Front.Comput.Sci.
DOI10.1007/s11704 − 019 − 9034 − z

RESEARCH ARTICLE

Understanding Urban Structures and Crowd Dynamics Leveraging Large-Scale Vehicle Mobility Data

Zhihan JIANG, Yan LIU, Xiaoliang FAN, Cheng WANG, Jonathan LI, Longbiao CHEN*

Fujian Key Laboratory of Sensing and Computing for Smart Cities,
School of Information Science and Engineering, Xiamen University, Xiamen 361005, China

Abstract

A comprehensive understanding of city structures and urban dynamics can greatly improve the efficiency and quality of urban planning and management, while the traditional approaches of which, such as manual surveys, usually incur substantial labor and time. In this paper, we propose a data-driven framework to sense urban structures and dynamics from large-scale vehicle mobility data. First, we divide the city into fine-grained grids, and cluster the grids with similar mobility features into structured urban areas with a proposed distance-constrained clustering algorithm (DCCA). Second, we detect irregular mobility traffic patterns in each area leveraging an ARIMA-based anomaly detection algorithm (ADAM), and correlate them to the urban social and emergency events. Finally, we build a visualization system to demonstrate the urban structures and crowd dynamics. We evaluate our framework using real-world datasets collected from Xiamen City, China, and the results show that the proposed framework can sense urban structures and crowd comprehensively and effectively.

Keywords vehicle mobility, big data, spatial clustering, event detection, urban computing, ubiquitous computing

1 Introduction

In order to facilitate efficient urban planning and effective city management, urban authorities need to understand the city functions of different areas [1], as well as the crowd dynamics moving around the city [2]. On one hand, urban planning, construction, and development have led to regular *urban structures* [3], such as central business districts (CBDs), residential areas, and transit hubs. Meanwhile, the occurrences of urban events may break the regular *crowd movement* patterns in different areas of the city [4, 5]. For example, holding a concert in a stadium may lead to abnormal human flow peaks around the stadium and the city's transit hubs. Due to the lack of a comprehensive understanding of the urban structures and crowd dynamics, urban authorities face difficulties in evaluating the impacts of urban social and emergency events, which affect the short-term event management and long-term urban planning. Therefore, it is of great importance for urban authorities to have a clear picture of urban structures, and to be able to analyze the crowd dynamics caused by urban events.

Urban structures are a reflection of social economy, culture and other factors in space [6, 7]. There are many different ways to analyze urban structures, such as divide the city into different functional districts. For example, Yuan et al. [8] analyzed the structures of urban areas by dividing urban roads into different blocks with image processing technologies. However, the fast pace of urban development has led to dynamic urban structures, and different urban areas exhibit complex spatial-temporal dependencies [9, 10]. Such dynamic and complex attributes can be reflected by human mobility patterns in and among these districts [11]. Therefore, we seek to incorporate human mobility data sources for urban structure analytics.

Urban events, on the other hand, may obstruct the normal operation of cities, and even cause great losses of lives and

Published: July 12, 2019

E-mail: *longbiaochen@xmu.edu.cn*

properties [12]. For example, a concert in a stadium may lead to a significant increase of vehicles in its surrounding areas and cause traffic congestions. Specifically, different kinds of urban social and emergency events may lead to different human mobility patterns in different urban areas [4, 5, 13]. A comprehensive perception of the urban structures can greatly benefit the detection and understanding of abnormal crowd dynamics, and help urban authorities make a quick response to urban events.

However, traditional methods of sensing urban structures and crowd dynamics mainly rely on manual surveys and statistical reports [14], which have following limitations. First, detailed surveys consume substantial time and labor, due to the large scale of cities [5]. Second, surveys are not comprehensive, since manual investigation may not cover the city-wide human activities and movements. Fortunately, with the advances of Internet of Things and big data technologies, various kinds of urban sensing data have been available, providing us with new opportunities to analyze urban structures and crowd dynamics in a data-driven manner [1, 13, 15].

In this work, we exploit large-scale vehicle mobility data extracted from city-wide taxi trajectories to sense urban structures and crowd dynamics. Meanwhile, we retrieve detailed information about urban social and emergency events from social media to verify the causes of abnormal crowd dynamics. In particular, we propose a two-phase framework to sense urban structures and crowd dynamics in a data-driven manner. In the first phase, we divide urban areas into geographical grids, and then map taxi GPS trajectories into the corresponding grids to obtain the spatial-temporal crowd mobility patterns. We then cluster grids with similar mobility patterns to obtain city functional regions. In the second phase, we detect abnormal traffic flows in each functional region, and verify the abnormal crowd dynamics with urban event information retrieved from social media. Finally, we build a visualization system to demonstrate the mobility patterns and impacting scopes of different types of urban events. In summary, the **main contributions** of this paper are summarized as follows:

1. We propose a data-driven methodology for sensing urban structures and crowd dynamics, which greatly benefits urban planning and development.
2. We propose a two-phase framework to analyze urban structures and crowd dynamics leveraging vehicle mobility data. In the *urban structure portrait* phase, we divide the city into fine-grained mobility grids, and propose a distance-constrained clustering algorithm (DCCA) to cluster the grids with similar mobility features into structured urban areas. In the *crowd dynamic characterization* phase, we detect irregular mobility traffic patterns in each area leveraging an ARIMA-based anomaly detection algorithm (ADAM), and correlate them to the urban social and emergency events extracted from social media.
3. We evaluate our method using real-world taxi GPS trajectory and social network data collected from Xiamen City, China. Results show that our method can sense the urban structures and characterize the crowd dynamics in urban events effectively and comprehensively, and consistently outperforms the baseline methods. A visualization system is developed to demonstrate the mobility patterns and impacting scopes of different types of urban events.

The remainder of this paper is organized as follows. We first review the related work in Section 2, and then describe the dataset in Section 3. We describe the two-phase framework in Section 4. In section 5 and Section 6, we present the proposed urban structure sensing and crowd dynamics characterization techniques. The results of experiments are shown in Section 7. Finally, we conclude the paper in Section 8.

2 Related Works

In this section, we survey the related research work from the following two aspects: (1) analysis of urban functions and structures, and (2) urban events and impacts.

