IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 6, Ver. II (Nov - Dec 2018), PP 27-34
www.iosrjournals.org

Detecting contiguous regions having traffic problems using taxi trajectories

Ngoc B. Do [1], Chi Q. Nguyen [1]
[1](Posts and Telecommunications Institute of Technology, Hanoi, Vietnam)

Abstract: *Nowadays, taxi is one of the most popular transportation modes. There is a large amount of passenger using taxi everyday and taxi trajectories represent the mobility of people. In the big cities, taxi is equipped GPS device and run during 24 hours per day, they may be used to extract reliable information for transportation status. This paper states our method using taxi trajectories in Hanoi, Vietnam during 4 weeks from September 18th to October 15th. In our method, Hanoi map is divided into the smaller regions with a predefined size. We identify the contiguous regions where jams happen during different time slots and their correlations.*

Date of Submission: 18-11-2018 Date of acceptance: 04-12-2018

I. Introduction

The rapid development of urban makes the popularity increase that leads to the increasing needs of transportation and the transportation jams in some areas. The problems in the transportation always exist and make bad affects to transportation, the moving time and air pollution [13, 14]. Therefore, the prediction of regions where the traffic jams always occur is very important.

In the big cities, there is a large amount of taxi running. To operate and supervise effectively, taxi is always equipped GPS device to report the location and status to servers with a specific frequency. A large amount of GPS device generates the large amount of trajectories every day [7, 9, 13].

Taxi which is equipped GPS can be considered as a popular mobile sensor indicating traffic status, simulating trajectory patterns of people. For example, there are about 19000 taxis with transportation license for 300000 passengers (each is equivalent to 4% of the population). Therefore, each taxi ride can be considered as a significant pattern to reflect the movement of the resident of the city and the traffic flow can be modeled by using the mobility of taxi running in the roads.

In this paper, we would like to find the regions where the traffic jams usually occur and their reasons, also the correlation between each pair of regions. From that, we build a model to predict the traffic status the next day, providing the information to help managers to find the appropriate solutions. Our problem is modeling traffics and detecting abnormal: We model the traffics between the contiguous regions by using region matrix. Each cell in the matrix contains a feature set representing the effectiveness of different regions. The values of the feature set are extracted from the taxis which go through the region. Next, we would like to look for pairs of regions which have traffic problems (called skyline) from region matrix of the duration using Skyline operator. By mining popular sample data of each time slot of a specific number of days, the results show pair of regions where the traffic problems (like jams) frequently occur and their correlations.

The remaining of this paper includes the following sections. Section 2 indicates some related work and some definitions. The problem with solution and experiment is showed in the section 3. The results and solution is stated in section 4 and the conclusion is in the section 5.

II. Related Works and Some Definitions

A large number of studies in the field of mining taxi trajectory have been presented for a variety of purposes. The document [14] provides driver assistance in picking up passengers for higher profits. Other studies have focused on the construction of intelligent transportation systems that help guide driving [11], intelligent intersections that minimize the impact of vehicle emissions on the air environment when vehicles are required to wait [3] [14]. Unlike researches focused exclusively on taxi drivers and drivers, our study included helping transportation managers find the areas where the problems occur and the cause.

The document [7] deals with detecting traffic anomalies such as accidents, congestion based on taxi tracking. Several other studies have attempted to evaluate the construction of transport works [15]. Studies in the Urban computing group, such as the exploration of human activities in urban areas, estimate the similarity level each day of the week [9] [13], study traffic flow, focus areas and images and its effect. Unlike studies that only detect problems when imminent, our study builds traffic prediction models. This model allows users to know in advance to avoid areas with poor traffic conditions and traffic managers offer the appropriate solution.

In the GPS data of taxi traffics, each trajectory includes a series of points (id, time, latitude, longitude, state, velocity, distance). A taxi has 3 operating status: no passenger, going to have passenger, having passenger.

Definition 1: Region. Map is divided into smaller regions with a predefined size, which includes road parts representing their traffic status.

