DETECTING JAM REGIONS CORRELATIONS AND PREDICTING TAXI TRANSPORTATION FLOW AND VELOCITY

Nguyễn Quỳnh Chi
Information Technology Department - Posts and Telecommunications Institute of Technology

Abstract: Nowadays, taxi is one of the most popular transportation modes. There is a large amount of commuter using taxi every day 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. Next, we identify the contiguous regions where jams happen during different time slots and their correlations. Finally, we develop a model predicting taxi transportation flow in each region and the velocity basing on historical and weather data.

Keywords: Taxi transportation flow prediction, contiguous regions jams, velocity.

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 [1, 2]. 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 [1, 3, 4].

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 taxi with transportation license for 300000 commuters (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 co-relation 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. We will implement 2 problems as the followings:

Problem 1. 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.

Problem 2. Predicting traffic flow and velocity: We develop traffic flow set and velocity in each region in combination with weather data to predict the traffic's status of the next day. The prediction results can be considered as the suggestions to help the transportation managers have solutions which make transport avoid these regions.

The taxi trajectory data, velocity data have been collected from http://gps.binhanh.vn in Hanoi during 4 weeks from September 18th to October 15th, 2018. All the data file is in the form .json of Java. We need to preprocess data to extract it and transform it into suitable form for all experiments in this paper.

The remaining of this paper includes the following sections. Section II indicates some related works and some backgrounds. The problem 1 with solution and experiment is showed in the section III and the problem 2 in the section IV. The conclusion is in the section V.

II. RELATED WORK AND BACKGROUNDS

A large number of studies in the field of mining taxi trajectory has been presented for a variety of purposes.

Contact author: Nguyen Quynh Chi,
Email: chinq@ptit.edu.vn
Arrival: 12/10/2019, Revised: 12/2019, Accepted: 12/2019.

The study [2] provides driver assistance in picking up passengers for increasing profits. Other studies have focused on the construction of intelligent transportation systems that help guide driving [5], intelligent intersections that minimize the impact of vehicle emissions on the air environment when vehicles are required to wait [2, 6]. Unlike only drivers were focused, our study can help transportation managers to find the regions where the problems occur and the cause.

The study [3] 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 [7]. 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 [1, 4], study traffic flow, focus on regions, images and their effect. Unlike studies that only detect problems when imminent, our study builds a traffic prediction model. 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 commuter, going to have commuter, having commuter.

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 $Tr: p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_n$, in which, each point p includes longitude, latitude, time, state, velocity, distance.

Definition 3. Trip and sub-trip: From a trajectory $Tr: p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_n$, by connecting GPS point to corresponding region codes (for example $\langle p_1, r_i \rangle \rightarrow \langle p_2, r_j \rangle \rightarrow ... \rightarrow \langle p_n, r_k \rangle$). 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

$$d(p_i, p_j) = p_j.d - p_i.d \quad (1)$$

$$v = d(p_i, p_j) / (p_j.t - p_i.t) \quad (2)$$

In Equation 2, velocity is calculated by d/t (d here is euclide distance) instead of calculating the average value sent from GPS. This makes the average velocity more exact because the traffic light waiting time (which GPS devices might ignore) is included.

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. PROBLEM OF MODELING TRAFFICS AND DETECTING JAMS

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 1 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

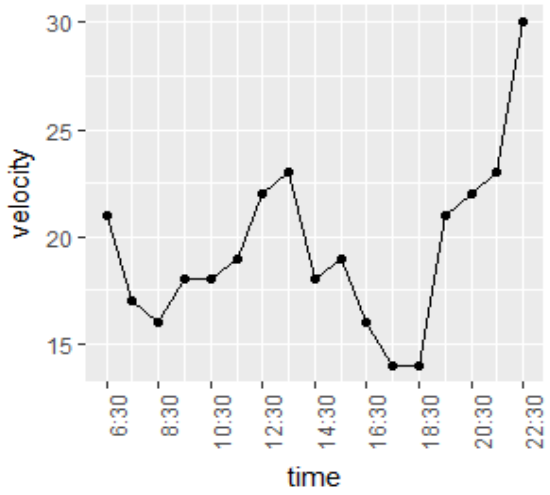
We partition the map of Hanoi including inner city and some areas with high population into squares sized 1km x 1 km (as showed in figure 1). Partitioning method is chosen instead of researching roads because the jams are the consequence while the entire regions bring the transportation information and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

