World Wide Web
DOI 10.1007/s11280-017-0487-4

LoTAD: long-term traffic anomaly detection based on crowdsourced bus trajectory data

Xiangjie Kong[1] · Ximeng Song[1] · Feng Xia[1] · Haochen Guo[1] · Jinzhong Wang[1,4] · Amr Tolba[2,3]

Received: 29 January 2017 / Revised: 27 May 2017 / Accepted: 25 July 2017

Abstract As the development of crowdsourcing technique, acquiring amounts of data in urban cities becomes possible and reliable, which makes it possible to mine useful and significant information from data. Traffic anomaly detection is to find the traffic patterns which are not expected and it can be used to explore traffic problems accurately and efficiently. In this paper, we propose LoTAD to explore anomalous regions with long-term poor traffic situations. Specifically, we process crowdsourced bus data into TS-segments (Temporal and Spatial segments) to model the traffic condition. Later, we explore anomalous TS-segments in each bus line by calculating their AI (Anomaly Index). Then, we combine anomalous TS-segments detected in different lines to mine anomalous regions. The information of anomalous regions provides suggestions for future traffic planning. We conduct experiments with real crowdsourced bus trajectory datasets of October in 2014 and March in 2015 in Hangzhou. We analyze the varieties of the results and explain how they are consistent with the real urban traffic planning or social events happened between the time interval of the two datasets. At last we do a contrast experiment with the most ten congested roads in Hangzhou, which verifies the effectiveness of LoTAD.

Keywords Traffic anomaly detection · Mobile crowdsourcing · Urban big data · Anomaly index

This article belongs to the Topical Collection: *Special Issue on Mobile Crowdsourcing*
Guest Editors: Bin Guo, Xing Xie, Raghu K. Ganti, Daqing Zhang, and Zhu Wang

✉ Feng Xia
f.xia@ieee.org

[1] The Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

[2] Riyadh Community College, King Saud University, Riyadh 11437, Saudi Arabia

[3] Mathematics Department, Faculty of Science, Menoufia University, Shebin El-Kom 32511, Egypt

[4] School of Management and Journalism, Shenyang Sport University, Shenyang 110102, China

Published online: 10 August 2017

1 Introduction

With the rapid development of cities, more and more urban problems appear, such as air pollution, traffic congestion and increased energy consumption, which are urgent for city planners to solve [1]. Nowadays, we enter into the era of data technology where we can figure out solutions by discovering knowledge from big data [9]. Compared with traditional artificial statistics, it solves the waste of resources and high cost issues and it can explore urban information more roundly and effectively. Mobile Crowdsourcing (MCS) is a new sensing paradigm which allows sensor-enhanced devices sharing information to make our city smart [11]. The idea of MCS is significant and applicable in many research areas for including traffic planning, mobile social recommendation, public safety and so on [12, 27, 33]. MCS provides novel thinking to solve problems. With the information explored by MCS, the urban planners have a more efficient and effective plan for the city. Consequently, it is essential for researchers to find methods of discovering and solving urban problems with technique of MCS and data mining effectively.

Nowadays, travel demand exceeds the existing travel capacity in many big cities, which leads to many urban traffic problems and negatively impacts citizens' standard of living [2, 18]. One of them is known as traffic anomaly. At the present stage, many researches focus on detecting traffic anomaly phenomena caused by big events, like the marathon, car accident, etc [23, 28]. Others focus on fraudulent behaviors happened when people travel out [8, 30, 31]. For instance, wicked drivers may detour to earn more money when they carry the passengers who are unfamiliar with the city. These studies mentioned above all focus on digging short-term anomaly from traffic data, which means the anomalous traffic situations they explore only exist in a short time. There are also many long-term traffic anomaly situations in urban area. For example, if a new shopping center is built in a region, then it will attract a host of people visiting the region, which leads to traffic flow increasing. As a result, it may cause traffic congestion, which impacts the citizens' life quality seriously. And this condition will not disappear automatically without proper urban traffic planning measures. Under the circumstances, it requires a method to detect these anomalous regions that have poor traffic situations automatically.

To our knowledge, the researches of traffic anomaly detection (TAD) at the stage all utilize taxi trajectory data to model urban traffic condition. It may lose much information because taxi drivers can choose route for themselves. If a taxi driver gets information ahead and chooses a road he doesn't mean to, we will not tell the actual road traffic situation from the dataset. On the contrary, the bus has a fixed route and the driver won't change route with the traffic condition's variation on road. So the crowdsourced data of bus trajectory can display traffic condition on the road factually. Big cities facing traffic problems have a perfect bus transfer system, so there exist sufficient bus lines and bus numbers overall the city and satisfy citizens' travel demand. In order to realize bus realtime monitoring, buses in big cities are usually equipped with global positioning system (GPS) sensors, which makes the possibility of obtaining mass bus trajectory data.

In this paper, we propose LoTAD (long-term traffic anomaly detection) which can detect long-term traffic anomaly with crowdsourced bus trajectory data. First, we use TS-segments extracted from bus trajectory data to describe the whole city traffic situation from both temporal and spatial aspects. We extract two features from the dataset, average velocity and average stop time which can describe traffic condition and travel demand respectively. Then we excavate poor TS-segments which are the bottleneck of travelling in one line by

calculating their anomaly index (AI). As traffic condition is related to travelling demand around the road, we explore the anomalous regions with the anomalous TS-segments in different lines. These results demonstrate which areas of the city have traffic trouble and provide city planners with advice of operational planning. Finally, we use real urban planning or social events in reality to verify the effectiveness of our method. We also compare our method with three contrast methods and we evaluate these two methods using three evaluation metrics.

Our major contributions are described as follows:

- We propose LoTAD to find long-term traffic anomalous regions in urban city. TS-segments are extracted from bus trajectory data to describe the traffic condition around the city. We then partition the urban road network into road segments based on the busline data and we partition the dataset by time slots. we get the TS-segments, which can describe the real traffic situation all around the city. Poor TS-segments in each bus line are explored by calculating their AI. Then we combine these poor TS-segments in different lines together to explore the anomalous regions.
- We propose a novel method to partition the whole city based on the transportation stations, which can divide regions rationally according to the travel demand. The method considers both the traffic condition and urban development and provides significant information for city planners.
- We evaluate our method with real bus trajectory data in Hangzhou. For the results, we verify the effectiveness of our method using real urban planning like new subway lines coming into use. And we use the most congested ten roads in Hangzhou to evaluate our method. We compare our method with three methods which are LOF, skyline and quartile deviation with three metrics including recall, precision and F1 score. The result shows the effectiveness and superiority of our method.

The rest of our paper is structured as follows. In Section 2, we introduce the related work about traffic anomaly detection. Our method is introduced detailedly and roundly in Section 3. Section 4 describes our experiment settings and results to verify the effectiveness of our method. In Section 5, we conduct a contrast experiment and compare it with our method. At last, in Section 6, we conclude our work and talk about future work.

2 Related work

Our work is inspired by the previous work of MCS and TAD. We take advantage of the two areas to work out the problem of long-term traffic anomaly detection.

2.1 Mobile crowdsourcing

With rapid advance in devices with embedded sensors, MCS becomes feasible and available now. Many researchers focus on the common problem of assigning tasks in MCS. Feng et al. design a mechanism called truthful auction which takes the crucial dimension of location information into consideration [10]. Guo et al. proposed a dynamic and quality-enhanced incentive mechanism [13] and multi-task allocation has been studied [14]. Li et al. investigate personalized influential topic search with crowdsourced data from social network [22]. MCS has been applied in a variety of applications in urban life, such

as environment monitoring, healthcare, location services and so on [11]. Transportation and traffic planning are crucial in MCS. The availability of spatial information from floating cars on road, location data from people's smartphone, video records of roads and so on provides opportunity to study urban transportation with MCS. B-Planner used crowdsourced GPS data from taxis to plan the routes of night-bus [7]. Time-location-relationship [19] combined model has been used to promote taxi service with taxi trajectory data. Liu et al. propose a bilevel optimization model to optimize the distribution of public electric vehicles across the city with trajectories of taxis [24]. Wolfson et al. proposed T-Share which can generate optimized ridesharing schedules based on crowd-powered data [25]. In our paper, we use the crowdsourced GPS data from bus spreading all over the city which can probe the traffic condition and we aim to detect traffic anomalous regions.

