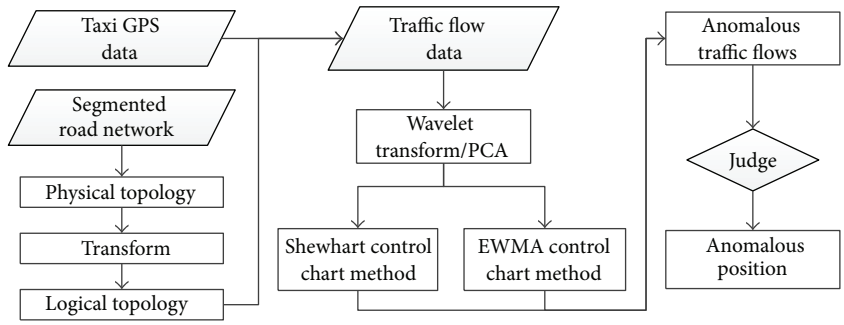


Figure 4: Traffic anomalies detection process.

as $\widehat{Q}_t$, and the actual value of the residual traffic flow at t is Q_t. Thus,

$$\widehat{Q}_{t+1} = \beta Q_t + (1-\beta)\widehat{Q}_t, \tag{5}$$

where $0 \le \beta \le 1$ is the weight of the historical data. The absolute value of the difference between the actual and predicted values $|Q_t - \widehat{Q}_t|$ is obtained, and the threshold value of EWMA can be defined as follows:

$$\psi = \mu_s + L \times \sigma_s \sqrt{\frac{\beta}{(2-\beta)T}}, \tag{6}$$

where μ_s is the mean value of $|Q_t - \widehat{Q}_t|$, σ_s is the mean square error, L is a constant, and T is the length of the time sequence. Thus, if $|Q_t - \widehat{Q}_t| \ge \psi$, an anomaly will appear.

The computational complexity of the proposed method is $O(Tp^2 + Tp)$, which mainly contains the wavelet transform and PCA process.

Currently, the paths which have traffic anomalies can be detected. However, the research goal is to determine which links between the adjacent regions are anomalous. Therefore, another method was designed to locate anomalous links based on the distribution of traffic flow in the next section.

2.2.3. Anomalous Position Locating. According to the analysis results, the paths of OD node pairs may have different traffic flow values at the same time. However, determining which paths are anomalous is not the purpose of this research. The anomalous position should be located to provide useful and clear information for transportation researchers and managers. The proposed method is different from other methods, which detect the anomalous road segment first and then infer the root cause of the traffic anomalies in the road network. Here, the paths with traffic anomalies can be detected and the anomalous position locating process was built as follows. First, the trips were connected with the paths that have traffic anomalies so that all links belonging to an anomalous path can be identified. Next, all links are assumed as potential anomalous links and stored into an anomalous pool. Next, the existing identification method is used to determine whether traffic anomalies exist on these links based on their historical data; this process ends until all of the links are tested. Finally, the links that are not anomalous are deleted and the other links are kept in the anomalous pool.

Links do not exist in the physical world. Thus, anomalous links need to be transformed into anomalous subregions. Based on the experience, the subregions that are connected by anomalous links will have the greatest probability of being anomalous. Thus, all of these subregions should be searched and considered as anomalous subregions. The traffic flow between them is anomalous. So far, the process of traffic anomalies detection has been completely presented.