2.1 Urban Functions and Structures

A series of existing work related to urban function and structure analysis have been conducted in the literature [8, 16, 16]. For example, Yuan et al. [8] employed image processing techniques to extract information from remote sensing images of urban roads, and then extract urban areas surrounded by road segments. This image-based method does not consider the latent social functions, populations, and human activities. Esch et al. [16] estimated the population distribution through the density of buildings and constructions, and mapped them to different urban structures. Chen et al. [17] attempted to depict the crowd dynamics and urban structures through the communication logs of mobile phone users. However, the fine-grained geographic information is difficult to obtain from mobile phone logs due to privacy concerns [18]. To address this problem, grid-based geographic segmentation approaches have been proposed [19, 20]. In this paper, we first divide urban areas into grids, and then characterize the daily traffic patterns of the grids.

2.2 Urban Events and Impacts

Researchers have exploited different methods to study the dynamics, causes and impacts of urban events [4, 5, 21, 22]. For example, Liang et al. [23] used the LBSN check-in data to model the size and duration of crowd gathering in urban events. However, the social network check-in data is usually biased [24], and it is difficult to estimate the event duration accurately. Another method is to detect urban events from a large number of text streams from social media. For example, Sakaki et al. [25] detected typhoon and earthquakes from tweets with locations, and Li et al. [22] investigated the similarity of users to find out unusual urban events mentioned in Twitter. Agarwal et al. [26] applied a graph clustering algorithm for event detection by finding dense sub-graph structures. Zhang et al. [4] proposed an approach to automatically discover the time, venue, and scale of urban events from abnormal human activities. In general, the above-mentioned works mainly focus on detecting urban events, but lack in the evaluation of their impacts on urban structures and crowd dynamics. In this paper, we correlate urban events with city structures via abnormal crowd movements, and analyze their impacts in a comprehensive manner

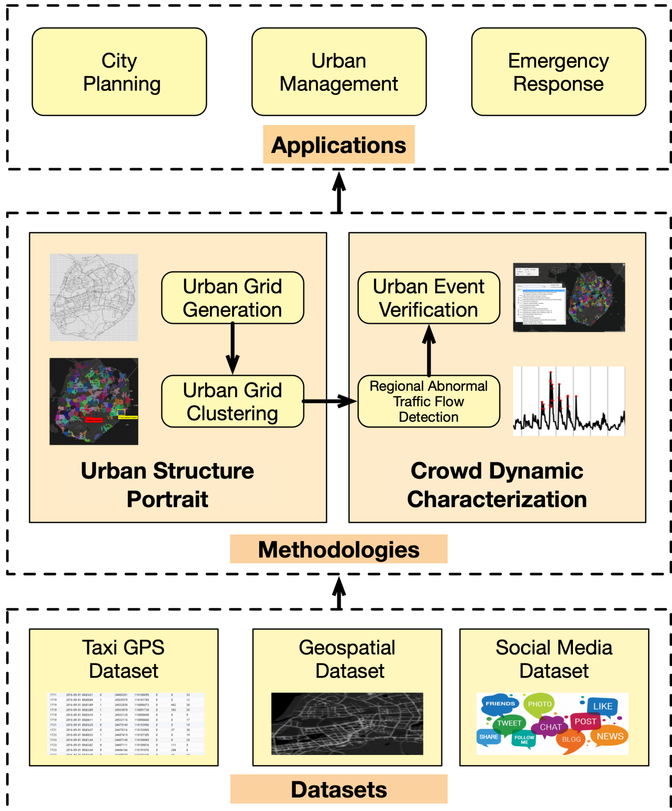


Fig. 1 Framework overview.

3 Framework Overview

The objective of this work is to characterize crowd dynamics in different urban structures in a low-cost and automatic manner. To this end, our framework consists of two phases, i.e., the urban structure portrait phase and the crowd dynamic characterization phase, as shown in Figure 1. Specifically, we first collect the relevant urban datasets, including taxi GPS trajectories, geographic boundaries, and social media contents. In the urban structure portrait phase, we first divide a city into equal-sized grids, and then map taxi GPS trajectories into the corresponding grids. Then we extract the spatial-temporal profiles of crowd movements in grids, and cluster them to obtain semantic *urban regions* leveraging a distance-constrained clustering algorithm (DCCA). In the crowd dynamic characterization phase, we detect anomaly crowd mobility flows for each urban region, and correlate these anomalies with real-world urban events discovered in the social media dataset. Finally, we build a dynamic visualization system to demonstrate the impact of urban events, and provide support for urban planning and management.

4 Dataset Description

Before elaborating the details of the proposed method, in this section, we present the datasets employed in this paper, including the taxi GPS trajectory dataset, geographic dataset, and the social media dataset collected in the city of Xiamen.

4.1 Taxi GPS Dataset

We collect the taxi GPS trajectories in September, 2016 from Xiamen transportation authority. The number of taxis in Xiamen City is 5,486. The GPS records about these taxis are uploaded to the authority in every 30 seconds. Subsequently, we obtain a dataset containing about 377 million records, with an average of 68,666 records per taxi. The raw data was stored in an Oracle database and the main fields of the taxi GPS trajectory data include:

- **ID**: the unique identity of a taxi.
- **LONGITUDE**: the longitude of a record in 10^6 degree.
- **LATITUDE**: the latitude of a record in 10^6 degree.
- **STATE**: the taxi service status, 1 if the taxi is occupied and 0 if it is vacant.
- **TIMESTAMP**: the time of a record in millisecond.

4.2 Social Media Dataset

Today, information about urban social and emergency events can be found in social networks via the posts, images, and location-based check-ins of their users [6,27,28]. Such social media information can be used to evaluate and characterize the time, venue, and impact of social events [4, 5]. In this work, we collect the posts, images, and check-ins in Xiamen during September, 2016 from Weibo, one of the largest social network in China. We store the information in a Mongodb database for query and maintenance.

5 Urban Structure Portrait

Due to the urban planning and the establishment of functional districts, different areas in the city show different structural characteristics with different daily traffic flow patterns. For example, there are large differences in traffic flow pattern between school districts and business districts. In order to demonstrate the urban structures clearly, we divide the whole city into smaller areas according to the spatial-temporal characteristics of vehicle trajectory data

In this chapter, our objective is to divide urban structures. Each of the divided area consists of small areas with similar structures. First, we preprocess the raw taxi GPS trajectory data. Secondly, we divide the urban areas into fine-grain grids. Finally, we cluster the urban grids with a distance-constrained clustering algorithm (DCCA).

5.1 Data preprocessing

Each record in the database contains its ID, timestamp, latitude, longitude and state. Since vehicles travel on urban roads, if the density of the trajectory data is large enough, the urban road network can be covered. Besides, the trajectories of taxis are different from those of private cars influenced by the commuting time, which are more flexible, and because of the shift system, the taxi track covers all 24 hours a day. Therefore, the trajectories of taxis have wide coverage of time and space. Then we extract the off-passenger trajectory from the dataset where the state of the vehicle jumped from 1 to 0.