2.2 Traffic anomaly detection

Anomaly detection is to find the patterns which are not expected from data [5], which belongs to the research area of data mining. With the variation of data used, anomaly detection can draw different conclusions and it can be used in many applications, like system diagnosis, biological mutation, and user anomalous action detection [15]. With the availability of big traffic data, many researches focus on finding anomalous traffic patterns in urban city, which may explore and solve problems we encounter in our real life.

Researchers can get the taxi trajectories easily with the obtaining of GPS data. Many researches aim to find meaningful patterns and develop valuable applications with these data. Based on two applications: detecting taxi driving frauds and road network changing, isolation-based anomalous trajectory (iBAT) [31] is proposed to discover anomalous driving patterns. And isolation-based online anomalous trajectory (iBOT) [8] improves the accuracy and real-time performance of iBAT. With taxi GPS trajectory data and crowd sensing data generated by smart phones, Jin et al. [17] develop a system called crowdTPR to find flagged taxis' passenger refusal behaviour in real time using a dynamic grid granularity selection method. Many researches partition the city into small regions and process traffic data into orientation-destination (OD) matrix using a statistical method to model the traffic situation. Many studies are carried out with this kind of data. Chawla et al. [6] mine the root cause of anomalies and propose a two-step mining and optimization framework which use the idea of principal component analysis (PCA). Zheng et al. [34] detect flawed urban planning using GPS trajectories of taxicabs. Kuang et al. [20] improve the performance of detecting traffic anomalies with PCA. In addition, they combine their method with wavelet transform and get more effectively results. Pang et al. [28] use an efficient pattern mining approach with likelihood ratio test statistic method to detect spatio-temporal outlier and monitor the anomalous traffic condition accurately and rapidly. Furthermore, they aim to improve traffic condition in advance. Huang and Wu [16] propose road segment-based outlier detection method to avoid the boundary problem and their method aims to find all outliers in the road network.

Our paper differs from the researches above in a few aspects. First we use the crowdsourced bus trajectory data in urban city and transfer it into TS-segments to model the whole city traffic situation, differing from the taxi trajectory or OD flow matrix in existing researches. Bus has its own route so the trajectory will not change according to the drivers' preference. Nowadays, bus lines are running through big cities. So with bus trajectory data

Table 1 Bus station-line dataset

Item	Format	Comment
LineID	Integer	Unique ID of a line
Direct	1 or 2	Direction of a line
StationIndex	Integer	Order of station in one line
StationName	String	Station's name
Longitude	Number	Longitude of bus station
Latitude	Number	Latitude of bus station
StartTime	Time	First vehicle hour in one day
EndTime	Time	Last vehicle hour in one day
TimeStamp	Date	Update time of the dataset

we can probe the city wide traffic situations factually. Second we aim to find long-term anomalous traffic phenomena which may exist for a long time and the traffic condition will not recover unless implementing urban planning, such as new subway line coming into use and new functional district region building. It differs from the traffic anomaly detection discussed above which is caused by short-term event like drivers' temporary changing route or some accidents. Information explored from our long-term traffic anomaly detection provides suggestions and directions to urban planners. Therefore, it is significant and practical.

3 Dataset

3.1 Dataset description

In our experiments, we use two datasets in Hangzhou. One is the bus station-line dataset whose format and comment are shown in Table 1. It contains the information of relationship of lines and stations and location of stations. The dataset is used to divide the road network to road segments and find stop points in Hangzhou.

The other datasets we use are the bus GPS dataset in Hangzhou whose format is shown in Table 2. This data contains temporal and spatial information of bus and is used to extract bus trajectories.

Table 2 Bus GPS dataset

Item	Format	Comment
LineID	Integer	Unique ID of a line
Numb	Integer	Unique ID of a bus
Longitude	Number	Longitude in bus GPS
Latitude	Number	Latitude in bus GPS
Time	Time	Time of the GPS generated

3.2 Dataset preprocessing

The original datasets we get are lack of accuracy. So we first clean the data by removing duplicate and dirty data such as the item whose GPS location is out of Hangzhou. Especially, we focus on the central urban areas in Hangzhou, which are located at 120.0905° E to 120.3155° E and 30.1115° N to 30.3565° N. As people's travel patterns differ between weekday and weekend, we separate the two time span's data. We choose crowdsourced bus trajectory data in October of 2014 and March of 2015 with a fixed scope. The reason that we choose October and March for our experiments is that the weather condition in these two months are similar, which means the travel behaviors and travel demand in cities resemble. It makes our results more impartial and meaningful. In order to make the data in two months on balance, we leave out the data in lines which only exist in one of the months. Finally the attributes of two datasets are shown in Table 3. As stations of two directions in one line may be different, so we regard two directions in one line as two different lines. The original dataset is disordered and we classify the data by busline ID and bus ID. Then we try to extract the items whose GPS location are in the range of bus stations and extract the items whose travel time is between 5:00 and 21:00 to get the bus trajectories between stations. We discard the trajectories whose velocity is less than 1m/s or stop time is over 600s at a station, which happens with an extremely low possibility in real life. This kind of data occurring is always by devices fault, so it needs to be cleaned out.

4 Method of LoTAD

4.1 Overview

4.1.1 Method overview

Figure 1 shows structure of LoTAD which will be introduced in this section generally. First we extract TS-segments from crowdsourced bus trajectory data and bus station-line data by time and space dividing. Two features are extracted from each TS-segment including average velocity which models the real traffic condition on the road and average stop time which represents for the passenger flow volume of the bus. Then we find the anomalous TS-segments by calculating the AI of each TS-segment according to its features. We explore traffic anomalous regions by a novel method of region partition and mapping the anomalous TS-segments to the corresponding region. We have a detailed description of these parts in the following subsections.

Table 3 Symbols used in the method

Dataset	2014.10	2015.03
Size	13.1G	8.37G
Days	31	31
lines	230	230
Bus Numbers	3812	6402
TS-segment number	95243	117542

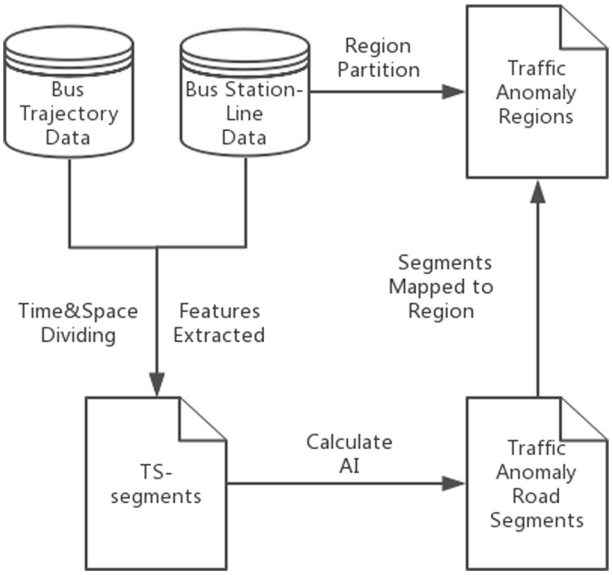


Figure 1 LoTAD overview

4.1.2 Symbols overview

We create a table of symbols used in this paper shown in Table 4 for better understanding of our method.