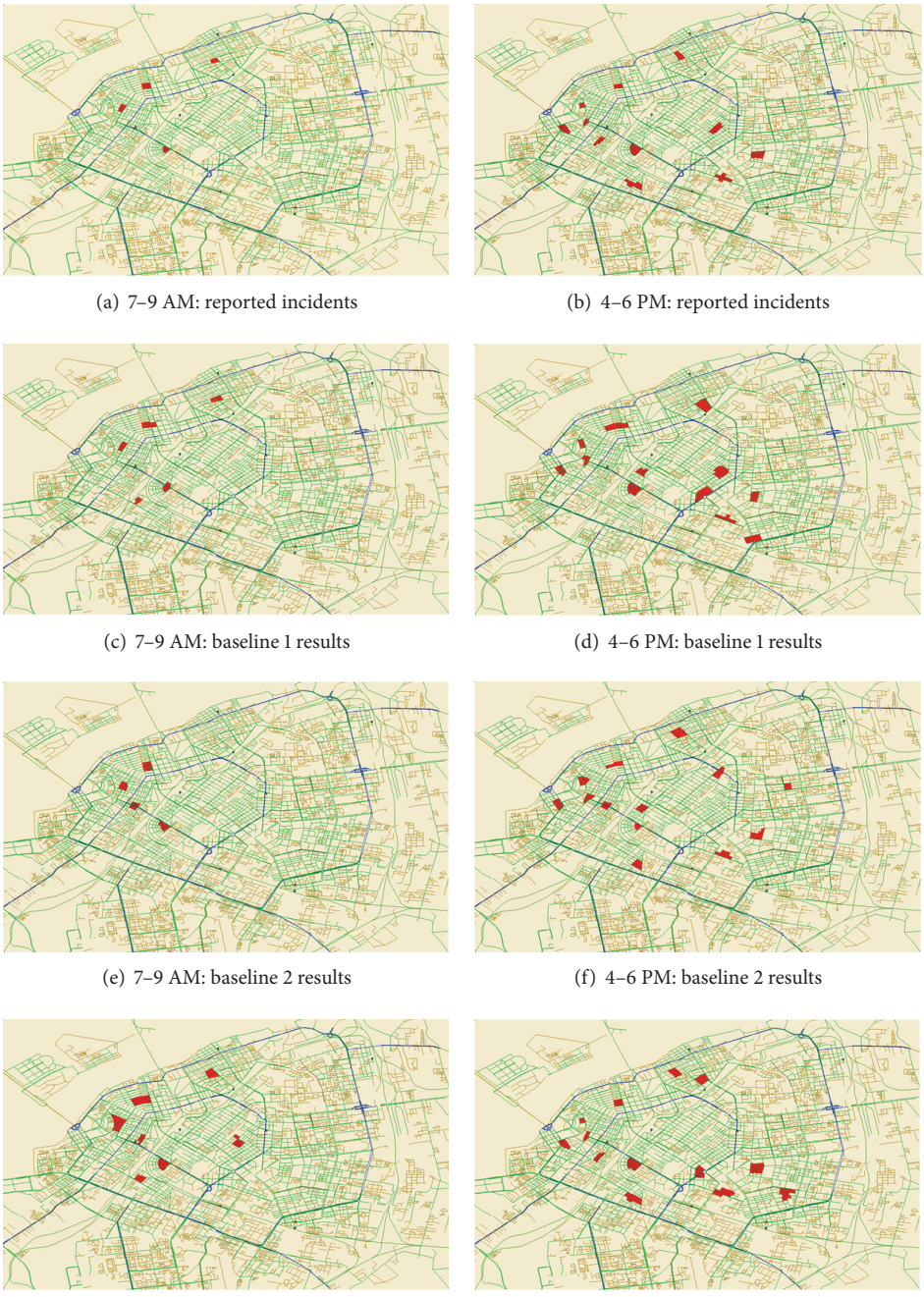
3. Results and Discussions

3.1. The Road Network and Data Preparation

3.1.1. Road Network. The road networks of Harbin were considered as the basic road networks, and the statistical information is shown in Table 2. To obtain a higher detection precision, minor roads and major roads were used to segment the urban area, as shown in Figure 5 (the green lines and blue lines are minor roads and major roads, resp.). Consequently, the area of the subregions became smaller so that the traffic anomalies can be located more accurately. Thus, the number of subregions significantly increases relative to the number shown in Figure 1.

3.1.2. Mobility Data. The taxi GPS data were used as mobility data, as shown in Table 2. Approximately 23% of the daily road traffic in Harbin is generated by taxies. Thus, taxi traffic can indicate the dynamics of all traffic. Although the mobility data were collected from taxies, it can be believed that the proposed method is general enough to use other data sources, which can reflect the characteristics of mobility on the road network, such as the public transit GPS data. All of these data require preprocessing to remove erroneous data and eliminate positioning deviations by map-matching technology.