Figure 1: Map which is partitioned

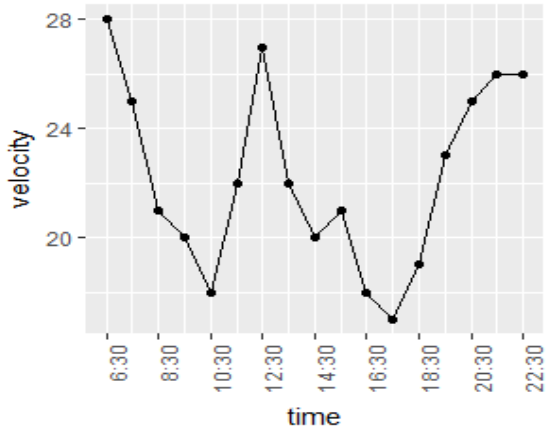
3.1.2 Constructing region matrix

Time division: Before constructing region matrix, we divide the taxi trajectors 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 (7 a.m to 10.30 a.m) is the lowest in the mornings. The velocity is the lowest in the afternoon during the time slot from 4p.m to 7.30 p.m, the time 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 hours slot in the morning and afternoon.



A) Business day



B) Weekend

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

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

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

Table 1: Time Division

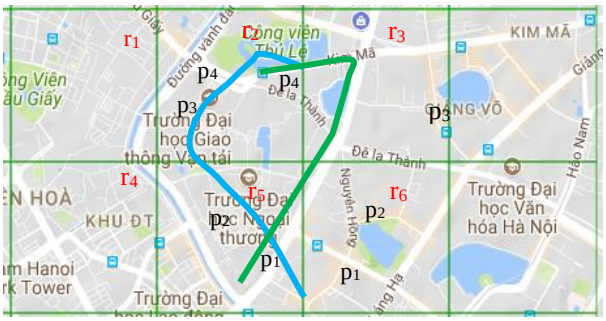


Figure 2: Put some trajectories into map

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

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

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

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

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 $a_{i,j} = <|S|, E(V), E(D)>$.

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

Figure 3: Region Matrix

3.2 Detecting Problem

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

3.2.1 Detecting skyline

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

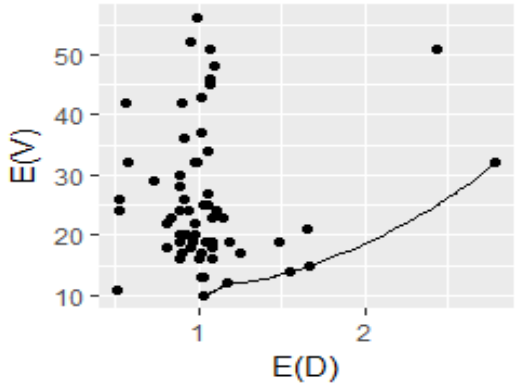
- The connection between 2 regions is represented by all the roads which can be moved because drivers sometimes can choose different roads to go to other regions to avoid the traffics jams.
- Although the shortest way between 2 regions is hard to move, the driver still decides to move through this way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

A small value of E(V) means the ways connecting regions are having bad traffic status. A large value of 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 [9] 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.



A) 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

B) Detecting Skyline

Figure 4: An example of 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 [10, 11] 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.

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

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.