4.2 Traffic modeling

4.2.1 Time and space dividing

We use the bus station-line data to partition the urban road network into road segments. Two adjacent stations which the bus will run sequentially between them and the route bus

Table 4 Information of two cleaned datasets for evaluation

Symbol	Description
t_k	Time slot
tra	Bus trajectory
$tras$	Bus trajectories which are belonged to one TS-segment
s_M	Manhattan distance
$\bar{v}$	One feature of TS-segment: average speed
$\bar{st}$	One feature of TS-segment: average stop time
M	A matrix which contains information of a TS-segment
d_c	Cut-off distance
t	A parameter in ADAI which identifies d_c
AI	Anomaly index of a TS-segment
$reAI$	Anomaly index of a region

travelling between the two stations are partitioned into a road segment. The position of a station is signed by its geographical coordinates. A road segment can be identified uniquely by the bus lineID, line direction and station number. In this step, we need to extract TS-segments from bus trajectory dataset. With the results of road network partition, the bus trajectory can be divided spatially. First, we search the stopping items whose location coordinate are in the range of the line's station area. Then according to the lineID and station number of the item, we find items who have adjacent stations in one line. The item pairs we get are the results of partition from spatial aspect.

Then we use (1) to have the temporal partition of bus trajectory,

$$t_k = [k\theta, (k+1)\theta), k = 5, 6...21 \tag{1}$$

where t_k is the number of time slots and θ is the duration of each time slot.

As the bus will not travel in a whole day, we only choose the time interval from 5:00 to 21:00 for our experiments, which fit most bus lines' timetable. And here we choose one hour as a time slot. As we know, the travel patterns have similarities at the same time of one day, so we only consider the temporal partition in one day ignoring the date varieties. But it is common that people have different daily routines on weekdays and weekends which makes the human travel patterns different, so the traffic condition will change too. So here we take the time of weekdays and weekends into considerations separately. A TS-segment ts is identified uniquely by lineID, line direction, station number, time slot number and it also has two variables: $loc1$ and $loc2$, which stand for the position of two stations in one TS-segment.

4.2.2 TS-segment features extracted

We define that a trajectory tra is a list of bus GPS points with chronological order as $p1, p2...pn$ between two adjacent stations. Any adjacent GPS points like $< p1, p2 >$ and the bus route between them make up an edge. Each TS-segment has lots of trajectories because there are many days in the dataset and there are numerous buses crossing a TS-segment in the same time interval. The trajectories of a TS-segment make up a set $tras$. We extract two features for each TS-segment: average velocity and average stop time. The speed of one trajectory is calculated as follow:

$$v = \frac{tra.s_M}{tra.t} \tag{2}$$

where $tra.s_M$ is the Manhattan distance [26] between the adjacent two stations of a TS-segment and $tra.t$ is the time a bus traveling between stations. Manhattan distance is also called taxi distance, which is widely used to calculate the distance of car driving or person walking on the road. Since the road in cities are straight from east to west and north to south generally, Manhattan distance is roughly equals to the driving distance.

Average velocity $\bar{v}$ is the expectation of all the trajectories' speed, which is related to the velocity and length of each trajectory. As the length of each trajectory is different, here we use an aggregating method putted forward by Zhang et al. [32]. They use this method to

estimate traffic state, which meets our requirements. Then we calculate the average velocity of the TS-segment as (3).

$$\bar{v} = \frac{\sum_{tra_i \in tras} tra_i.v * tra_i.w * g(tra_i.v)}{\sum_{tra_j \in tras} tra_j.w * g(tra_j.v)} \tag{3}$$

$$g(tra_i.v) = tra_i.v^{0.9 - tra_i.w} \tag{4}$$

where $tra_i.v$ is the velocity of trajectories of $tras$, $tra_i.w$ represents weighted coefficient which is equal to the ratio between the travelled length of the edge and the sum of a TS-segment's all edges' length, the constant 0.9 is an empirical number obtained from previous work.

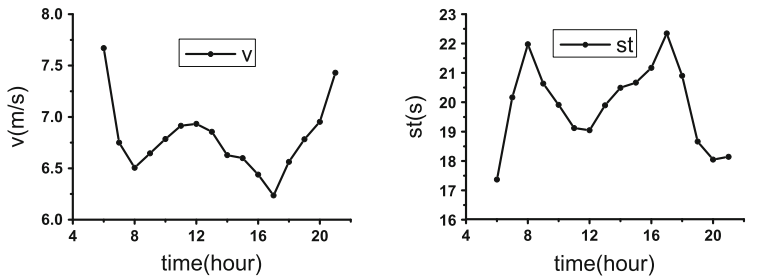
After we extract the trajectory items which mean the bus stay at a station, we calculate the stop time of the adjacent stations st_1, st_2 easily by subtracting the time in trajectory items. Then we can use (4) to calculate the average stop time $\bar{st}$.

$$\bar{st} = \frac{\sum_{tra_i \in tras} \frac{tra_i.st_1 + tra_i.st_2}{2}}{|tras|} \tag{5}$$

So we get the features of a TS-segment: $< \bar{v}, \overline{st} >$. Figure 2 shows the average $\bar{v}$ and $\overline{st}$ of all the ST-segments in every time slot of our dataset. The two subgraphs show the opposite variation trend. We can see that the average $\bar{v}$ come to its low ebb around 8:00 a.m and 17:00 p.m. Meanwhile the average of $\overline{st}$ reach its peak at the same time period. $\bar{v}$ represents the driving status on the road and $\overline{st}$ reflects travel demand. The bigger $\overline{st}$ is, the more people are willing to travel out in this TS-segment. By analysing these two features from Figure 2, we find that traffic condition is terrible during these two time slots, and it conforms to our acknowledgement that the periods are the daily travel peak time. TS-segments can model the city-wide traffic condition from both time and space, with features $\bar{v}$ and $\overline{st}$ describing the traffic condition factually and appropriately.

4.2.3 Bulid TS-segment matrix

In this part, we formulate a matrix M for each bus line, as demonstrated in Figure 3. Each item in the matrix presents a TS-segment which is a tuple here. a_{ij} is the expression of



(a) average $\bar{v}$ of all TS-segments (b) average $\overline{st}$ of all TS-segments

Figure 2 TS-segment features

Figure 3 Matrix of TS-segments

$$M = \begin{array}{c} \\ r_0 \\ r_1 \\ \dots \\ r_m \end{array} \begin{array}{c} \begin{array}{cccc} t_0 & t_1 & \dots & t_n \end{array} \\ \begin{pmatrix} a_{00} & a_{01} & \dots & a_{0n} \\ a_{10} & a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{m0} & a_{m1} & \dots & a_{mn} \end{pmatrix} \end{array}$$

$< \bar{v}, \overline{st} >$, denoting the average velocity and average stop time at road segment r_j during the time slot t_i. Since the units of velocity and stop time are different, we do max-min normalizing of values in a_{ij}. Assuming there are x bus lines in our dataset, then there will be x matrices built at last. These matrices are treated as input data to find the anomalous TS-segments.

4.3 Anomalous TS-segments detection

In this part, we try to explore anomalous points of the TS-segment matrix. Anomalous TS-segments are the bottleneck segments which effect peoples travel behavior negatively in a long-term. Anomaly means the points which are different from the others. We use the two features which are $\bar{v}$ and $\overline{st}$ of a TS-segment to find the anomalous TS-segments by calculating its AI in each line. In the experiment part, we use the matrices M built above to do calculation.

We propose a novel method to do anomaly detection of calculating anomaly index of each point(ADAI). AI represents the degree of a points anomaly. Our work is on the basis of the idea of a distance based anomaly detection method called local outlier factor (LOF) [3]. But LOF has some characters, which make it not applicable to our problem. First, LOF does anomaly detection by calculating local outlier factor of each point which means it can find both global and local anomalous points. In our problem, we only need to find the global outliers. Second, a parameter k must be identified previously to indicate the range of local area. The value of k affects the result a lot and it is an empirical value which varies with different datasets. So we can't tell the value of k automatically by program. Third, we can't tell which points are anomalous, because the boundary of LOF value is hard to tell.