Definition 2: Trajectory. A trajectory is a series of GPS points along the time $Tr : p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$, in which, each point includes longitude, latitude, time, state, velocity, distance

Definition 3: Trip and sub-trip. From a trajectory $Tr : p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$, by connecting GPS point to corresponding region codes (for example $\langle p_1, r_i \rangle \rightarrow \langle p_2, r_j \rangle \rightarrow \dots \rightarrow \langle p_n, r_k \rangle$). A sub-trip $s : r_1 \rightarrow r_2$ is created if p_i and p_j (from *Tr*) are the first point in r_1 and r_2 (i<j), where distance and velocity of sub-trip s are calculated by Equation 1 and 2

$$d(p_i, p_j) = p_j.d - p_i.d \tag{1}$$

$$v = d(p_i, p_j) / (p_j.t - p_i.t) \tag{2}$$

In Equation 2, velocity is calculated by d/t instead of calculating the average value sent from GPS, reflecting the average velocity more exactly because the traffic light waiting time also is included which GPS devices might ignore.

Each trajectory can produce many sub-trips but only one trip, the sub-trip between the beginning region and the ending region of one trajectory is a trip. At the following sections, we will call both "trip" and "sub-trip" as "trip".

III. Modeling Traffics and Detecting Problems

When going through road parts where traffic jams occur frequently, people can choose a longer road but higher speed. This is one of the reasons which make some roads stuck due to the jams from other roads. The problem helps to detect pair of regions which have traffic jams and the correlation between two regions.

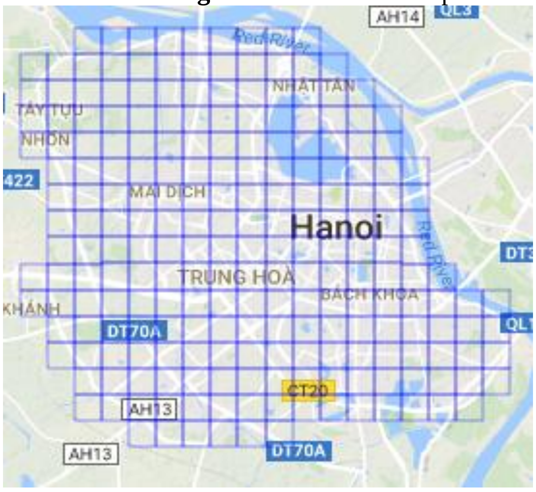
3.1 Traffic Modeling

In this section, firstly we divide the city map into many regions, then construct region matrix with each different time slot.

3.1.1 Partitioning maps

We partition the map of Hanoi including inner city and some areas with high population into squares sized 1km x1 km (as showed in figure 1). Partitioning method is chosen instead of researching roads because the jams are the consequence while the entire regions bring the transportation information and the roots of problems. Moreover, partitioning maps can help us to find the place which the jams exactly occur.

Figure 1. Partitioned Map

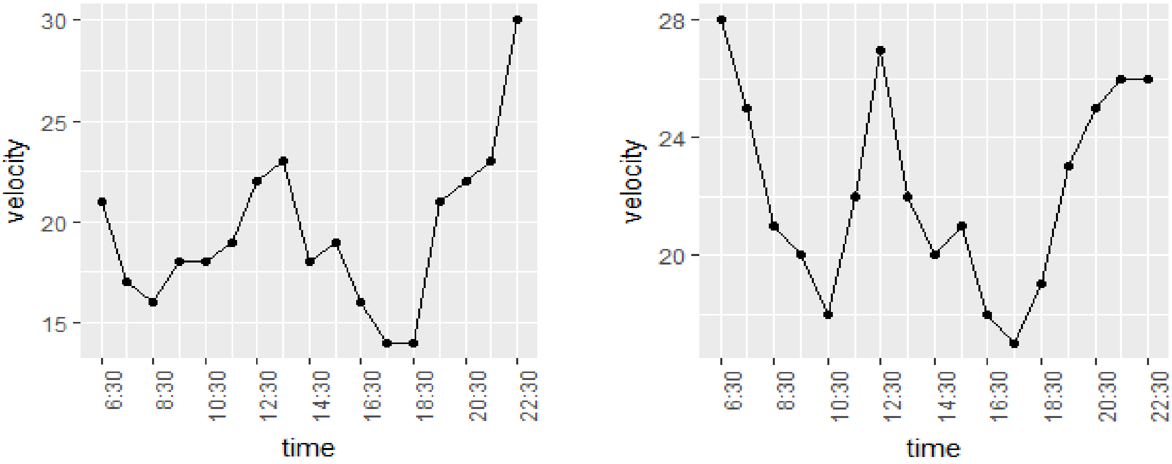


3.1.2 Constructing region matrix

Time division: Before constructing region matrix, we divide the taxi trajectories according to each day in week and different time slots in a day because the traffics in different days and times are different and the traffics status are also different [8].