3.2. Evaluation Approach. In the numerical experiment, the traffic anomalies reported during the half-year period were used as real data to evaluate the detecting effectiveness and performance of this approach. In practice, continuous execution is unrealistic due to the need for large amounts of



(a) 7–9 AM: reported incidents

(b) 4–6 PM: reported incidents

(c) 7–9 AM: baseline 1 results

(d) 4–6 PM: baseline 1 results

(e) 7–9 AM: baseline 2 results

(f) 4–6 PM: baseline 2 results

(g) 7–9 AM: proposed method results

(h) 4–6 PM: proposed method results

Figure 5: Reported traffic anomalies and detection results.

computation; thus, time discretization was used to overcome this fault. The time interval of algorithm execution is 15 minutes. It means the detection method was executed every 15 minutes with the data collected during the latest period as current data. All of the previous data were stored as historical data in the database and used for experimental calculations. In addition, the length of the time interval can be determined based on the actual demand (it is a tradeoff process; readers can refer to Ziebart et al. [11]).

3.2.1. Measurement. In the process of evaluating the effectiveness of the proposed traffic anomalies detection method, traffic anomaly reports were used as a subset of real traffic anomalies because not all traffic anomalies can be recorded in reports. The evaluation method consists of comparing the detection results with the reports to determine how many real traffic anomalies can be detected. Thus, the R parameter was defined to measure the accuracy, which can be expressed as $R = C_d/C_r$, where C_d is the number of reported anomalies

Table 2: Dataset statistics.

	Data duration	Mar.–Aug. 2012
GPS data	Taxies	15,210
	Effective days	74
	Trips	21,510,880
	Avg. sampling interval	60 s
Road network	Road grade	Major and minor roads
	Subregions	387
Reports	Avg. reports per day	28

that can be detected using the proposed method and C_r is the number of anomalies in the reports. This parameter is not a precision measurement because a traffic anomalies report may not provide a complete set of all real traffic anomalies. It is possible that some traffic anomalies can be detected by using the proposed method but should not be recorded in the report, as shown in Figure 5.

3.2.2. Baselines. The accuracy of the proposed method should be evaluated in this process. Two anomalous traffic detection methods were used as baselines: a method based on the likelihood ratio test statistic (LRT) [17] and a modified version of PCA [14]. The ideas used in these two methods are similar to ours; thus, these methods were applied to the matrixes of all subregions to find out the subregions which have an anomalous number of taxies based on our segmentation. Next, the accuracy can be obtained by comparing the results of the three methods.

3.3. Numerical Experiments

3.3.1. Effectiveness. To accurately evaluate the proposed method, two "peak-hour" time intervals on 11/5/2012 were chosen as study period, which are presented in Figure 5 (the red regions of all eight figures indicate the anomalies). Figures 5(a) and 5(b) show the anomalies that were reported during these two time intervals. Figures 5(c) and 5(d) show the anomalies that were detected by using baseline 1 method (the method based on LRT), and Figures 5(e) and 5(f) show the anomalies that were detected by using baseline 2 method (the modified version of PCA). In addition, Figures 5(g) and 5(h) show the detection results of the proposed method.

According to Figure 5, the proposed method detected more traffic anomalies than the baseline methods during each time interval. From 7 AM to 9 AM, baseline 1 method and the proposed method detected all anomalies in the report. However, baseline 2 method only detected 75% of the anomalies. In addition, the results show that the proposed method detected 2~3 more anomalies (which could be potential anomalies) than the baseline methods. From 4 PM to 6 PM, the proposed method can detect 10 reported anomalies. However, baseline 1 and 2 methods resulted in 8 and 9 reported anomalies, respectively. Thus, the proposed method can detect 90.91% of all reported anomalies in this special time interval, which is 18.18% more than the value of baseline 1 method and 9.09% more than the value of baseline 2 method. In the experiments of different time intervals on 11/5/2012, the average R value of the proposed method is 82.37%, but the value of baseline 1 method is only 63.74% and the value of baseline 2 method is 72.70%. When the experiment was extended to another 73 effective days from March to August, as shown in Table 3, the average R value of the proposed method is 74.62%, the value of baseline 1 method is 56.33%, and the value of baseline 2 method is 63.29%. This phenomenon indicates that the detection rate of the proposed method improved by 32.47% and 17.90% relative to baseline 1 and baseline 2 methods, respectively. In addition, according to the R value of each day, the proposed method can detect more reported anomalies than the baselines. Thus, it can be concluded that the proposed method is significantly better than the baseline methods.