Time	Day 1	Day 2	Day 3
Slot 1	$r_1 \rightarrow r_3$ $r_2 \rightarrow r_3$ $r_4 \rightarrow r_5$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$
Slot 2	$r_4 \rightarrow r_5$ $r_5 \rightarrow r_7$	$r_1 \rightarrow r_4$ $r_4 \rightarrow r_5$ $r_6 \rightarrow r_8$	$r_1 \rightarrow r_4$ $r_6 \rightarrow r_8$ $r_2 \rightarrow r_3$
Slot 3	$r_1 \rightarrow r_3$ $r_1 \rightarrow r_4$ $r_2 \rightarrow r_6$	$r_1 \rightarrow r_3$ $r_2 \rightarrow r_6$ $r_4 \rightarrow r_5$	$r_1 \rightarrow r_4$ $r_3 \rightarrow r_6$ $r_4 \rightarrow r_2$

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

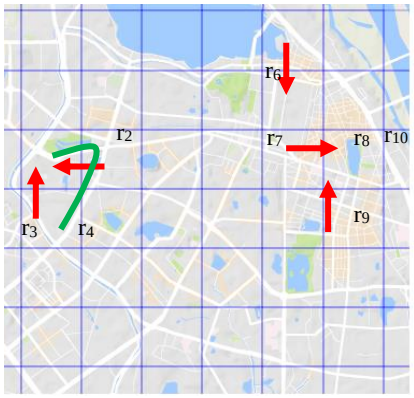
Figure 5: Association rule mining

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 confidene of 2/3 while the rule $((r_4 \rightarrow r_5) \Rightarrow (r_1 \rightarrow r_3))$ has the confidence of 1.

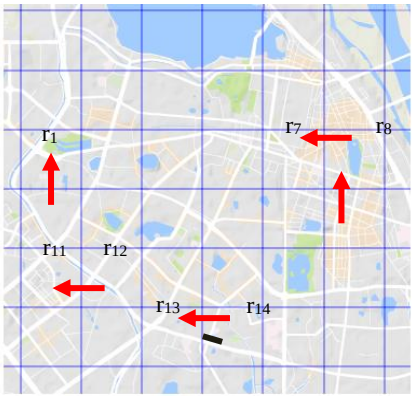
The association rules with their supports and confidence larger than a given threshold can show the cause and effect information about the pairs of regions. Then, we continue to mine patterns of pairs of problem regions during each time slot. The pairs of regions satisfied the final conditions and the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

3.3 Results and solution

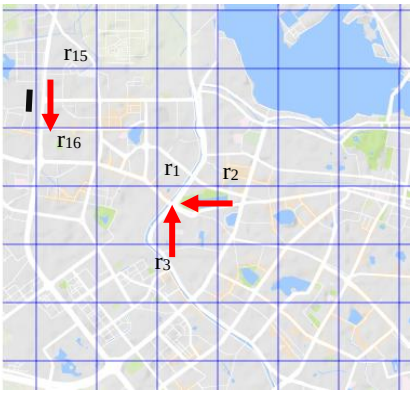
The traffic jams usually occur in business days and rush hours. To find the frequent jam regions, we create skylines for time slot 2, 3, 4 of business days in a week (Monday-Friday). During a time slot, each pair of region occur jams more than twice a week can be considered as problem regions.



A) 7a.m-10:30 a.m



B) 10:30a.m-4p.m



C) 4p.m-7:30p.m

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 two main groups and some individual regions. The first group is (r_1, r_2, r_3) and the second group is (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 10.30 a.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 10.30 a.m to 4 p.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.

The detection of jams computed basing on regions instead of the connecting ways can provide a general view on traffic status, however there are many ways between two regions, even they are in reversed directions. In this situation, the connection between two regions could not offer some useful suggestions for drivers if the real traffics in these ways are different.

IV. PREDICTING TRAFFIC FLOW AND VELOCITY

Each geographic region has different traffic characteristics, and these characteristics vary from time to time. Some areas have poor traffic conditions in the morning but are good at noon and afternoon. In addition, traffic conditions are influenced by a number of factors, such as the weather or the day of the week. For example, a person who regularly travels by motorbike but due to the weather is too hot, this person decides to move by taxi or due to good weather most people decide to use personal vehicles to move. Every weather change affects the state of traffics, people will want to know what the impact of weather and how much traffic is expected tomorrow in weather conditions. The purpose of Problem 2 is to predict the flow and velocity of the taxi in each region, which determines the traffic conditions in each region, and gives recommendations to drivers and managers.