To make algorithm be suitable for our problem, we propose ADAI, whose pseudo-code is as following. To solve the first and second problem, we compare the density of a point to all the other points using gaussian kernel function to calculate points' density [21]. Gaussian kernel function represents a point's neighbouring points' number and their relative distance. And we get a density result of continuous numerical, so it reflects a point's density properly. Then we use a numerical analysis method to find breakpoint of AI to solve the last problem. Given matrix of TS-segments M, parameter t, ADAI explore list of anomalous TS-segments $AIList$. t is the parameter to identify cut-off distance d_c. Here we follow the rule of thumb which the value of d_c makes the average number of neighbors is around 1% to 2% of the total number of points in the dataset [29], which can be determined by parameter t.

Algorithm 1 ADAI algorithm

input: M, t
output: $anomalyList$:

for point i in M **do**
  for point j in M **do**
    if $i < j \&\& i != j$ **then**
      $distanceArray.add(distance(i, j))$;
    end if
  end for
end for
$sort(distanceArray)$;
$M \Leftarrow distanceArray.size()$;
$d_c \Leftarrow distanceArray.get(Mt)$;
for point i of M **do**
  $density(i) \Leftarrow \sum_{j \in M \setminus \{i\}} e^{-(\frac{d_{ij}}{d_c})^2}$;
end for
for point i of M **do**
  $AI(i) \Leftarrow \frac{\sum_{j \in M} \frac{density(i)}{density(j)}}{M.size}$;
  $AIList.add(AI(i))$;
end for
$sort(AIList)$;
for value i in AIlist **do**
  $divFactor = \frac{AIlist.get(i+1) - AIlist.get(i)}{AIList.get(i)}$;
  $divFactorList.add(div)$;
end for
$avgDivFactor \Leftarrow averagevalueofdivFactorList$;
for value i in divFactorList **do**
  $sumDiv+ = i$;
end for
$divSize \Leftarrow divFactorList.size$
$avgDivFactor \Leftarrow \frac{sumDiv}{divSize}$;
for value i in AIlist **do**
  $divFactor = \frac{AIlist.get(i+1) - AIlist.get(i)}{AIList.get(i)}$;
  if $divFactor > avgDivFactor$ **then**
    $breakPoint = i$;
    break;
  end if
end for
for $i = breakPoint, i < AIlist.size, i++$ **do**
  $anomalyList.add(l.get(i))$;
end for
return anomalyList ;

For the two features of a segment, $\overline{v}$ reflects the traffic congestion on the road and $\overline{st}$ reflects travel demand of the segment. After the traffic anomalous TS-segments detecion, we have three kinds of anomalous TS-segments shown in Figure 4. Figure 4 shows all the

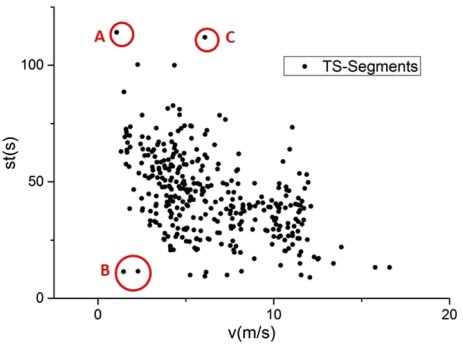


Figure 4 An example of TS-segments

TS-segments in one line and a point stands for a TS-segment. The figure shows how the TS-segments distribute in the dataset. First, we have the anomalous TS-segments shown as point A who has a smaller $\overline{v}$ and a bigger $\overline{st}$. It means the road is congested and the travel demand in this segment is high. It indicates the traffic and road capacity is not enough for the travel demand and it may help to build new roads or subways to relieve traffic pressure. Second, TS-segments shown as point B who has a smaller $\overline{v}$ and a smaller $\overline{st}$ are also detected. It means the road is congested and travel demand in this segment is low. This situation happened when people are willing to drive their cars or the road is narrow. It will work if the government try to improve the environment and the convenience of public transit system. Building more roads will also work. Finally, we detect TS-segments shown as point C who has a bigger $\overline{v}$ and a bigger $\overline{st}$. It means the travel demand is high in this region and more public transport services are needed.

To classify the three kinds of anomalous TS-segments above, we use the original k-means algorithm to cluster the anomalous points. There exist four kinds of points in the output of ADAI. ADAI finds anomalous TS-segmets which are different from the most segments, including defected ones shown as A, B, C in Figure 4 and also the perfect one. So at last we have four different kinds of points, three bad and one good of traffic condition. After do k-means with parameter four, we remove the better segments from our result. At last, we explore the anomalous TS-segments and they are the travel bottleneck of each bus line which negatively effect travel efficiency. If the travel problem of these key segments can be settled, then the traffic condition in each line will improve a lot.

4.4 Traffic anomalous regions detection

After exploring anomalous TS-segments in each line, we have the outlier collection of TS-segments. In this paper, we aim to explore the anomalous traffic regions in city for two major reasons. First, hundreds of bus lines exist in big city and these lines may have connections with each other like sharing same stations, so it is necessary to consider the TS-segments in one area comprehensively. Second, traffic situation on road tends to be associated with the number of cars apparently. But the root cause is the variation of the nearby regions' function, which leads to the travel demand increasing or decreasing. For example, if a WanDa Square which is the most famous district chain in China opens in a region, then the travel demand in this region will increase rapidly. It results that traffic capacity near the region will not be sufficient to satisfy the increased travel demand and it may cause traffic problems. So it is essential to detect anomalous regions on map, which may provide significant information for urban planners.

4.4.1 Region partition

First, we should find a proper way to partition the whole city. Public transportation is established to make convenience to citizens' travelling, so it represents travelling demand in an area. A common and easy way to divide map is using grids, but this method can't satisfy our requirements. Traffic density is different around the city, so it is unreasonable to partition the city into same area. In the center of city, the traffic condition is more complex and needs to be analyzed meticulously. On the contrary, urban fringe is supposed to partition into bigger areas. Here we propose a method to partition the city into small regions according to the placement of the station of bus and subway. Since some of the stations are located in the same place or the distance among the stations is short, we first cluster the stations according to the stations' location. The clustering centers are treated as traffic center of regions. K-means is a famous clustering algorithm, but it has two drawbacks if we use it directly to solve our problem. First, if we use K-means directly, we have to know the value of k previously, which is the number of centers. But the value can't be confirmed. Second, initialization of cluster center randomly makes the cluster result variable. So in this paper we use an improved algorithm based on K-means. The process is shown in Figure 5. The input of algorithm is the stations of public transportation as nodes. With these nodes we do the improved cluster algorithm and get the cluster center as output. In this algorithm, we choose the two farthest nodes as the initial centers of k-means which can solve the previous two problems. Because we can't tell the value of k previously, we choose the minimum number

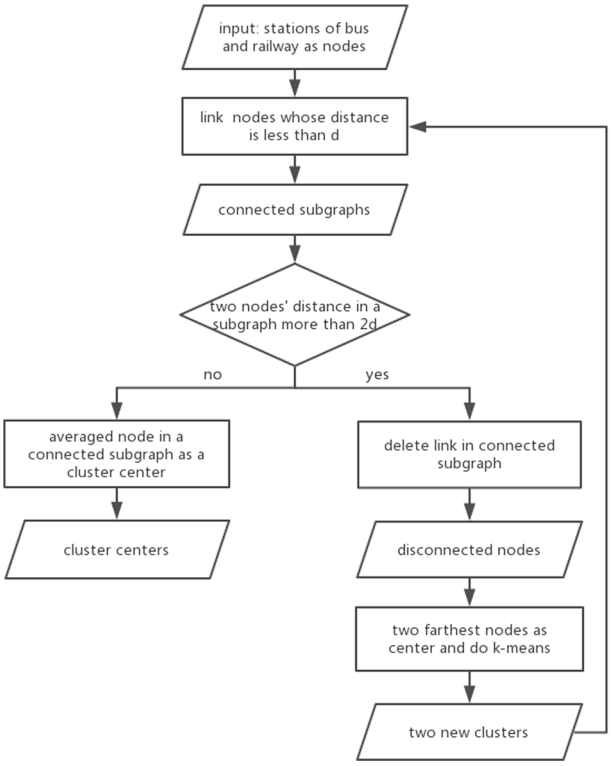


Figure 5 Explore region centers by improved k-means

two as the initial k and do the loop. Finally, the algorithm converges with the value of d and obtain clusters. In cluster results, the farther of two nodes are, the higher of possibility that the two nodes are in different clusters. So the algorithm chooses the farthest nodes as initial cluster center. In the algorithm, we have a new parameter d which can be determined by travel requirement. In daily life, people go to the nearest station to take a bus by walking, so the range of d is the distance of a human walking in 5 minutes to 10 minutes which is about 400 meters to 800 meters.