To further illustrate the feasibility and superiority of the proposed method, an anomalous subregion was chosen between 7:30 AM and 9:30 AM. In this case, three anomalous paths can be observed in the subregion (their traffic flow is shown in Figure 6). Thus, the path that causes traffic is obvious, and the transportation managers can guide the traffic to the regions that have less traffic pressure.

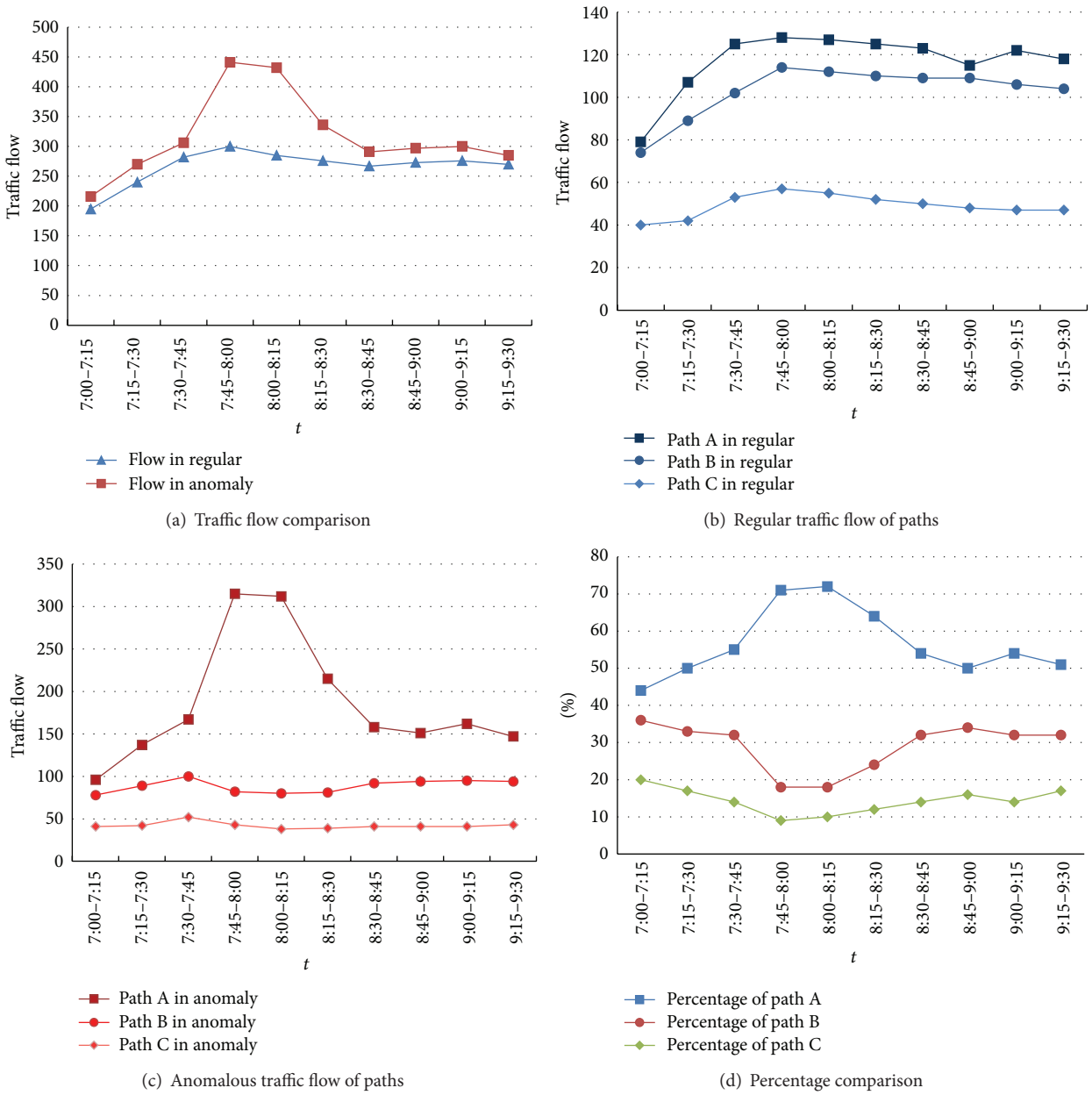
According to Figure 6(a), the overall traffic flow did not differ much from the regular overall traffic flow between 7:00 AM and 7:45 AM. However, between 7:45 AM and 8:30 AM, a significant difference was observed between the two curves. By comparing Figures 6(b) and 6(c), this traffic anomaly resulting from the traffic flow of path A can be observed obviously. According to Figure 6(d), the percentages of the traffic flow in paths B and C declined between 7:45 AM and 8:30 AM because some taxi drivers changed their routes to avoid this anomalous region. After this period, the traffic flow gradually returned to the normal status, as shown in Figure 6(a). Consequently, in the directions with more potential capacity for sharing more traffic flows, such as path B in Figures 6(c) and 6(d), the traffic flow and percentages all decreased during the anomalous interval; thus, a portion of the traffic flow can be guided to this direction to reduce the traffic pressure of anomalous region.