During a same period of time, the traffic status and transportation of the people are similar and the traffics problem also can occur during this time. So, time division can help explore the problems in more details. As can be shown in figure 2A, average velocity in the city during the early morning of business days (7a.m to 10.30a.m) is lowest in the mornings. Similarly, the time slot from 4p.m to 7.30p.m is for coming back home. The results have described exactly the traffics status in rush hours is lower than the different time slots. Figure 2B represents the average velocity during weekends, showing that the velocity during 2 weekend days is similar in which the lowest velocities are of 2 rush hour slot in the morning and afternoon.

Figure 1. Taxi Velocity during the different time slots in Hanoi



A) Business day

B) Weekend

From figure 2, we suggest to divide time as the table 1

Table 1. Time Division

Time	Business day	Weekend
Slot 1	00:00 – 7:00	00:00 – 08:00
Slot 2	07:00 – 10:30	08:00 – 11:00
Slot 3	10:30 – 16:00	11:00 – 16:00
Slot 4	16:00 – 19:00	16:00 – 19:00
Slot 5	19:00 – 24:00	19:00 – 24:00

Constructing region matrix: Firstly, we choose the trajectories having passenger, these trajectories represent the transportation of a person. Then, we put these trajectories into the map and construct trips between two regions (according to definition 3).

Figure 2. Put some trajectories into map

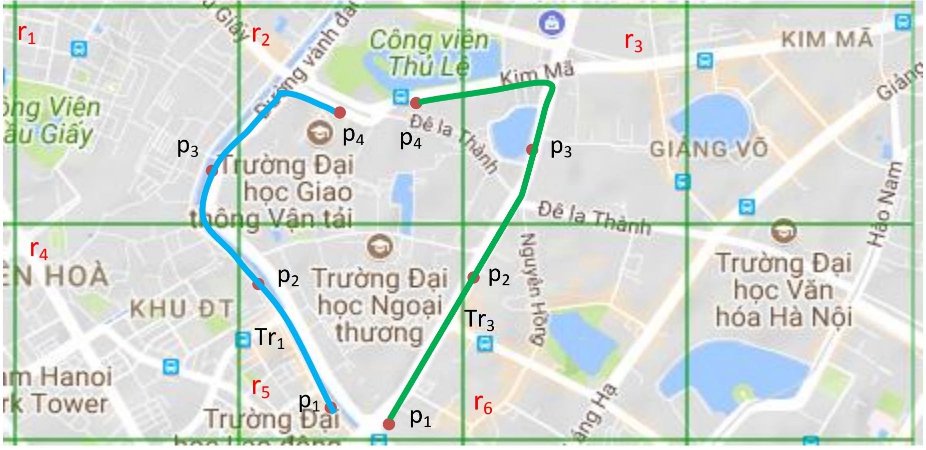


Figure 3 describes 2 trajectories in the map with blue and green, GPS points is orange, regions is showed by red color. The trajectory Tr_1 going through $r_5 \rightarrow r_2 \rightarrow r_1$ constructs 3 trips $r_5 \rightarrow r_2$, $r_2 \rightarrow r_1$ and $r_5 \rightarrow r_1$, Tr_2 going through $r_5 \rightarrow r_6 \rightarrow r_3 \rightarrow r_2$ constructs 6 trips. Two trajectories with different roads can construct the trip $r_5 \rightarrow r_2$. Note that trajectory Tr_1 does not construct $r_5 \rightarrow r_4$ since there is no GPS point from Tr_1 in r_4.

Each pair of regions $r_1 \rightarrow r_2$ has a set of trips between them, by summarizing these trips in this set, each a pair of regions has a feature set: the number of trips |S| representing traffic flow, average velocity E(V) and average moving distance E(D). This feature set is calculated in Equation 3 and 4 with S is the set of trips

$$E(V) = \frac{\sum_{s_i \in S} S_i.v}{|S|} \tag{3}$$

$$E(D) = \frac{\sum_{s_i \in S} S_i.d}{|S|} \tag{4}$$

Region matrix M is constructed as in figure 4 from each time slot and each day, each value in the matrix is corresponding to each pair contiguous regions, is denoted as feature $a_{i,j}$ = < |S|, E(V), E(D) >.