4.1 Creating feature sets

The flow of taxi passing through the r region is determined by the trajectory of passing passengers r1. By aggregating points from these trajectories on r, we can calculate the velocity of the taxi through Equation 8. Taxi traffic flow represents the change in traffic flow over time and speed represents the traffic condition here.

$$M(V)_r = \sum p_i \in P_r \quad (8)$$

In this case, P_r is the set of GPS points located in the right trajectory in r region

In this problem, we build the feature set in every 1 hour because the traffic characteristics change enough to see the difference from the previous time. In addition, within one hour, changes in weather conditions may be different and impacts on traffic with varied levels. Table 2 shows an example of a feature set of a region.

Weather is always one of the main factors of traffic. Many studies have examined the effects of direct weather conditions on traffics, such as pavement conditions, rain and snow [12, 13, 14]. Rain is considered the most influential factor in traffic in Hanoi due to tropical climate. Here, the average annual rainfall is 1800mm and in the rainy season in July, August, the rainfall can reach 500mm / month (data from the Statistics General Office 2016). Rain causes the area of the road to be reduced, moving difficult due to being limited by water and slowing people down due to dressing and feeling.

In addition to the direct impact elements, several studies conducted to determine the effect of weather on the driver [15]. In addition, weather can affect the decision to participate in human traffic and indirectly affect traffic. In this study, we use the following information and indicators

Heat Index: The heat index is a combination of temperature and relative humidity. This index considers the comfort of the body. For example, when the body feels hot it will sweat to lower body temperature. When the humidity is high, the rate of sweat decreases making the body feel hotter. The Heat Index is calculated by Equation 9 where T is the temperature measured in degrees F, R is the relative humidity.

$$HI = -42.379 + 2.04901523T + 10.14333127R - 0.22475541*TR - 6.83783 * 10^{-3} T^2 - 5.481717 * 10^{-2}R^2 + 1.22874 * 10^{-3}T^2R + 8.5282 * 10^{-4}TR^2 - 1.99x * 10^{-6}T^2R^2 \quad (9)$$

Dew Point: Dew point is a combination of heat, humidity, it refers to the temperature at which steam condenses into liquid water, which can be changed into rain. Dew Point is calculated by Equation 10 with a = 17.27, b = 237.7.

$$T_{dewpoint} = \frac{b\left(\frac{aT}{b+T} + \ln(RH)\right)}{a - \left(\frac{aT}{b+T} + \ln(RH)\right)} \quad (10)$$

Table 2 shows an example of the change in flow and velocity of days in the week that combined the weather data. In the table 2, T (C) is the temperature in degrees celsius, P (MM) is the rainfall in millimeter, HI and DP are the temperature and dew point, and M (V) is the average taxi flow and velocity. On rainy days (3-8 / 10), people usually take more taxis and the speed of travel is also lower than the sunny days (1.2 / 10, 9/10).

Table 1: An example of feature sets and weather

Day	Time	Outlook	T(C)	P(MM)
1/10	7:00	Sunny	29	0
2/10	7:00	Sunny	28	0
3/10	7:00	Moderate rain shower	28	1.4
4/10	7:00	Moderate rain shower	28	1.4
5/10	7:00	Patchy rain possible	27	0.6
6/10	7:00	Moderate rain shower	27	1.3
9/10	7:00	Partly cloudy	27	0
10/10	7:00	Light rain shower	26	2.9
11/10	7:00	Torrential rain shower	26	12.5

12/10	7:00	Light rain shower	27	1
13/10	7:00	Cloudy	24	0

Day	Time	HI(°C)	DP(°C)	|S|	M(V)
1/10	7:00	34	25	50	23
2/10	7:00	33	24	55	25
3/10	7:00	33	24	63	13
4/10	7:00	32	24	69	12
5/10	7:00	31	23	72	15
6/10	7:00	31	23	65	17
9/10	7:00	31	23	56	21
10/10	7:00	29	23	68	14
11/10	7:00	29	24	71	17
12/10	7:00	30	23	64	19
13/10	7:00	26	19	56	24

4.2 Building machine learning models

To build machine learning models for predictive work, we first transform the data to fit the model by dividing the information and indexes into some groups. Table 3A shows rainfall classification with P is the rainfall in mm/h. Table 3B shows the classification of temperature, Table 4 shows the classification of heat index and dew point.