3.3.2. Performance. In the experiments, the hardware/software configuration and average processing time for anomaly detection are shown in Tables 4 and 5, respectively. The urban area was segmented into a number of subregions in the first step, and the following study was affected by the segmentation results. The computing times for different steps are related to the numbers of subregions. Thus, the computing times will be significantly different when the urban area is segmented according to different levels of roads. Specifically, the computing time will increase as the road level decreases, as shown in Figure 7.

3.4. Case Study. In this section, two cases were used to further evaluate the detection method. In the first case, an anomalous region was detected and reported. In another case, the detected anomalous region does not exist in the report; these two cases are shown in Figures 8 and 9,

Table 3: R values of the detection results.

Number	Date	R value of each day		
		Baseline 1 method	Baseline 2 method	Proposed method
1	4/3/2012	59.27%	62.97%	83.17%
2	6/3/2012	64.18%	64.52%	75.86%
3	7/3/2012	53.44%	70.20%	88.49%
⋮	⋮	⋮	⋮	⋮
32	11/5/2012	63.74%	72.70%	82.37%
⋮	⋮	⋮	⋮	⋮
74	31/8/2012	47.28%	77.37%	78.88%
Average R value		56.33%	63.29%	74.62%



(a) Traffic flow comparison

(b) Regular traffic flow of paths

(c) Anomalous traffic flow of paths

(d) Percentage comparison

Figure 6: Effects of time intervals.

Table 4: Hardware/software configuration.

Hardware/software name	Version/size
Server	64-bit
Operating system	Windows Server 2008
CPU	2.50 GHz
Memory	16 Gb

Table 5: Average processing time for anomaly detection.

Procedure name	Time (s)
GPS data transform (one day)	19.17
Wavelet transform/PCA	<2.00
Shewhart & EWMA	2.32

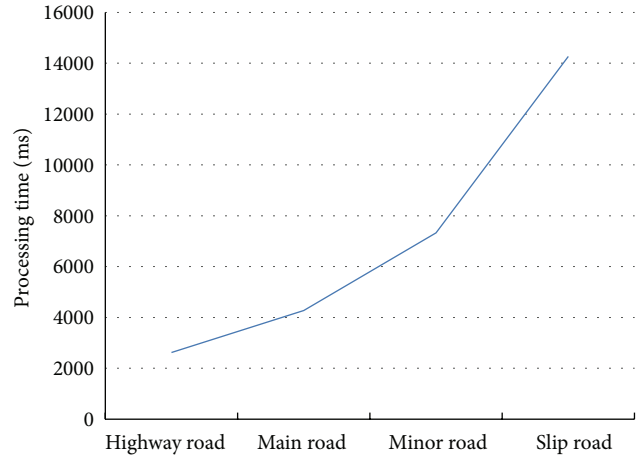


Figure 7: Processing time for anomaly detection.

respectively. Each figure contains three subfigures, with Figures 8(a) and 9(a) presenting the detection results of baseline 1 method, Figures 8(b) and 9(b) presenting the detection results of baseline 2 method, and Figures 8(c) and 9(c) presenting the anomalous subregions detected using the proposed method.