After determining center of each region, we try to explore areas' boundary. In daily life, people are willing to choose the nearest placement to travel out and it conforms to the idea of Voronoi plane partition [4]. So we also use the clusters to define small areas by Voronoi. In our method, each center represents a small region and placements in this region are closest to the center. At last, we partition the whole city conforming to traffic condition. In our method, center of the city is divided to smaller regions and urban fringe is divided to larger regions, which helps the analysis of traffic situation.

4.4.2 Traffic anomalous regions explored

Based on the results of anomalous segments and region partition, we try to explore the traffic anomalous regions. The segments' AI explored above is in a single line, so here we consider the difference between lines. A segment can go through a few regions, so we use trajectories between two stations to allocate segments' region-AI to different regions. And we accumulate them to calculate regions' anomaly index as $reAI$. At last, we calculate $reAI$ of each region in (6, 7).

$$re(ts) = \{ts | ts \in O(ts) \& ts.tra \in re\} \tag{6}$$

where $ts.tra \in re$ means that the bus trajectories of TS-segment go through the area.

$$reAI(re) = \sum_{ts \in re(ts)} \frac{ts.AI * ts.ratio * ts.st_l}{ts.v_l} \tag{7}$$

where st_l and v_l is the average stop time and average speed of all the TS-segments in the line which is the TS-segment belonged to. These two factors aim to reflect the difference of traffic situation among different lines. $ts.ratio$ equals to the ratio between length of trajectories passing the area and the distance of the two stops in a TS-segment.

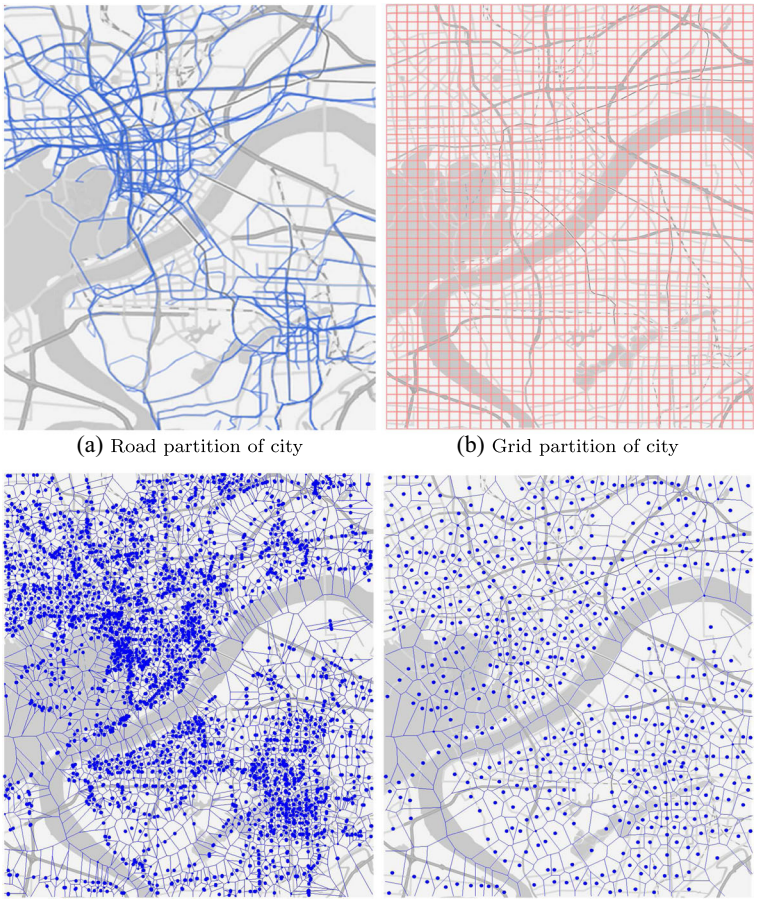
$reAI$ reflects the traffic condition of each small area in the whole city and the traffic condition is worse with the rise of $reAI$. Regions with a big $reAI$ are eager for valid traffic planning measures. So we can find regions of the city which have serious traffic problem and are urgent for city planners to resolve from the results.

5 Experiments and analysis

In this section, we give the results and analysis of our experiment. With some real urban planning and social events in Hangzhou, we prove that LoTAD is effective.

5.1 Region partition

Figure 6 shows four ways to do region partition. Figure 6a shows the bus lines we use to divide the urban traffic road network. We can see that the bus lines go through the major



(a) Road partition of city (b) Grid partition of city

(c) Region partition with original stations (d) Region partition with clustering points

Figure 6 Ways of dividing city spatially

district of city. And it can be seen from the diagram that city line distribution density is different overall the city. So the traffic density is different and it is rational to analysis traffic status in a small region with high traffic density and large region with low traffic density. Figure 6b is the result of meshing area which is used frequently. The whole city area is divided into rectangle areas with same size. Through the analysis of the original lines, we can find that this way is not reasonable. For example, traffic is dense in the downtown area, so it requires a more detailed analysis within a smaller area. On the contrary, traffic flow is more sparser in tourist attraction areas and small range of traffic detection is meaningless. So they need analysis on a larger scope. Figure 6c is the result which is based on the original. From figure it can be seen that some areas are too dense to do analysis and display. Figure 6d is the result of our method. The parameter d is chose as 400m. Regions near the city center around the West Lake, Xiaoshan business area are divided into smaller regions. On the contrary, regions around the Qiantang River, scenic spot like Xixi Wetland are divided into larger regions. These conform to the traffic condition in reality. From the experimental results and comparison chart we can see, the proposed method of transportation

center clustering do regions partition according to the actual traffic distribution and travel demand. At the same time, the station redundancy problem can be effectively solved.

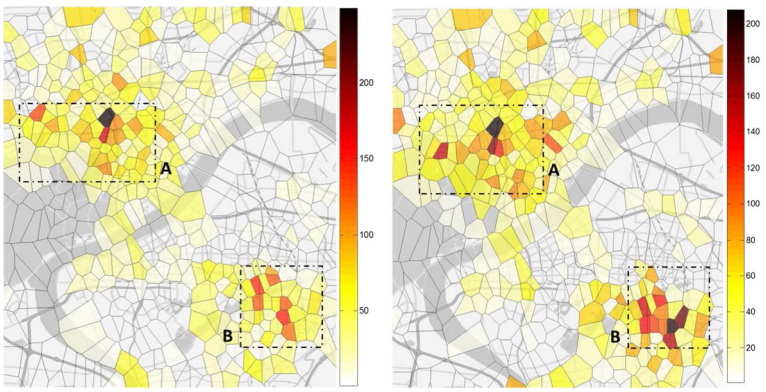
5.2 Traffic anomaly regions detection results

Figures 7 and 8 show our results of weekday and weekend respectively. And anomalous regions we explore of the datasets of two month are shown in the figures. The color represents the $reAI$ of the small regions we divide, which mean that anomalous regions have a deeper color are having a worse traffic condition. In our experiment, 17 time slots are generated and a time slot is one hour. In the ADAI algorithm, the parameter t is chose as 2%. The running time of LoTAD on bus data of 2014.10 and 2015.03 are 41.319s and 53.711s respectively. It is clear from the two results that there are mainly two parts in Hangzhou that have traffic serious problems. The first part A contains Xihu district and Xiacheng district. Many downtowns like Wulin Square, schools like Zhejiang University, Zhejiang province and Hangzhou government buildings and hospitals are located in this region, and this is also the old district of Hangzhou. The other region shown in part B is around the Xiaoshan commercial city. In this region there are also many office buildings, hotels and banks. The function of these regions results in the amount traveling demands acquiring in these two regions, which make the traffic condition here worse. Urban planners should pay more attention to these areas' traffic condition. Traffic condition is also different between weekday and weekend, which is caused by human travel behavior.