Fig. 2 An illustrative visualization of the taxi drop off points in Xiamen City from the dataset.

The $trajec\_drop = (ID, lon, lat, t\_drop)$ represents a drops off record from a taxi with ID number. And the drops off location is (lon, lat), lon represents longitude, lat represents latitude, and t_drop represents the time for alighting. As shown in Figure 2, the drop off points can almost cover all roads in the city, which shows that the data are large enough to sensing the structure of the whole city.

5.2 Urban Grid Generation

In order to divide urban areas into smaller ones, we grid the whole urban areas [29]. The whole city is divided into $M \times N$ matrix according to its geographical characteristics, which is the formation of a consistent size and arrangement of small rectangular. Then, according to the latitude and longitude range of each rectangle, the drop off information are mapped into the corresponding rectangles, sorted by the timestamps. Then we count the number of records in grid i within a certain duration t as $g_{i,t}$. Specifically, we calculate the number of drop off vehicles in the grid i with latitude between lat_min and lat_max and longitude between lon_min and lon_max as follows:

$$\begin{aligned} g_{i,t} = \{&\text{count}(trajec\_drop) | t\_drop \in t, \\ &lon_{min} \leqslant lon \leqslant lon_{max}, lat_{min} \leqslant lat \leqslant lat_{max}\} \end{aligned} \tag{1}$$

And the set $G_i = \{g_{i,1}, g_{i,2}, g_{i,3}, ..., g_{i,n}\}$ describes the drop off feature of grid i. Then we form a $M \times N \times T$ tensor. As show in Figure 3.

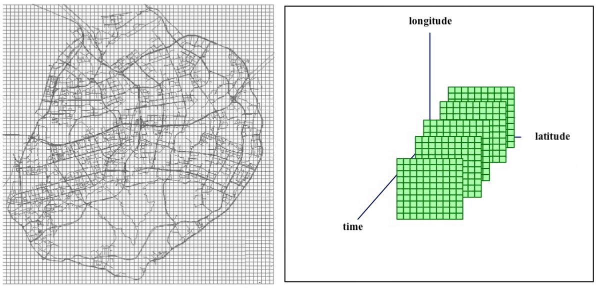


Fig. 3 (a) Grids mapping to city areas. (b) Analytics of the time series for each grid to form a $M \times N \times T$ tensor.

5.3 Urban Grid Clustering

After dividing the whole city into grids, we observed that the traffic pattern of a grid is highly dynamic under different temporal contexts. As shown in the Figure 4.

It is obviously that weekday-weekend patterns are regular and there is a negative correlation between the similarity of traffic flow patterns and the distance of the areas. For example, the traffic flow patterns of grids from one place, such as railway station or convention center, are quite similar, and the traffic flow patterns from different places, such as one from railway station and another from convention center, are very different, which shows the spatial-temporal characteristics of the grids' daily flow patterns. We characterize the flow pattern of each grid using a temporal-context-based profile. More specifically, given a grid g_i and extract its flow vector measured in hours from its time series G_i. We aggregate and average the traffic flow from Monday to Friday in each week to build a typical weekday traffic flow profile, i.e.

$$f_w(g_i) = [u_1, u_2, u_3, u_{120}] \tag{2}$$

Similarly, we build a typical weekend traffic flow profile by aggregating and averaging the flow in Saturday and Sunday of each week, i.e.

$$f_n(g_i) = [v_1, v_2, v_3, v_{24}] \tag{3}$$

Finally, we concatenate the weekday and weekend traffic flow profiles to obtain the temporal-context-based traffic flow profile, i.e.

$$f(g_i) = [f_w(g_i), f_n(g_i)] \tag{4}$$

Due to the spatial dependency, the grids close to each other should be clustered together. Firstly, we construct a weighted graph $G = (V, E)$, where $V = \{g_1, g_2, ..., g_N\}$ denotes the set of N grids and E denotes the set of links between every two grids, to represent the relationship of grids.

Secondly, we define the adjacency matrix A of graph G, which is an asymmetric $N \times N$ matrix with entries $a_{i,j}$ = 1 when there is a link between grid g_i and grid g_j, and $a_{i,j} = 0$ otherwise. We use the geographic distance of two grids to determine whether they are adjacent or not. More specifically, for grid g_i and grid g_j, we define:

$$a_{i,j} = \begin{cases} 1, & if \quad dist(g_i, g_j) \leqslant \tau \\ 0, & otherwise \end{cases} \tag{5}$$

where $dist(g_i, g_j)$ is the geographic distance between the two grids, and τ is a neighborhood threshold controlling the geographic distance of neighboring grids.

Given two adjacent grids, we use their similarity measurement to determine their link weight, i.e.

$$w(g_i, g_j) = (SIMDDIST\{f(g_i), f(g_j)\}) \times a_{i,j} \tag{6}$$

We note that $w(g_i, g_j) = 0$ when $a_{i,j} = 0$, which means that there is no link between g_i and g_j.

Given a graph $G = (V, E)$, we define a set of clusters $R = \{C_l, ..., C_k\}$, where $\cup_{\forall C_k \in R} = V$, and $\cap_{\forall C_k \in R} = 0$. Then, given a grid v, we define the connectivity of v to a cluster C as the sum of link weights between v and the grids in the cluster C, $con(v, C) = \sum_{v' \in C} w_{v,v'}$. Finally, we define the adjacent clusters $C(v)$ of v as $C(v) = \{C | con(v, C) > 0, C \in R\}$.

With the definitions above, our objective is to find an optimal set of clusters R, so that the connectivity within a cluster is higher than the connectivity between different clusters, i.e.

$$\forall v \in C_k, con(v, C_k) \geqslant max\{con(v, C_l), C_l \in R\} \tag{7}$$

We bound the distance span of a cluster within the threshold, i.e.

$$\forall v, v' \in C_k, dist(v, v') \leqslant \tau \tag{8}$$

Based on [30], we use a Distance Constrained Clustering Algorithm (DCCA) to cluster area grids. The basic idea of DCCA is iteratively assigning grids to the adjacent clusters, where the gain of assigning grid v to cluster C is iteratively evaluated by a value function as follows:

$$value(v, C) = con(v, C) \times log(\frac{\tau}{max(dist(v, v'))}) \tag{9}$$

The DCCA greedily assigns the grids to the adjacent cluster with highest value until none of the grids are moved among clusters. As the convergence of such a greedy approach is difficult to prove, the algorithm is stopped in the following two cases: (1) the user specified maximum iteration number max_iter is reached, or (2) none of the grids are moved among clusters.