Table 2: Rain and Temperature classification

Order	Level	P(mm)/1h
1	No rain	0
2	Small rain	Less than 0.25
3	Heavy rain	0.25 to 2.0
4	Very heavy rain	More than 2.0

A) Rain Classification

Order	Temp (°C)	Perception
1	Less than 10	Very cold
2	10 to 19	Cold
3	20 to 25	Cool
4	26 to 33	Normal
5	More than 33	Hot

B) Temperature Classification

Table 3: Heat Index and Dew Point Classification

Heat Index (°C)	Perception
27 to 32	Feeling tired
32 to 39	Heat shock, loss of strength
39 to 51	Heat cure
More than 51	Heat shock may occur

A)Heat Index Classification

Dew Point (°C)	Perception
Greater than 27 °C	Serious
21–26 °C	Very annoyed
16–21 °C	Pretty annoyed
10–15 °C	Comfortable

B)Dew Point Classification

Next, we classify traffic flow and velocity by value because the days having similar weather patterns will have similar taxi's flow and similar taxi's moving speeds. Finally, with the feature set that changed during each time slot, we used two algorithms, K nearest neighbor (KNN) and random forest (RF) for predictions.

4.3 Experimental results and evaluation

To evaluate the effectiveness of the model, we use Accuracy measurement. The accuracy (denoted ACC) is calculated by Equation 11.

$$ACC = \frac{number\ of\ correct\ predictions}{number\ of\ predictions} \quad (11)$$

Table 5 and table 6 show the accuracy of built models for predicting flows and velocity in 10 high traffic areas and poor traffic conditions. Where the blue columns represent the K-Nearest Neighbor (KNN) algorithm with different K values, the green column represents the Random Forest (RF) algorithm, the final line is the average ACC of each color model in which red marks the best model.

Table 5 shows that the taxi flow prediction model with the KNN method and K = 7 gives the best average result. Table 6 shows that the velocity prediction model with the best ACC is KNN with K = 8. However, ACC's predictions in some areas are not high because of these chaotic traffic or speed changes due to other factors (such as traffic accidents or some events).

In this study, KNN is most likely to produce better results because each weather stage will have different weather patterns and usually lasts from one week to two weeks. During this time, the weather will be similar each day so the rules of travel will also be similar. KNN uses similar dates for predictions so it can be seen that KNN has the practical implementation approach. The RD results are less exact than the KNN's because RD considers each factor and can ignore some elements in the training process.

Table 4: Accuracy of models predicting taxi traffic flow

Test	K=3	K=4	K=5	K=6	K=7
1	72.5	66.7	70.6	66.7	70.6
2	60.8	74.5	72.5	68.6	72.5
3	88.2	86.3	86.3	90.2	88.2
4	58.8	64.7	62.7	58.8	62.7
5	74.5	70.6	74.5	78.4	76.5
6	80.4	76.5	76.5	78.4	86.3
7	84.3	86.3	86.3	86.3	88.2
8	60.8	64.7	60.8	62.7	62.7
Mean	72.54	73.79	73.78	73.76	75.96

Test	K=8	T=64	T=96	T=128
1	66.7	72.5	72.5	70.6
2	74.5	60.8	60.8	56.9
3	90.2	82.4	84.3	82.4
4	56.9	60.8	62.7	56.9
5	78.4	72.5	78.4	74.5
6	82.4	82.4	82.4	78.4
7	84.3	84.3	86.3	86.3
8	58.8	58.8	56.9	54.9
Mean	74.03	71.81	73.04	70.11

Table 6: Accuracy of models predicting velocity

Test	K=3	K=4	K=5	K=6	K=7
1	68.8	72.7	76.7	70.8	70.8
2	74.7	80.6	82.5	80.6	86.5
3	78.6	74.7	74.7	72.7	74.7
4	59	51.2	59	62.9	53.1
5	66.9	61	62.9	66.9	59

6	64.9	59	64.9	51	62.9
7	64.9	64.9	59	68.8	70.8
8	68.8	70.8	72.7	68.8	74.7
Mean	68.33	66.86	69.05	69.06	69.06