Figure 9 shows the traffic condition differences between 2014.09 and 2015.03. The color blue represents that the traffic condition in 2015.03 is becoming better compared with 2014.09. On the contrary, the purple regions have worse traffic condition. The variety may cause by some social events or traffic administrative measures. Next we will explain our results locally.

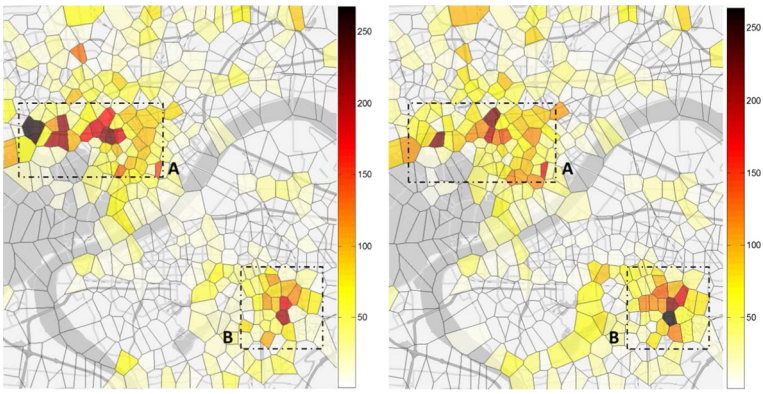
5.3 Analysis results locally

In this part, we will analyse the results of anomalous regions contrasted using some social events and urban planning such as green subway line coming into use in the future to verify our idea.



(a) Anomalous regions of 2014.10 (b) Anomalous regions of 2015.03

Figure 7 Anomalous regions of weekday

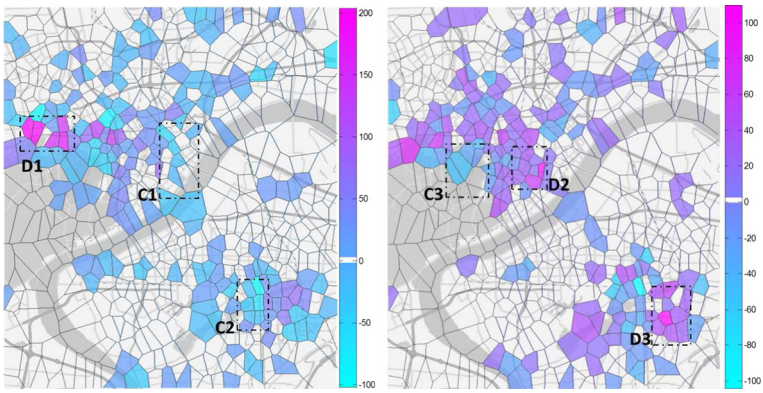


(a) Anomalous regions of 2014.10 (b) Anomalous regions of 2015.03

Figure 8 Anomalous regions of weekend

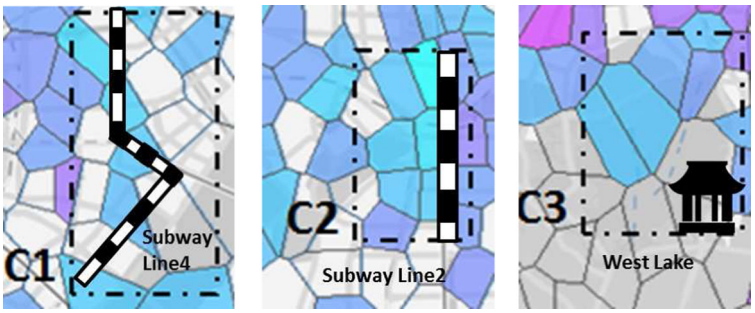
(1) We analyse regions which have a color of blue on the map which means that some events happened in the region leading the traffic condition relieving. In Figure 10a we can see that most small regions get a better traffic condition with the subway line 4 coming into use. Subway line 4 in Hangzhou is coming into service on 2015.02.02, which just happens between the time interval of our datasets. Before the running of subway line 4, people travelling out with public transport in this area have to choose bus or taxi, which makes great pressures to the traffic condition on the road. After the subway line 4 running, more people would like to choose this faster and more comfortable way to travel out, which makes the regions around the new subway line have a better traffic condition. The same reason is for regions around subway line 2 which is opened on 2014.11.24 and is shown in Figure 10b.

Another example is given in Figure 10c. The area is the famous scenic spot in Hangzhou called West Lake, which attracts many tourists to visit every year. The National day in China is in October when people have a seven days long holiday and most people may choose to have a journey. So the travel demands around West Lake increase during this period, making



(a) Contrast results of weekday (b) Contrast results of weekend

Figure 9 Anomalous regions contrasted of the two months



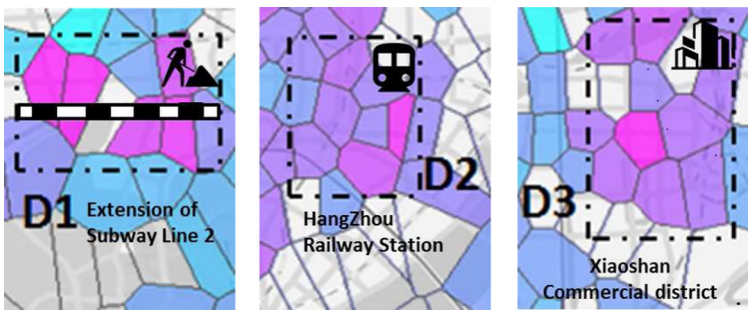
(a) Regions around subway line 4. (b) Regions around subway line 2. (c) Regions near West Lake.

Figure 10 Regions have better traffic condition in 2015.03

the traffic condition worse in October. In contrast, March is an ordinary month for West Lake. So it makes sense the regions around West Lake are blue in our results.

(2) Regions which have a worse traffic condition in 2015.03 than in 2014.09 or the regions both have poor traffic situation during the two months are analysed. An example is presented in Figure 11a. In this region, the traffic condition is becoming worse in 2015.03. And we can see that the urban planners also realize this problem. As shown in the graph, the extension of subway line 2 is now under constructing in this area in order to relieve traffic pressure on road. And the constructing of subway takes up a part of the road, which is another reason makes traffic condition worse here.

In Figure 11b we can see the traffic condition near Hangzhou Railway Station is becoming worse in 2015.03. This is resulted in the most important festival Spring Festival in China, which was in February 19 in 2015. It is traditional for Chinese to go hometown during Spring Festival and return home afterthe holiday. So the railway station is expected to be very busy in March. March is likewise the time when college opens and many colleges like Zhejiang University are situatedin Hangzhou. These two events both lead to the increasing passenger flow volume near the railway station, making our result reasonable.



(a) Regions around extension of subway line 2. (b) Regions around Hangzhou railway station. (c) Regions around Xiaoshan district.

Figure 11 Regions have worse traffic condition in 2015.03

Figure 11c shows a commercial district in Xiaoshan. In this area, there exist many building supply companies. Many people choose to spruce up their house in Spring, which make the travel demand in this area increase and traffic condition become worse.