Figure 3. Region Matrix

$$M = \begin{array}{c|ccccc|}
 & r_0 & r_1 & \ldots\ldots & r_{n-1} & r_n \\
\hline
r_0 & \emptyset & & \ldots\ldots\ldots\ldots\ldots & & a_{0,n} \\
r_1 & a_{1,0} & & \ldots\ldots\ldots\ldots\ldots & & a_{1,n} \\
\vdots\ \vdots & & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\
r_{n-1} & a_{n-1,0} & & \ldots\ldots\ldots\ldots\ldots & & a_{n-1,n} \\
r_n & a_{n,0} & & \ldots\ldots\ldots\ldots\ldots & & \emptyset \\
\hline
\end{array}$$

3.2. Detecting Problems

Firstly, we detect the skyline from region matrix in each time slot. Then we mine the patterns to find pairs of regions where traffic jams occur frequently and the relation between them.

3.2.1 Detecting skyline

The traffic problem between pairs of regions can be described as the followings:

- The connection between 2 regions is represented by all the roads which can be moved because drivers sometimes can choose different roads to go to other regions to avoid the traffics jams.
- Although the shortest way between 2 regions is hard to move, the driver still decides to move through this way instead of the round ways

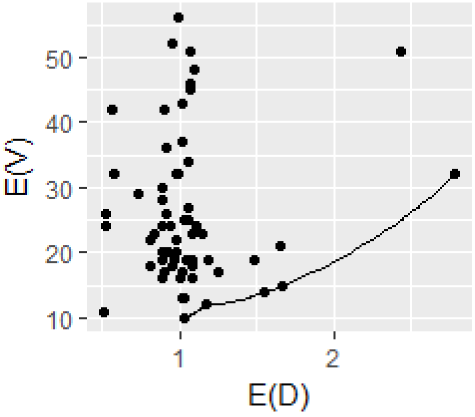
A small E(V) means the ways connecting regions are having bad traffic status. A large E(D) means that the taxi must go around way and the shortest way between 2 regions has a problem. So, E(V) and E(D) are used to find the problems. The tuple < |S|, E(V), E(D) > indicates the model of connection and traffics between 2 regions. E(D) shows the geometric feature of the connection between 2 regions, a large E(D) means that we need to go a longer way to move to another region, E(V) and |S| represent the traffics features.

At the beginning, we choose pairs of regions which have the number of trips larger than the average number from matrix M, these pairs of regions are considered as crowded and having big effect regions if the some problem occurs. Then, we use Skyline operators [2] to detect pairs of regions according to E(V) and E(D).

Definition 4. Skyline L is a set of points which are not dominated by any other point. A point dominates another point if it is better in all dimensions or at least one dimension.

In this problem, a pair of regions $a_{i,j} \in L$ if there is no any pair of region $a_{p,q} \notin L$ in which E(V) is smaller and E(D) is larger than $a_{i,j} \in L$. Figure 5A shows Skyline is the black line in the lower right conner, we can see that there is no point outside which has smaller E(V) and larger E(D) than any point in the skyline.

Figure 4. An example of detecting skyline



Point	E(V)	E(D)
1	10	1.026
2	12	1.176
3	14	1.552
4	21	1.66
5	19	1.481
6	17	1.023
7	15	1.673
8	32	2.79
9	51	2.44

A) Skyline

B) Detecting skyline

Figure 5 shows an example of skyline: E(V) and E(D) in the figure 5B and the picture of a skyline in figure 5A. In this example, point 1 and 8 are in the skyline because 2 these points are not affected by any other point due to they have the smallest E(V) and the largest θ.

Point 6 is not in the skyline due to it is affected by point 1. Point 2 and 3 are also detected being in the skyline but point 4 and 5 are not due to point 2, point 9 is not due to point 8.

3.2.2 Mining patterns

First, we build skyline for each day and each time slot. Then, we apply Apriori algorithm to mine patterns ([1, 5]) to find the pairs of regions which frequently occur traffic jams because the jams sometimes occur only in a specific time slot. This method helps to find the association rules between pair of regions then pair of problem regions during the time of each day, then pair of problem regions during a time slot. Finally, the remaining pairs of popular regions are the pairs of problem regions.