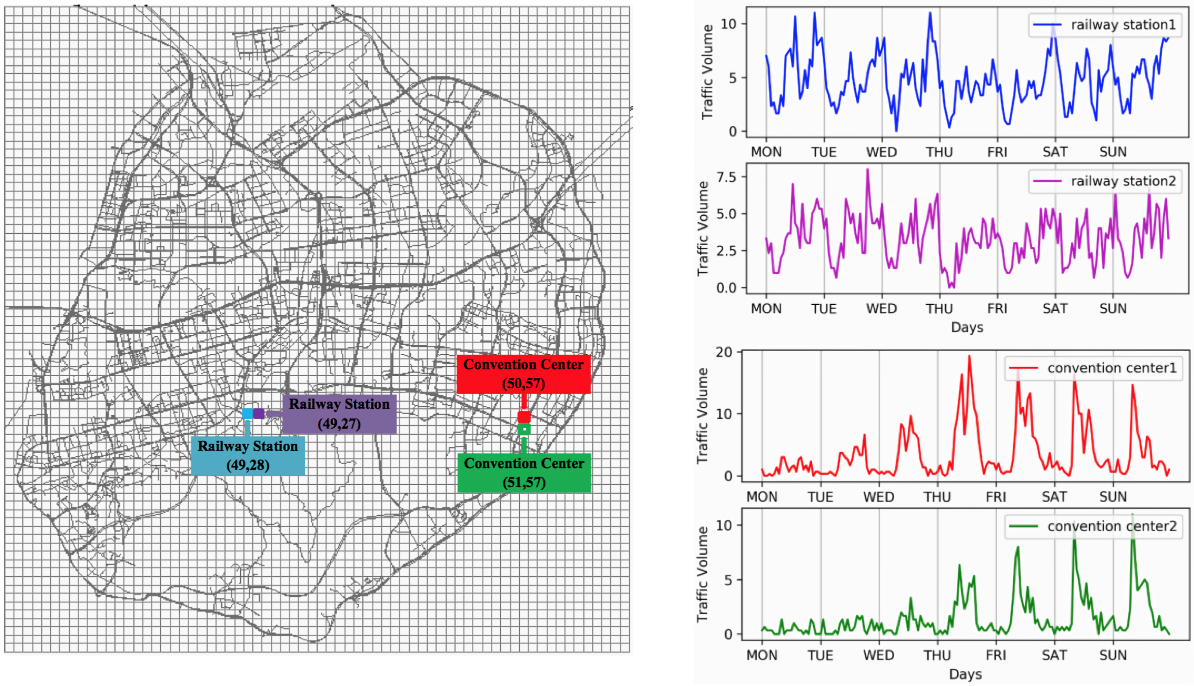


Fig. 4 The flow patterns of the four sample regional grids. Convention and Exhibition Center accounts for two grids, while the Railway Station accounts for the other two grids.

6 Crowd Dynamics Characterization

The daily state of the city shows a certain regularity. However, the occurrence of an urban event often breaks this regularity. In this section, the main objective is to explain the impacts of the urban events on crowd dynamics by detecting the anomalous traffic flow in urban structures. The occurrences of urban events can affect the daily traffic characteristics and different types of events have different impacts on crowd dynamics. For example, a concert is held in one place and the number of people in the vicinity of this area increases before the concert starts, resulting in increasing activities in the area. In addition, if there is a natural disaster in the city, such as a typhoon, it will lead to the stagnation of urban production, paralysis of traffic and reduction of people's mobility, and many areas of the city will be less active.

Through analyzing the impacts of urban events on the crowd dynamics, we can help the city managers to improve the efficiency of the city and reduce the loss caused by emergencies. The analysis consists of two steps. First, we use anomaly detection algorithms to detect abnormal traffic from daily traffic patterns in each region. Second, we verify every abnormal traffic point with urban events based on the social media news, and then analyze the impacts of urban events on the dynamics of urban areas.

6.1 Regional Abnormal Traffic Flow Detection

Through the structure analysis of the urban areas in the previous section, we can extract the daily historical traffic flow characteristics of each region. If an influential urban event occurs, the flow characteristics of the region will change greatly, as show in Figure 5. By observing the 30-day total flow chart of a region and decomposing it into daily flow chart, we can find red anomalous traffic points. We can detect those irregularities from all basic patterns at once using the ARIMA Outlier Detection [31] method.

More specifically, we first extract the daily historical traffic flow characteristics $A(r,n)$ of each area r, n represents the area r consists of n area grids, i.e.

$$A(r,n) = \{a_1, a_2, ..., a_k\} \tag{10}$$

Where $a_i (1 \leqslant i \leqslant k)$ represents the sum of the flow in the ith hour of n regional grids. The traffic flow characteristics vector $A(r,n)$ is divided into testing set and training set. Then we use ARIMA to train the basic flow pattern model, $Model(r, j)$, where there is no urban event.

$$Train = \{a_1, a_2, ..., a_j\} \tag{11}$$

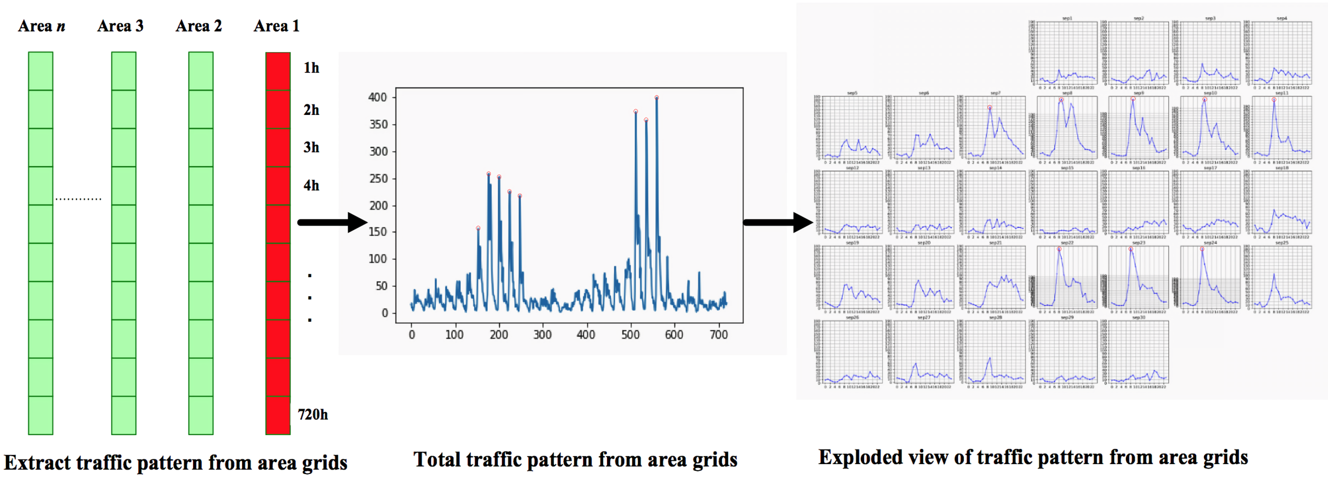


Fig. 5 The flow pattern of the extracted area grid with a 30-day flow characteristic in one area along with a daily flow chart.