In the first case, road reconstruction occurred on Liaohe Road between 9:00 AM and 11:00 AM on Jun 17, 2012. As shown in Figure 8, the red line presents the work zone and the orange region represents the detected anomalous subregions. In Figures 8(a) and 8(b), the total areas of the anomalous subregions around the work zone are small. However, using the detection results of the proposed method (as shown in Figure 8(c)), a larger collection of anomalous subregions was obtained and all of the paths through these affected subregions can be determined. In contrast with the results from the baseline methods, our advisory paths can avoid the anomalous subregions that were not detected by the baseline methods. Thus, the advisory paths can be more accurate and useful for drivers or management departments to actively avoid the anomalous subregions, such as the black lines in Figure 8(c). These advisory paths can change the actual driving routes of some vehicles, and this effect can reduce the traffic pressure in this area while accelerating the dissipation of anomalies.

In the second case, the proposed method detected a traffic anomaly near the Harbin International Conference and Exhibition Center (HICEC) from 8:30 PM to 10:00 PM on Jul 30, 2012. However, this anomaly was not reported by the traffic management department. As shown in Figures 9(a) and 9(b), baseline 1 method cannot be used to detect any anomalies around the HICEC (gray region), and baseline 2 method can only detect a small region adjacent to the HICEC. However, according to the daily news on the Internet, the Harbin International Automobile Industry Exhibition (HIAIE) was held in the HICEC. The HIAIE is one of the largest exhibitions in Harbin and can attract many dealer and automobile manufacturers that exhibit their products. Thus, a large number of citizens attend this grand exhibition. To ensure safety, the management department deploys many police officers in this area. Thus, the traffic anomalies in this area may be ignored in the reports because it can be assumed that this area is effectively controlled. However, good control does not mean that no traffic anomaly occurs. Large traffic pressure can result in short-term and large-scale traffic anomalies. Thus, the results of these two baseline methods are not sufficient for supporting traffic management and emergency treatment. However, as shown in Figure 9(c), the proposed method detected a large-scale anomalous region around the HICEC, which corresponds better with the actual traffic; thus, the accuracy of the proposed method is much higher than the baseline methods. Consequently, the proposed method is more sensitive to short-term traffic anomalies, and the development and dissemination of traffic anomalies can be controlled well by using the proposed method.

4. Conclusions

A traffic anomalies detection method that uses taxi GPS data was presented to explore one aspect of urban traffic dynamics. And a novel approach based on the distribution of traffic flow was used for locating and describing traffic anomalies. This method provides an effective approach for discovering traffic anomalies between two adjacent regions. The effectiveness and computing performance of this method were evaluated by using a taxi GPS dataset of more than 15,000 taxies for six months in Harbin. This method detected most of the reported anomalies because it combines the advantages of the Shewhart control chart method and the EWMA control chart method. Thus, this method can detect the anomalies caused by rapidly changing traffic flows and slowly changing traffic flows. According to the experimental results, 74.62% of the anomalies reported by the traffic administrative department were identified, which is much higher than the existing methods based on LRT and PCA. Compared with other anomalies detection methods, this method can identify traffic flows that cause traffic anomalies and provide effectiveness information for managers to solve traffic jam or emergency response problems. Furthermore, this method can change the granularity of region segmentation based on the actual

(a) Baseline 1 results

(b) Baseline 2 results

(c) Proposed method results

Figure 8: Case 1 detection results.

(a) Baseline 1 results

(b) Baseline 2 results

(c) Proposed method results

Figure 9: Case 2 detection results.

demand, which satisfies the requirements of traffic anomalies detection for different purposes. The average execution time of this method is less than 10 seconds, and the effectiveness is high enough to support real-time detection of anomalies.

Conflict of Interests

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

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Project no. 71203045), Heilongjiang Natural Science Foundation (Project no. E201318), and the Fundamental Research Funds for the Central Universities (Grant no. HIT.KISTP.201421). This work was performed at the Key Laboratory of Advanced Materials & Intelligent Control Technology on Transportation Safety, Ministry of Communications, China.

References

[1] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '13)*, pp. 334–343, ACM, New York, NY, USA, 2013.