The mining pattern process uses the following information: the support shows the frequencies of occurrence of pair rp (according to formula 5). The pairs with their supports larger than a particular threshold δ are considered as the problem pairs in the duration of time

$$Support(rp) = \frac{|rp|}{number\ of\ days} \quad (5)$$

Association rule mining find patterns according to formula 6, 7 in which $|rp_1 \cap rp_2|$ is the number of days during that rp$_1$ and rp$_2$ regions occur. $Support(rp_1 \Rightarrow rp_2)$ indicates the frequency of co-occurrence of rp$_1$ and rp$_2$. $Confidence(rp_1 \Rightarrow rp_2)$ indicates the probability of occurrence of rp$_2$ given the occurrence of rp$_1$.

$$Support(rp_1 \Rightarrow rp_2) = \frac{|rp_1 \cap rp_2|}{number\ of\ days} \quad (6)$$

$$Confidence(rp_1 \Rightarrow rp_2) = \frac{|rp_1 \cap rp_2|}{|rp_1|} \quad (7)$$

Figure 5. Association rule mining

Time	**Day 1**	**Day 2**	**Day 3**	**Support >=2/3**	**Support=1/3**
Slot 1	$r_1 \rightarrow r_3$ $r_2 \rightarrow r_3$ $r_4 \rightarrow r_5$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$	$r_2 \rightarrow r_3$
Slot 2	$r_4 \rightarrow r_5$ $r_5 \rightarrow r_7$	$r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$ $r_6 \rightarrow r_8$	$r_1 \rightarrow r_4$ $r_6 \rightarrow r_8$ $r_2 \rightarrow r_3$	$r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$ $r_6 \rightarrow r_8$	$r_2 \rightarrow r_3$ $r_5 \rightarrow r_7$
Slot 3	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_2 \rightarrow r_6$	$r_1 \rightarrow r_3$ $r_2 \rightarrow r_6$ $r_4 \rightarrow r_5$	$r_1 \rightarrow r_4$ $r_3 \rightarrow r_6$ $r_4 \rightarrow r_2$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_2 \rightarrow r_6$	$r_3 \rightarrow r_6$ $r_4 \rightarrow r_2$ $r_4 \rightarrow r_5$

Figure 6 represents an example of association rule mining from skyline through a number of days in the duration of time.

In time slot 1, a pair of regions $r_1 \rightarrow r_3$ occurs in 3 days so the support being 1, $r_1 \rightarrow r_{4,}$ $r_4 \rightarrow r_5$ occur in 2 days so the support is 2/3, $r_2 \rightarrow r_3$ occur only the first day so the support is 1/3. Similarly, according to formula 6, the rule $(r_1 \rightarrow r_3) \Rightarrow (r_4 \rightarrow r_5)$ has the support of 2/3, the confidence of 2/3 while the rule $(r_4 \rightarrow r_5) \Rightarrow (r_1 \rightarrow r_3)$ has the confidence of 1.

The association rules with their supports and confidence larger than a given threshold can show the cause and effect information about the pairs of regions. Then, we continue to mine patterns of pairs of problem regions during each time slot. The pairs of regions satisfied the final conditions and the association rules of these regions can be considered as problem regions during all time slots.

IV. Result and Solution

The traffic jams usually occur in business days and rush hours. To find the frequent jam regions, we create skylines for time slot 2, 3, 4 of business days in a week (Monday-Friday). During a time slot, each pair of region occur jams more than twice a week can be considered as problem regions.