$$Test = \{a_{j+1}, a_2, ..., a_k)\} \tag{12}$$

$$Model(r, j) = ARIMA(Train) \tag{13}$$

In this way, we use the model $Model(r, j)$ to predict traffic $a'_{j+1}, a'_{j+2}, ..., a'_{j+t}$ for the next time t and detect the abnormal traffic flow with the anomaly detection algorithm [32], which requires individual threshold δ_m to control the irregularity significance for each basic pattern of each region.

$$\begin{aligned} &a'_m \in (a'_{j+1}, a'_{j+2}, ..., a'_{j+t}) \\ &a_m \in (a_{j+1}, a_{j+2}, ..., a_{j+t}) \in Test, |a_m - a'_m| \geqslant \delta_m \end{aligned} \tag{14}$$

If $|a_m - a'_m| \geqslant \delta_m$, it means that the abnormal traffic flow point appeared in the moment of m. We can find out the abnormal traffic flow in all areas at any time by this method.

6.2 Urban Event Verification

6.2.1 Identify Points of Interest

POIs(Points of Interest) mainly include the scenic spots, governmental agencies, school districts, companies and other geographical entities used in the daily lives of urban residents. According to the category of POI in the area, we can determine the location of the urban event and connect the anomaly traffic points with the urban event. We download the POI distribution dataset of Xiamen from Baidu, including categories of catering, landscape, company, medical, education, and each category has fields shown in Table 1. According to the latitude and longitude range of urban areas ,we can count the number of POIs in each area, which brings convenience to search urban events.

Table 1 The fields specifications of the POI dataset.

Order	Field Name	Instruction	Field Type
1	uid	identifier	int
2	name	the name of POI	varchar(128)
3	longitude	the longitude of POI	float
4	latitude	the latitude of POI	float
5	telephone	the telephone of POI	varchar(50)
6	address	the address of POI	varchar(100)

6.2.2 Search and Verify Urban Events

We search the social media for urban events that coincides with the time and location of anomalous flow peaks, verify each abnormal flow point with urban events in reality and correlate the urban events to the results. Then we analyze the impacts of such urban events on crowd dynamics, including the scopes of the affected areas and durations, and match the anomaly traffic points with urban events to build the urban event dataset.

By utilizing the visual display technology, we show the urban structures and the impact areas of urban events, which can help city managers and the general public have a better understanding of the urban structures and dynamics, prepare for the similar events in advance to keep some accidents from happening and take steps in time to reduce the loss.

7 Evaluation

In this section, we evaluate our method using real-world taxi CPS trajectory data and social media data from Xiamen island. We introduce the experiment settings first. Then we present the evaluation results. Finally, we display our analysis results on the visualization platform.

7.1 Experiment Settings

7.1.1 Evaluation Plan

Firstly, we map the grids to the coverage areas of Siming and Huli districts in Xiamen and aggregate the traffic data of the corresponding grids. We select the data between 09/01/2016 and 09/30/2016 to generate traffic profile of the grids and cluster the grids into urban structure regions with the DCCA algorithm. Then we extract the daily traffic characteristics of each clustered region, and use anomaly detection algorithm to detect abnormal traffic points. Finally, we verify each abnormal traffic point with urban events and display the distribution of the impact of each urban event on the dynamic of urban regions.

7.1.2 Evaluation Metrics

We evaluate the anomaly detection algorithm that we used by computing the accuracy of the anomaly traffic detection results and the precision and recall of the anomaly traffic flow detection results. If a detected traffic abnormal point has a temporal overlapping with the real world urban event, we mark the detection as a hit. In this paper, the precision and recall are calculated as follows:

$$precision = \frac{A}{B} \tag{15}$$

$$recall = \frac{A}{C} \tag{16}$$

where A denotes the detected abnormal points that corresponding to urban events, B denotes all detected abnormal points and C denotes the abnormal points caused by all urban events.

In addition, we calculate the *F1-Score* as

$$F1\text{-}Score = \frac{2 \times precision \times recall}{precision + recall} \tag{17}$$

to assess the performance of the anomaly detection algorithm that we used.

7.1.3 Baseline Method

We compare our abnormal traffic detection algorithm with several baseline methods as follows:

1. **ULQT (Upper and Lower Quartile Threshold)**: This method establishes a threshold range with upper and lower quartiles, and regards the traffic values above the threshold range as outliers. The model predicts the flow value in the next time range, compares it with the real value, and then detect the abnormal value compared with the threshold.
2. **S-H-ESD (Seasonal-Hybrid-ESD)**: This anomaly detection algorithm is from Twitter, which employs statistical learning to detect anomalies in both application, and system metrics. It employs seasonal decomposition to filter the trend and seasonal components of the time series, and use median and median absolute deviation to detect anomalies.
3. **iForests-Based (Isolation Forests-Based) Method**: This method splits the points in the data into outliers or inliers, depending on how long it takes to separate the points [33]. A non-outlier will have many points around, so it will be really difficult to isolate it, while if a point is an outlier, it will be alone and can be found easily. The advantage of this method is that it can work with a huge dataset and several dimensions.

Fig. 6 Results of urban structure portrait.

7.2 Evaluation Results

7.2.1 Urban Structure Portrait Results

The geographical area of Xiamen is partitioned into 77×68 grids with the grid size about 200×200 square meters. In each grid, the normalized traffic volume is recorded on an hourly basis. It is meaningless to analyze all grids and it will waste computing resources because the number of grids is large and the small traffic volume in some grids cannot affect the traffic

Table 2 The Results of Anomalous Traffic Flow Detection

Methods	**Precision**	**Recall**	**F1-Score**
ULQT	31.6%	90.7%	46.9%
S-H-ESD	80.0%	62.5%	70.2%
iForests	50.0%	58.1%	53.7%
ARIMA (Proposed)	71.8%	92.0%	**80.7%**

greatly. Therefore, we check the traffic volume of all grids and select 1,655 grids.