In this section, we demonstrate the anomalous regions we detected. By analysing the reason of traffic condition variation in October and March, we verify the effectiveness of our method.

6 Performance evaluation

We evaluate LoTAD based on monthly traffic situation reports from comprehensive transportation research center of Hangzhou. The reports investigate the most congested ten roads of every month in Hangzhou (http://hznews.hangzhou.com.cn/chengshi/content/2015-04/09/content_5721813.htm and http://ori.hangzhou.com.cn/ornews/content/2014-11/10/content_5520842.htm). So we pick up the reports of corresponding month to evaluate our results. We map the congested road to our small regions with the roads' geographic position and get the congested regions.

Table 5 represents a region's possible result with the evaluation. Detected means a region is detected as an anomalous region with the algorithms and congested means that a region is in the range of the ten congested roads.

We use the following three metrics which are Recall Rate, Precision and F1 Score to evaluate our proposed method.

– Recall Rate. It is used to show the rate of congested regions we detect.

$$R = \frac{TP}{TP + FN} \tag{8}$$

– Precision. It represents the rate of congested regions in our detected regions.

$$P = \frac{TP}{TP + FP} \tag{9}$$

– F1 Score. It is the comprehensive result considering both Recall Rate and Precision.

$$F1 = \frac{2PR}{P + R} \tag{10}$$

We also do three contrast experiments which are LOF, skyline and quartile deviation. We improve algorithm LOF and propose ADAI, so we also evaluate the performance of LOF. Skyline is proposed by Zheng et al. [34], which used to find defected traffic areas of the city. The skyline is a collection of the points which are not dominated by any other point. Quartile deviation is used widely to find anomalous points. In our contrast method, we use the division of average stop time and average speed as the input of quartile deviation and explore anomalous TS-segments. The results of the three metrics of the two datasets are shown in Figures 12 and 13 respectively. We rank the anomalous region with their

Table 5 Possible results

	Congested	No Congested
Detected	TP	FP
No Detected	FN	TN

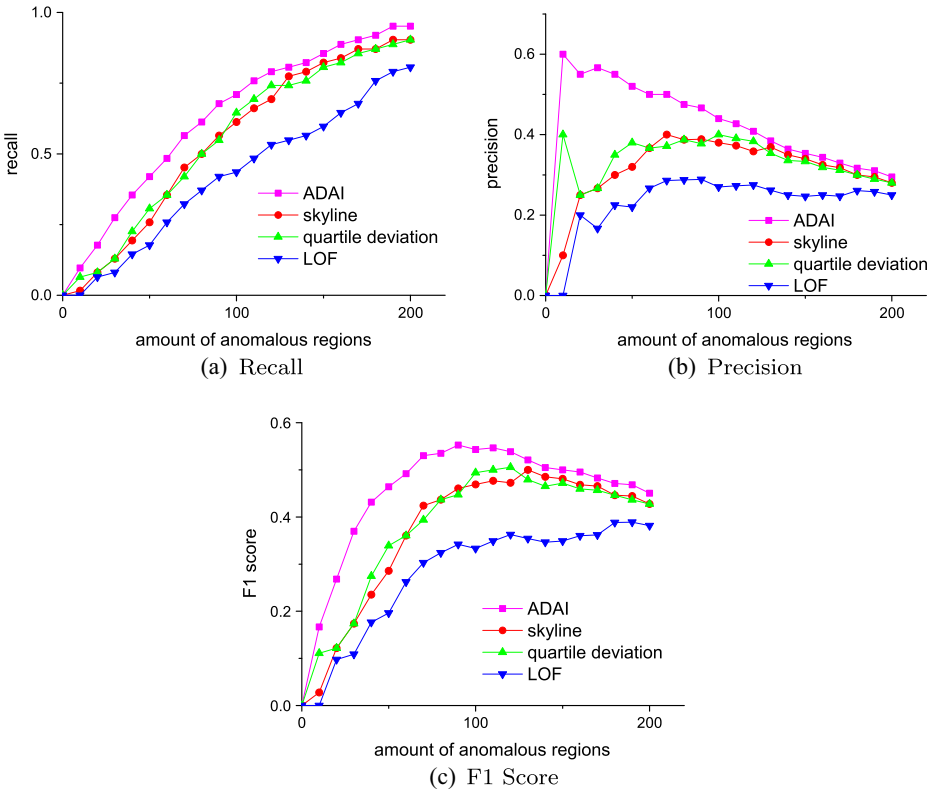
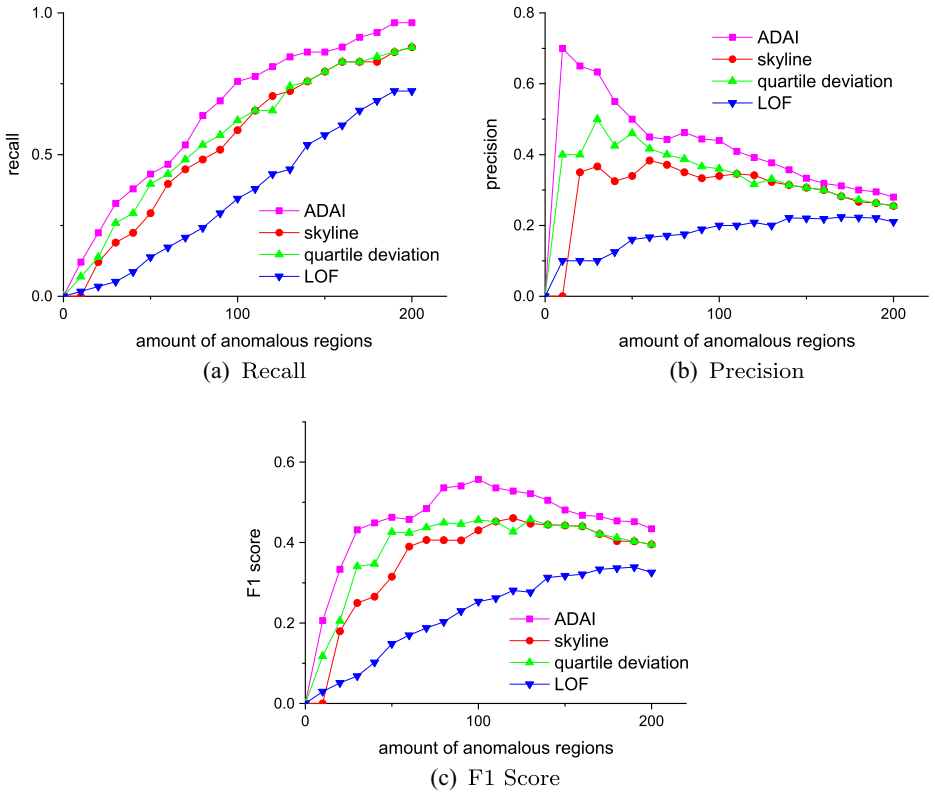


Figure 12 Evaluation results of 2014.10

anomaly index and the horizontal axis represents amount of top anomalous regions. We can see that the ADAI-method with magenta line performs better on the three metrics which are Recall, Precision and F1 Score both in October and March. It means that our method can find more anomalous regions compared with the contrast methods and our method performs better. Specially, we can see that our method perform better with the top 100 regions.

From the results above and analysis, we can conclude three points where our method is superior to the contrast method.

- With the results of three metrics from Figures 12 and 13, we conclude that our method performs better on the two datasets.
- LoTAD-method can present the anomalous degree of a region by numeric, while the contrast method can only distinguish whether a region is anomalous or not. So it can be tell intuitively which part of the city has seriously traffic problem from our results.
- There exist some situations that the contrast method cannot detect, but our method can. For example, if the average speed of a TS-segment is normal but the average stop time is large, it means that this segment is lack of effective public transport devices. So it is an anomalous condition and it may help to build subways around this place. For this condition, LoTAD can detect it but the contrast method cannot.