Test	K=8	T=64	T=96	T=128
1	72.7	70.8	68.8	68.8
2	82.5	74.7	74.7	74.7
3	76.7	64.9	64.9	74.7
4	61	53.1	57.1	53.1
5	68.8	61	57.1	61
6	66.9	47.3	45.3	45.3
7	66.9	62.9	64.9	64.9
8	76.7	70.8	70.8	62.7
Mean	71.53	63.19	62.95	63.15

From the built-up models, we will proceed to predict the velocity in the next time and the next day. Figure 8 shows the heat map of some regions in an hour with the red areas having the lowest velocity then increasing, the colorless areas being the places where no taxis moves.

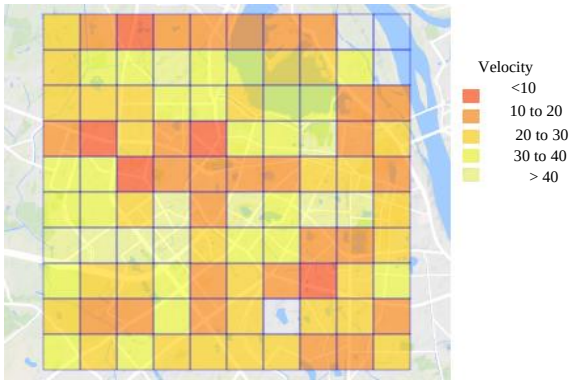


Figure 8: Heat map of average velocity of taxis

V. CONCLUSION

In this study, we have solved two problems using GPS data of taxi.

1) Finding 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, the cause can not be identified and solutions can not be offered for some specific pairs of regions.

2) Developing traffic flow and velocity models with the accuracy of 76% and 72% allowing drivers and managers to know future traffic patterns in regions which have not frequent good traffic conditions. However, the predicted results in some areas are not high.

In the future, we intend to continue to work on Problem 1 with a number of other ways of partitioning regions to better pinpoint the problem and propose solutions. We will also continue to work on Problem 2, taking into account the direction of travel in each region, proceeding with different group divisions, and incorporating more elements, applying additional learning algorithms to increase the accurancy.

REFERENCES

[1] 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).

[2] 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.

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

[4] 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)

[5] 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)

[6] 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)

[7] 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

[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] 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

[10] 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

[11] 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

[12] 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

[13] 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)

[14] 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)

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

PHÁT HIỆN TƯƠNG HỖ VÙNG ÙN TẮC VÀ TIÊN ĐOÁN LƯU LƯỢNG VÀ VẬN TỐC DI CHUYỂN CỦA TAXI

Tóm tắt: Ngày nay, taxi là một trong những phương tiện di chuyển phổ biến nhất. Có một số lượng lớn người đi lại sử dụng taxi hàng ngày và hành trình của taxi thể hiện sự di chuyển của con người. Ở những thành phố lớn, taxi đều được trang bị thiết bị GPS và chạy 24 giờ mỗi ngày, chúng sẽ được sử dụng để trích ra những thông tin tin cậy về tình trạng giao thông. Bài báo này phát biểu phương thức của chúng tôi dùng hành trình của taxi tại Hà nội, Việt nam trong 4 tuần từ 18 tháng 9 tới 15 tháng 10. Trong đó, bản đồ Hà nội được chia thành các vùng nhỏ hơn với một kích cỡ được xác định trước. Sau đó, chúng tôi xác định các vùng kế cận nhau nơi có các ùn tắc xảy ra trong các khoảng thời gian khác nhau và mối tương quan giữa chúng. Cuối cùng, chúng tôi xây dựng một mô hình tiên đoán lưu lượng di chuyển của taxi trong mỗi vùng và vận tốc của chúng dựa trên dữ liệu lịch sử và thời tiết.

Từ khoá: Tiên đoán lưu lượng di chuyển của taxi, ùn tắc vùng kế cận, vận tốc.

Nguyen Quynh Chi is currently a senior lecturer of the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam. She received a M.Sc. in Computer Science in University of California, Davis, USA (UCD) and became PH.D Candidate at UCD in 2004 and 2006, respectively. Her research interests include machine learning, data mining, and testing algorithms.