We chose the maximum number of iterations max_iter = 20 and distance threshold $\tau = 1km$, and then 1665 grids were clustered into 147 regions using the DCCA clustering algorithm. The clustering regions are shown in the Figure 6, where adjacent grids of the same color represent a region.

We find that the flow characteristic curve of a single region after clustering is relatively stable and has spatial-temporal characteristics. There are differences in the traffic characteristics at different locations, and the trends of traffic volume at different times in the same location are regular. As show in Figure 7.

7.2.2 Anomalous Traffic Flow Detection Results

The performance comparison of two abnormal traffic flow detection algorithms is shown in the TABLE 2. ARIMA anomaly detection algorithm achieves the best performance, except that its precision is lower than than of the S-H-ESD method. In particular, the F1-Score of ARIMA anomaly detection algorithm achieves 80.7%, which is the highest. So in this paper, we detect abnormal traffic flow through ARIMA anomaly detection algorithm mainly.

Firstly, we use ARIMA to train the model in terms of traffic characteristics from 09/01/2016 to 09/07/2016 in each clustered region, and then predict the traffic flow next 6 hours using the training model. Secondly, the real traffic flow of six hours is added to the original training data set to establish a new model. Finally, we set the threshold and detect abnormal points by comparing the original and predicted traffic flow data, and verify the outliers with urban events collected from social media. As shown in the Figure 8, this is the result of anomaly detection on the traffic characteristics of Xiamen Convention Center. We can see that there are abnormal points from 09/08/2016 to 09/11/2016 and from 09/23/2016 to 09/25/2016, during which time large-scale commercial exhibitions were held in Xiamen Convention Center, as shown in Table 3.

7.3 Visualization Platform of Crowd Dynamic

The occurrences of urban events often leads to the changes of crowd dynamics, resulting in the abnormal traffic flow in some regions. Firstly, we find the corresponding urban events to the abnormal traffic flow and then count the impacted regions of each urban event by analyzing the abnormal traffic flow points in each region. In order to demonstrate the impacts of urban events on city dynamics more clearly, we have built a visualization platform [34] to display the impacted areas of each urban event and show the traffic flow changes in each area.

As shown in Figure 9(a), this platform is mainly divided into two parts. In the left frame, you can select the urban event, the event date, the number of the regions affected by the event and the flow curve of each region will be displayed below. The map on the right shows the influenced regions of the urban event. As show in Figure 9(b), we can arbitrarily select an event from the event list and a detailed description of the event will appear. At the same time, the affected region of this event will be displayed on the map.

As shown in Figure 9(c) and Figure 9(d), we can select a region from the list of regions influenced by the event. Then the changes of the traffic flow within 22 days and the distribution of abnormal traffic points in this region will be displayed.

8 Discussions

In this section, we discuss several issues and concerns of this work.

- **Area partitioning scheme**. The traffic flow on the road network is closely related to the urban areas around the road, and the traffic flow patterns in different regions show different characteristics. For example, if a concert is to be held in a Stadium, the traffic flow on the surrounding roads of the stadium will increase significantly before or after the competition compared to usual days. Therefore, instead of using road segments as the basic elements to partition the city region, we map road traffic to the corresponding urban areas using grid-based coordination systems. In this way, we aim to reveal the urban structures and dynamics impacted by the intrinsic city functions and points of interests, etc.
- **Grid size**. The size of girds will influence the results of analysis. On the one hand, if the gird size is too small, the efficiency of the proposed method will

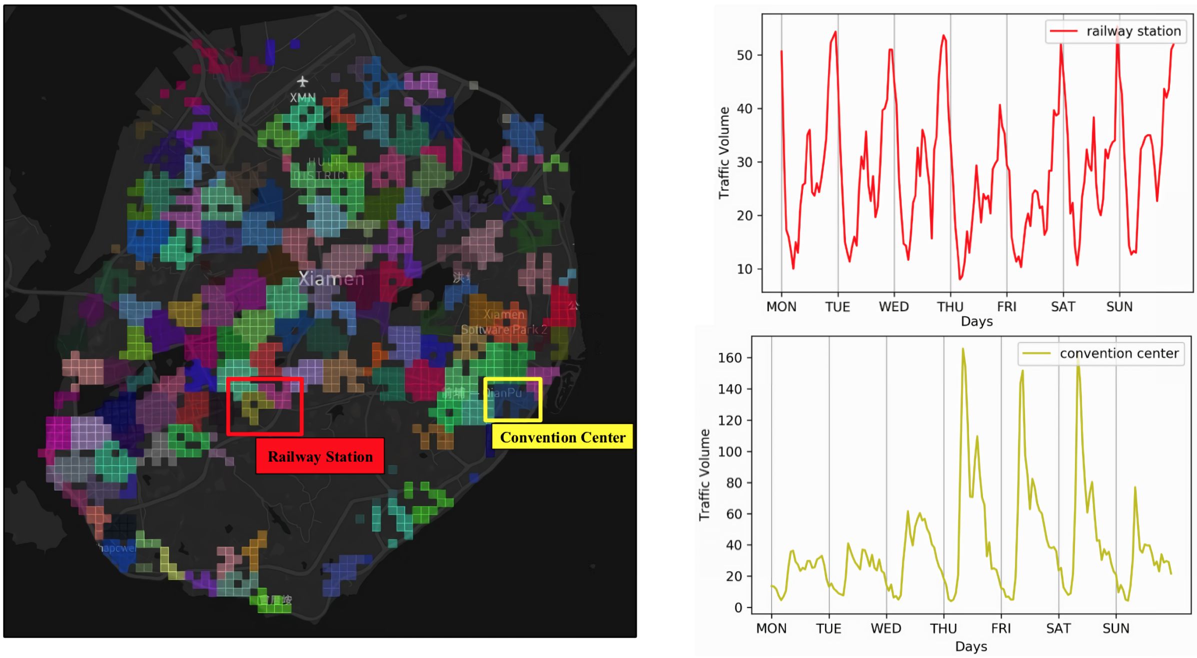


Fig. 7 Illustrative traffic patterns of the Railway Station and the Convention Center formed by the regional grid clustering.

Table 3 Detailed information about the urban events.