Figure 6. Problem regions in business days

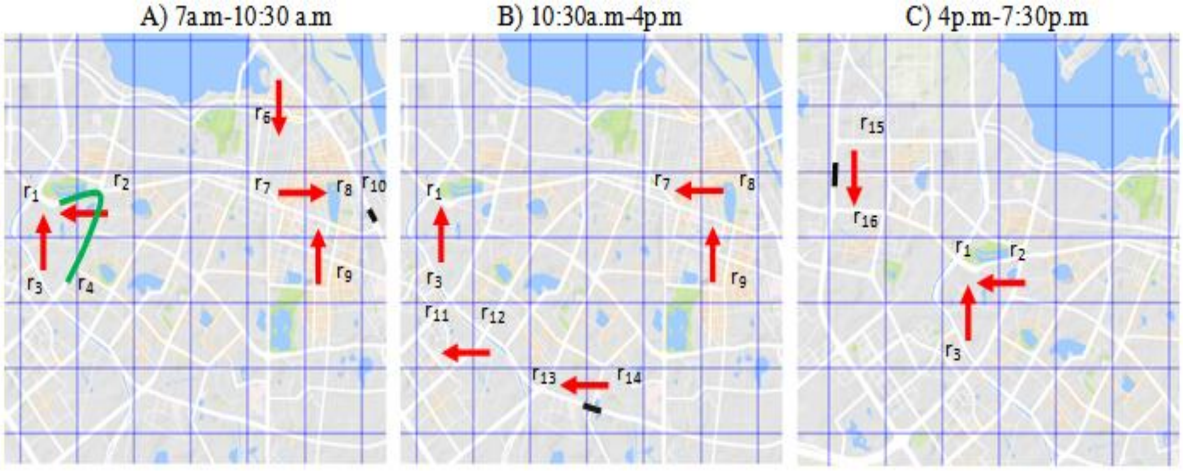


Figure 7 represents frequent problem regions in business days. According to the map, the problem regions can be divided into 2 main groups and some individual regions. The first group is r_1, r_2, r_3 the second group is the regions of r_7, r_8, r_9. The individual pairs of regions are $r_5 \rightarrow r_6$, $r_{12} \rightarrow r_{11}$, $r_{14} \rightarrow r_{13}$, $r_{15} \rightarrow r_{16}$.

Look at group 1 of 3 regions (r_1, r_2, r_3), we can see that during the time from 7a.m to 10.30a.m (fig 7A), the moving direction from region r_3 and r_2 to r_1 has traffic jams but the directions from r_1 to others regions have not any jam because from here people can move towards many different directions. In addition, the moving direction from r_3 to r_1 is shortest and most reasonable if moving to the left of r_1. The fact that the pair of region $\{r_1 \rightarrow r_3\}$ continues to appear at noon and rush hour of the afternoon indicates the traffics jams in this region gradually occur during all the time of days, the pair of region $\{r_2 \rightarrow r_1\}$ does not occur at the time slot from 10.30a.m to 4p.m (Fig 7 B) shows that this region has the traffics jams during the rush hour.

The problems in these regions can be explained as the followings: the shortest way connecting $\{r_3 \rightarrow r_1\}$ has jams all the time of days and especially during rush hour. So, during this time, the around way $r_4 \rightarrow r_2 \rightarrow r_1$ (the green line in figure 7A) is chosen. When taxies move along this way to the square of r_2 the traffic flow increases a lot that causes the problem for the pair of region of $\{r_2 \rightarrow r_1\}$. If the problem of $\{r_3 \rightarrow r_1\}$ is solved then the problem of $\{r_2 \rightarrow r_1\}$ also is solved.

In the group 2 the region r_9 and r_7 towards to r_8 occur the problem in the morning. As can be seen in the map, people want to move towards region r_{10} and larger roads (black line in figure 7A) to move more easily. At noon and in early afternoon, the moving direction from r_9 to r_8 still has problem while the direction from r_8 to r_7 has problem in the morning. This fact is because people want to return after finishing morning activities and move to urban. In this group, the pair $\{r_9 \rightarrow r_8\}$ is considered as the key reason of the problems, so we need to solve the problem of this pair first then the problem of this group.

Among the remaining individual regions, the pair $\{r_{15} \rightarrow r_{16}\}$ occurs during the rush hour in the afternoon. Since there is no other pair in this area having jams and there is only one connecting way, we can conclude that the problem of this way is due to the way capacity cannot afford the number of vehicles here. The solution is to extend the way. The pair $\{r_{14} \rightarrow r_{13}\}$ is rather similar to the pair of $\{r_{15} \rightarrow r_{16}\}$, the given solution is similar to the pair of $\{r_{14} \rightarrow r_{13}\}$. The pair of regions $\{r_5 \rightarrow r_6\}$ has no direct connecting so people have to use around way leading to waste fuel and time, this pair also should be solved. The remaining pair $\{r_{12} \rightarrow r_{11}\}$ has not been able to find the reasons and solutions because there are some different ways and directions to go.