(a) Recall (b) Precision

(c) F1 Score

Figure 13 Evaluation results of 2015.03

7 Conclusion and future work

In this paper, we propose LoTAD, a method to detect long-term traffic anomaly in cities, which is different from the researches in anomaly detection area at the stage. Moreover, we use crowdsourced bus trajectory data to model traffic condition in urban city, which can reflect the traffic condition on road more accurately. Then we propose LoTAD to explore anomalous regions which has flawed traffic condition and the results provide suggestions for future urban traffic planning. At last, we use real bus station line dataset and bus trajectory dataset in Hangzhou to do the experiments. By analysing, we find the anomalous regions we explored in two months vary in accordance with real traffic planning and social events. At last, we do three contrast experiments and evaluate our results with the ten most congested roads in Hangzhou, which verify the effectiveness of LoTAD.

In the next step, we will consider to use multi-source data such as taxi data, POI data, and smart card data to improve the effectiveness of our results, which can show the traffic condition and travel demand in a more comprehensive way. We also consider to combine our research results with quite a few applications in daily life.

Acknowledgment The authors extend their appreciation to the International Scientific Partnership Program ISPP at King Saud University for funding this research work through ISPP#0078. This work was partially supported by the National Natural Science Foundation of China under Grants no. 61572106, the Natural Science Foundation of Liaoning Province, China under Grants no. 201602154, and the Dalian Science and Technology Planning Project under Grant no. 2015A11GX015 and 2015R054.

References

1. Batty, M.: Big data, smart cities and city planning. Dialog. Human Geograph. **3**(3), 274–279 (2013)
2. Borg, D.L., Scerri, K.: Efficient traffic modelling and dynamic control of an urban region. Transp. Res. Procedia **6**, 224–238 (2015)
3. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. SIGMOD **29**(2), 93–104 (2000)
4. Cao, T.T., Edelsbrunner, H., Tan, T.S.: Triangulations from topologically correct digital voronoi diagrams. Comput. Geom. **48**(7), 507–519 (2015)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
6. Chawla, S., Zheng, Y., Hu, J.: Inferring the root cause in road traffic anomalies. In: ICDM, pp. 141–150. Brussels (2012)
7. Chen, C., Zhang, D., Zhou, Z.H., Li, N., Atmaca, T., Li, S.: B-planner: Night bus route planning using large-scale taxi gps traces. In: 2013 IEEE International Conference on Pervasive Computing and Communications, pp. 225–233. California (2013)
8. Chen, C., Zhang, D., Castro, P.S., Li, N., Sun, L., Li, S., Wang, Z.: iboat: Isolation-based online anomalous trajectory detection. IEEE Trans. Intell. Transp. Syst. **14**(2), 806–818 (2013)
9. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mob. Netw. Appl. **19**(2), 171–209 (2014)
10. Feng, Z., Zhu, Y., Zhang, Q., Ni, L.M., Vasilakos, A.V.: Trac: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing. In: IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, pp. 1231–1239. Toronto (2014)
11. Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N.Y., Huang, R., Zhou, X.: Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM Comput. Surv. **48**(1), 7:1–7:31 (2015)
12. Guo, B., Chen, H., Han, Q., Yu, Z., Zhang, D., Wang, Y.: Worker-contributed data utility measurement for visual crowdsensing systems. IEEE Trans. Mob. Comput. **PP**(99), 1–1 (2016)
13. Guo, B., Chen, H., Yu, Z., Nan, W., Xie, X., Zhang, D., Zhou, X.: Taskme: Toward a dynamic and quality-enhanced incentive mechanism for mobile crowd sensing. Int. J. Human-Comput. Stud. **102**(6), 14–26 (2017)
14. Guo, B., Liu, Y., Wu, W., Yu, Z., Han, Q.: Activecrowd: A framework for optimized multitask allocation in mobile crowdsensing systems. IEEE Trans. Human-Mach. Syst. **47**(3), 392–403 (2017)
15. Gupta, M., Gao, J., Aggarwal, C., Han, J.: Outlier Detection for Temporal Data, vol. 5 (2014)
16. Huang, C., Wu, X.: Discovering road segment-based outliers in urban traffic network. In: 2013 IEEE Globecom Workshops, pp. 1350–1354. Atlanta (2013)
17. Jin, L., Han, M., Liu, G., Feng, L.: Detecting cruising flagged taxis' passenger-refusal behaviors using traffic data and crowdsourcing. In: UTC-ATC-ScalCom, pp. 18–25. Bali (2014)
18. Kong, X., Xu, Z., Shen, G., Wang, J., Yang, Q., Zhang, B.: Urban traffic congestion estimation and prediction based on floating car trajectory data. Futur. Gener. Comput. Syst. **61**, 97–107 (2016)
19. Kong, X., Xia, F., Wang, J., Rahim, A., Das, S.K.: Time-location-relationship combined service recommendation based on taxi trajectory data. IEEE Trans. Indus. Inform. (2017). doi:10.1109/TII.2017.2684163
20. Kuang, W., An, S., Jiang, H.: Detecting traffic anomalies in urban areas using taxi gps data. Math. Probl. Eng. **2015**(2015), 1–13 (2015)
21. Kumar, G.R., Nimmala, M., Narasimha, G.: An approach for intrusion detection using novel gaussian based kernel function. J. Univ. Comput. Sci. **22**(4), 589–604 (2016)
22. Li, J., Liu, C., Yu, J.X., Chen, Y., Sellis, T., Culpepper, J.S.: Personalized influential topic search via social network summarization. IEEE Trans. Knowl. Data Eng. **28**(7), 1820–1834 (2016)
23. Liu, W., Zheng, Y., Chawla, S., Yuan, J., Xing, X.: Discovering spatio-temporal causal interactions in traffic data streams. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1010–1018. California (2011)
24. Liu, C., Deng, K., Li, C., Li, J., Li, Y., Luo, J.: The optimal distribution of electric-vehicle chargers across a city. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 261–270 (2016)
25. Ma, S., Zheng, Y., Wolfson, O.: T-share: A large-scale dynamic taxi ridesharing service. In: 2013 IEEE 29th International Conference on Data Engineering, pp. 410–421. Brisbane (2013)
26. Mohibullah, M., Hossain, M.Z., Hasan, M.: Comparison of euclidean distance function and manhattan distance function using k-mediods. Int. J. Comput. Sci. Inf. Secur. **13**(10), 61 (2015)

27. Mrazovic, P., Matskin, M., Dokoohaki, N.: Trajectory-based task allocation for reliable mobile crowd sensing systems. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 398–406. NJ (2015)
28. Pang, L.X., Chawla, S., Liu, W., Zheng, Y.: On detection of emerging anomalous traffic patterns using gps data. Data Knowl. Eng. **87**(9), 357–373 (2013)
29. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
30. Sun, L., Zhang, D., Chen, C., Castro, P.S., Li, S., Wang, Z.: Real time anomalous trajectory detection and analysis. Mob. Netw. Appl. **18**(3), 341–356 (2012)
31. Zhang, D., Li, N., Zhou, Z.H., Chen, C., Sun, L., Li, S.: ibat: Detecting anomalous taxi trajectories from gps traces. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 99–108. Beijing (2011)
32. Zhang, J.D., Xu, J., Liao, S.S.: Aggregating and sampling methods for processing gps data streams for traffic state estimation. IEEE Trans. Intell. Transp. Syst. **14**(4), 1629–1641 (2013)
33. Zhang, D., Wang, L., Xiong, H., Guo, B.: 4w1h in mobile crowd sensing. IEEE Commun. Mag. **52**(8), 42–48 (2014)
34. Zheng, Y., Liu, Y., Yuan, J., Xie, X.: Urban computing with taxicabs. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 89–98. Beijing (2011)