Time	Detailed Information
09/08/2016 - 09/11/2016	1. The twelfth China International Investment and Trade Fair
	2. Xiamen Overseas Property Investment and Immigration Exhibition
09/23/2016 - 09/25/2016	1. China (Xiamen) International Leisure Tourism Expo
	2. New Silk Road (Xiamen) International Commodities Fair

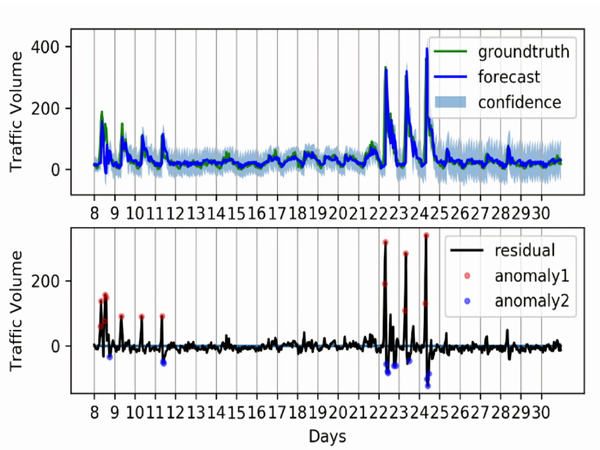


Fig. 8 Anomaly detection in Xiamen Convention Center.

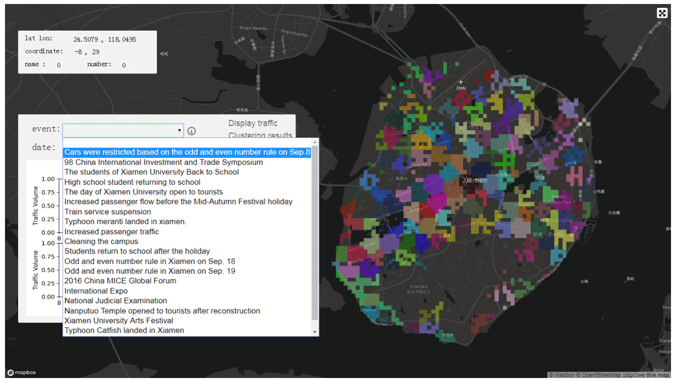
decrease. On the other hand, if the grid size is too large, many valuable information will be missed and the results of the analysis will be meaningless. In fact, this problem is coined as the modifiable areal unit problem (MAUP) [35]. We should take the geographical and cultural environments of the cities into consideration while choosing the grid size in different cities. To this end, we have conducted a series of empirical studies to find the proper grid size. In Xiamen City, the grid size about 200×200 square meters is suitable in this work.

9 Conclusions

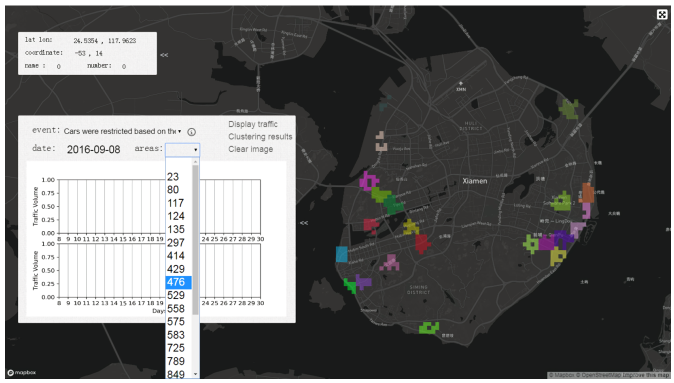
In this paper, we propose a data-driven framework to sense urban structures and dynamics from large-scale vehicle mobility data in a systematic manner. Based on large-scale vehicle sensing data, we extract the regular mobility patterns and exploit their similarities to discover urban area functionalities. The irregular crowd movements are also investigated to analyze urban social and emergency events. Specifically, we propose a distance-constrained clustering algorithm (DCCA) to cluster the grids with similar mobility features into structured urban areas, and leverage an ARIMA-based anomaly detection algorithm (ADAM) to correlate crowd movement anomalies to the urban social and emergency events. Finally, we build a visualization system to demonstrate the urban structures and crowd dynamics. We evaluate our framework using real-world datasets collected from Xiamen City, China. Results show that our approach can sense urban structures and crowd dynamics for urban planning and city management comprehensively and effectively.

In the future, we plan to build a real-time analytics system for the urban crowd dynamics, and involve more intelligent algorithms (e.g., deep learning techniques) to correlate events with anomalies.

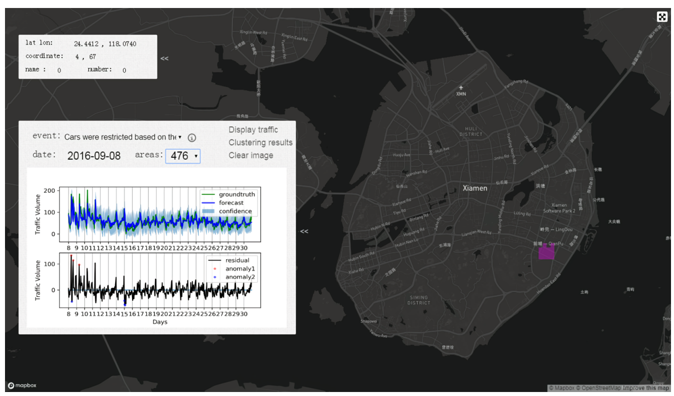
(a) The overview of the system.



(b) Urban events demonstration.



(c) Affected areas demonstration.



(d) Traffic patterns demonstration.

Fig. 9 The crowd dynamics visualization system.