[2] Y. Yue, H.-D. Wang, B. Hu, Q.-Q. Li, Y.-G. Li, and A. G. O. Yeh, "Exploratory calibration of a spatial interaction model using taxi GPS trajectories," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 140–153, 2012.

[3] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai," *Landscape and Urban Planning*, vol. 106, no. 1, pp. 73–87, 2012.

[4] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the International Workshop on Trajectory Data Mining and Analysis (TDMA '11)*, pp. 23–30, ACM, September 2011.

[5] C. Chen, D. Zhang, P. S. Castro et al., "Real-time detection of anomalous taxi trajectories from GPS traces," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 63–74, Springer, Berlin, Germany, 2012.

[6] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "A taxi driving fraud detection system," in *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM '11)*, pp. 181–190, December 2011.

[7] D. Zhang, N. Li, Z. H. Zhou et al., "iBAT: detecting anomalous taxi trajectories from GPS traces," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, pp. 99–108, ACM, 2011.

[8] J. Zhang, "Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pp. 157–162, ACM, August 2012.

[9] H. Wang and R. L. Cheu, "A microscopic simulation modelling of vehicle monitoring using kinematic data based on GPS and ITS technologies," *Journal of Software*, vol. 9, no. 6, pp. 1382–1388, 2014.

[10] J. Yuan, Y. Zheng, C. Zhang et al., "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*, pp. 99–108, ACM, New York, NY, USA, November 2010.

[11] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior," in *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08)*, pp. 322–331, ACM, September 2008.

[12] H. Yoon, Y. Zheng, X. Xie, and W. Woo, "Smart itinerary recommendation based on user-generated GPS trajectories," in *Ubiquitous Intelligence and Computing*, vol. 6406 of *Lecture Notes in Computer Science*, pp. 19–34, Springer, Berlin, Germany, 2010.

[13] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 316–324, ACM, August 2011.

[14] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM '12)*, pp. 141–150, December 2012.

[15] J. A. Barria and S. Thajchayapong, "Detection and classification of traffic anomalies using microscopic traffic variables," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 695–704, 2011.

[16] Q. Chen, Q. Qiu, H. Li, and Q. Wu, "A neuromorphic architecture for anomaly detection in autonomous large-area traffic monitoring," in *Proceedings of the 32nd IEEE/ACM International Conference on Computer-Aided Design (ICCAD '13)*, pp. 202–205, IEEE, November 2013.

[17] C. Chen, D. Zhang, P. S. Castro, N. Li, L. Sun, and S. Li, "Real-time detection of anomalous taxi trajectories from GPS traces," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, vol. 104 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 63–74, Springer, Berlin, Germany, 2012.

[18] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, pp. 89–98, ACM, September 2011.

[19] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1010–1018, ACM, New York, NY, USA, August 2011.

[20] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. V. D. Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2159–2168, 2013.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, pp. 851–860, ACM, April 2010.

[22] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions," in *Proceedings of the 18th International Conference on Intelligent User Interfaces (IUI '13)*, pp. 203–212, ACM, March 2013.

[23] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, article 15, 2009.

[24] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[25] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On detection of emerging anomalous traffic patterns using GPS data," *Data & Knowledge Engineering*, vol. 87, pp. 357–373, 2013.

[26] D. Jiang, P. Zhang, Z. Xu, C. Yao, and W. Qin, "A wavelet-based detection approach to traffic anomalies," in *Proceedings of the 7th International Conference on Computational Intelligence and Security (CIS '11)*, pp. 993–997, December 2011.

[27] A. Gran and H. Veiga, "Wavelet-based detection of outliers in financial time series," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2580–2593, 2010.

[28] N. J. Yuan, Y. Zheng, and X. Xie, "Segmentation of urban areas using road networks," Tech. Rep. MSR-TR-2012-65, Microsoft Research, 2012.

[29] S. G. Mallat, "Theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[30] B. R. Bakshi, "Multiscale PCA with application to multivariate statistical process monitoring," *AIChE Journal*, vol. 44, no. 7, pp. 1596–1610, 1998.

[31] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 219–230, 2004.

[32] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: an overview," *Quality and Reliability Engineering International*, vol. 23, no. 5, pp. 517–543, 2007.