V. Conclusion

In this study, we have solved a problem using GPS data of taxi: Find out where frequent problems occur in cities, using taxi GPS data by modeling traffic between geographic pairs of regions. The results reveal the problems of each pair of regions such as insufficient response to demand or lack of direct connection and the relationship of pairs of regions. After grouping these pairs and identifying the problem, we proposed a solution for each group. However, some pairs cannot identify the cause and offer solutions.

In the future, we intend to continue to work on this problem with a number of other ways of partitioning regions to better pinpoint the problem and propose solutions.

References

[1]. Rakesh Agrawal, Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, California, United States on June, (1994)

[2]. Stephan Börzsönyi, Donal Kossmann, Konrad Stocker, The skyline operator, Proceedings of the 17th International Conference on Data Engineering, Washington, United States on April 02 – 06, (2001)

[3]. Jonathan J. Buonocore, Harrison J. Lee, Jonathan I. Levy, The Influence of Traffic on Air Quality in an Urban Neighborhood: A Community–University Partnership, Am J Public Health, Volume 99, pp 629-635 (2009)

[4]. Luchao Cao, Latita Thakali, Liping Fu, Garrett Donaher, Effect of Weather and Road Surface Conditions on Traffic Speed of Rural Highways, Annual Meeting of the Transportation Research Board, Washington, United States on January 13-17, (2013)

[5]. Jochen Hipp, Ulrich Güntzer, Gholamreza Nakhaeizadeh, Algorithms for association rule mining - a general survey and comparison, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1, pp 58-64 (2000)

[6]. Markku Kilpeläinen, Heikki Summala, Effects of weather and weather forecasts on driver behaviour, Science Direct, Volume 10, Issue 4, pp 288-299 (2007)

[7]. Weiming Kuang, Shi An, Huifu Jiang, Detecting Traffic Anomalies in Urban Areas Using Taxi GPS Data, Mathematical Problems in Engineering, Volume 2015 (2015)

[8]. Eric M.Laflamme, Paul J.Ossenbruggen, Effect of time-of-day and day-of-the-week on congestion duration and breakdown, Science Direct, Volume 1, pp 31-40 (2017)

[9]. Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, Yuan Tian, Understanding intra-urban trip patterns from taxi trajectory data, Journal of Geographical Systems, Volume 14, Issue 4, pp 463-483 (2012)

[10]. Nordiana Mashros, Johnnie Ben- Edigbe, Sitti Asmah Hassan, Norhidayah Abdul Hassan, Nor Zurairahetty Mohd Yunus, Impact of Rainfall Condition on Traffic Flow and Speed: A Case Study in Johor and Terengganu, Jurnal Teknologi, Volume 70 (2014)

[11]. Mostofa Kamal Nasir, M. A. Kalam, B. M. Masum, Rafidah Md. Noor, Reduction of Fuel Consumption and Exhaust Pollutant Using Intelligent Transport System, The Scientific World Journal, Volume 2014 (2014)

[12]. Ju Sam Oh, Yong Un Shim, Yoon Ho Cho, Effect of weather conditions to traffic flow on freeway, KSCE Journal of Civil Engineering, Volume 6, pp 413-420 (2002)

[13]. Chengbin Peng, Xiaogang Jin, Ka-Chun Wong, Meixia Shi, and Pietro Liò, Collective Human Mobility Pattern from Taxi Trips in Urban Area, PLOS ONE, Volume 7, Issue 4 (2012)

[14]. Marco Veloso, Santi Phithakkitnukoon, Carlos Bento, Pedro d'Orey, Mining Taxi Data for Describing City in the Context of Mobility, Sociality, and Environment, IEEE Intelligent Transportation Systems Conference, Rio de Janeiro, Brazil on November, (2016)

[15]. Yu Zheng, Yanchi Liu, Jing Yuan, Xing Xie: Urban Computing with Taxicabs, 13th International Conference on Ubiquitous Computing, Beijing, China on September 17-21 (2011)

Ngoc B. Do. " Detecting contiguous regions having traffic problems using taxi trajectories." IOSR Journal of Computer Engineering (IOSR-JCE) 20.6 (2018): 27-34.