References

1. Zheng Y. Urban computing: Enabling urban intelligence with big data. Frontiers of Computer Science, 2017, 11(1): 1–3
2. Miyazawa S, Song X, Xia T, Shibasaki R, Kaneda H. Integrating GPS trajectory and topics from Twitter stream for human mobility estimation. Frontiers of Computer Science, 2018
3. Wang L, Zhang D, Wang Y, Chen C, Han X, M'hamed A. Sparse mobile crowdsensing: Challenges and opportunities. IEEE Communications Magazine, 2016, 54(7): 161–167
4. Zhang W, Qi G, Pan G, Lu H, Li S, Wu Z. City-scale social event detection and evaluation with taxi traces. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 40
5. Chen L, Jakubowicz J, Yang D, Zhang D, Pan G. Fine-Grained Urban Event Detection and Characterization Based on Tensor Cofactorization. IEEE Transactions on Human-Machine Systems, 2017, 47(3): 380–391
6. Zhang D, Guo B, Yu Z. The emergence of social and community intelligence. Computer, 2011, 44(7): 21–28
7. Chen C, Chen X, Wang Z, Wang Y, Zhang D. ScenicPlanner: Planning scenic travel routes leveraging heterogeneous user-generated digital footprints. Frontiers of Computer Science, 2017, 11(1): 61–74
8. Yuan N J, Zheng Y, Xie X. Segmentation of urban areas using road networks. MSR-TR-2012–65, Tech. Rep., 2012
9. Yang D, Zhang D, Qu B. Participatory Cultural Mapping Based on Collective Behavior Data in Location Based Social Networks. 2016
10. Wang L, Zhang D, Yang D, Pathak A, Chen C, Han X, Xiong H, Wang Y. SPACE-TA: Cost-Effective Task Allocation Exploiting Intradata and Interdata Correlations in Sparse Crowdsensing. ACM Trans. Intell. Syst. Technol., 2017, 9(2): 20:1–20:28
11. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C. Geo-spotting: Mining online location-based services for optimal retail store placement. In: Proc. KDD. 2013, 793–801
12. Chen L, Fan X, Wang L, Zhang D, Yu Z, Li J, Nguyen T M T, Pan G, Wang C. RADAR: Road Obstacle Identification for Disaster Response Leveraging Cross-Domain Urban Data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2018, 1(4): 130:1–130:23

13. Wang J, He X, Wang Z, Wu J, Yuan N J, Xie X, Xiong Z. CD-CNN: A Partially Supervised Cross-Domain Deep Learning Model for Urban Resident Recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence. April 2018
14. Getz D. Event Management & Event Tourism. Cognizant Communication Corporation New York, 1997
15. Chen C, Ding Y, Xie X, Zhang S, Wang Z, Feng L. TrajCompressor: An Online Map-matching-based Trajectory Compression Framework Leveraging Vehicle Heading Direction and Change. IEEE Transactions on Intelligent Transportation Systems, 2019, 1–17
16. Esch T, Schmidt M, Breunig M, Felbier A, Taubenböck H, Heldens W, Riegler C, Roth A, Dech S. Identification and characterization of urban structures using VHR SAR data. In: Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International. 2011, 1413–1416
17. Chen S, Wu H, Tu L, Huang B. Identifying hot lines of urban spatial structure using cellphone call detail record data. In: Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom). 2014, 299–304
18. Cici B, Gjoka M, Markopoulou A, Butts C T. On the Decomposition of Cell Phone Activity Patterns and Their Connection with Urban Ecology. In: Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'15). 2015, 317–326
19. Krumm J, Horvitz E. Predestination: Where Do You Want to Go Today? Computer, 2007, 40(4): 105–107
20. Chen L, Yang D, Zhang D, Wang C, Li J, Nguyen T M T. Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization. Journal of Network and Computer Applications, 2018, 121: 59–69
21. Chen C, Jiao S, Zhang S, Liu W, Feng L, Wang Y. TripImputor: Real-Time Imputing Taxi Trip Purpose Leveraging Multi-Sourced Urban Data. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(10): 3292–3304
22. Li C, Sun A, Datta A. Twevent: Segment-based event detection from tweets. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012, 155–164
23. Liang Y, Caverlee J, Cheng Z, Kamath K Y. How big is the crowd?: Event and location based population modeling in social media. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. 2013, 99–108
24. Yang D, Zhang D, Chen L, Qu B. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. Journal of Network and Computer Applications, 2015, 55: 170–180
25. Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. 2010, 851–860
26. Agarwal M K, Ramamritham K, Bhide M. Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. Proceedings of the VLDB Endowment, 2012, 5(10): 980–991
27. Yu Z, Zhang D, Wang Z, Guo B, Roussaki I, Doolin K, Claffey E. Toward Context-Aware Mobile Social Networks. IEEE Communications Magazine, 2017, 55(10): 168–175
28. Han X, Wang L, Farahbakhsh R, Cuevas Á, Cuevas R, Crespi N, He L. CSD: A multi-user similarity metric for community recommendation in online social networks. Expert Systems with Applications, 2016, 53: 14–26
29. Tostes A I J, de LP Duarte-Figueiredo F, Assunção R, Salles J, Loureiro A A. From data to knowledge: City-wide traffic flows analysis and prediction using bing maps. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. 2013, 12
30. Chen L, Zhang D, Wang L, Yang D, Ma X, Li S, Wu Z, Pan G, Nguyen T M T, Jakubowicz J. Dynamic Cluster-based Over-demand Prediction in Bike Sharing Systems. In: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16). 2016, 841–852
31. Ali S M. Time series analysis of Baghdad rainfall using ARIMA method. Iraqi Journal of Science, 2013, 54(5): 1136–1142
32. Li H, Wu Q, Dou A. Abnormal traffic events detection based on short-time constant velocity model and spatio-temporal trajectory analysis. JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE, 2013, 10(16): 5233–5241
33. Liu F T, Ting K M, Zhou Z H. Isolation-Based Anomaly Detection. ACM Trans. Knowl. Discov. Data, 2012, 6(1): 3:1–3:39
34. Jiang Z, Liu Y. Visualization platform. `https://github.com/longbiaochen/usd`, 2019
35. Wong D. The modifiable areal unit problem (MAUP). The SAGE handbook of spatial analysis, 2009, 105: 23

Zhihan Jiang received the B.Sc. degree in computer science and technology from Xiamen University, China in 2018. She is currently pursuing the master's degree in Department of Computer Science, Xiamen University.

Yan Liu received the master's degree in Department of Computer Science, Xiamen University, in 2018. Her research interests include big data analytics and urban computing.

Xiaoliang Fan is a senior research engineer with Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University. He received his Ph.D. degree at University Pierre and Marie CURIE, France in 2012. His research interests include services computing, and big data analytics.

Cheng WANG received the Ph.D. degree in information and communication engineering from the National University of Defense Technology, China in 2002. He is a Professor with Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China. His research interests include remote sensing image processing, mobile LiDAR data analysis, and multisensor fusion.

Jonathan LI received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000. He is currently a Professor with Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China. He is also a Professor and the Head of the WatMos Lab, Faculty of Environment, University of Waterloo, Canada. His current research interests include information extraction from LiDAR point clouds and from earth observation images.

Longbiao CHEN is an assistant professor with Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University. He obtained his Ph.D. degree in computer science from Sorbonne University, France in 2018 and Zhejiang University, China in 2016. His research interests include Ubiquitous Computing, Urban Computing, and Big Data Analytics.