Riding Pattern Analysis of Taxi Passengers

(A Case Study of a Korean Taxi Company)

Fransiskus Tatas Dwi Atmaji
Engineering Department, Industrial and Systems Engineering School, Telkom University
40257, Bandung, Indonesia

Kwon Young Sig
Information System Laboratory (ISL), Industrial and Systems Engineering Department, Engineering School, Dongguk University,
Pil-dong, Jung-gu, Seoul, 100-715, Republic of Korea

***Abstract*— In an urban area, including Seoul city, many people prefer to use taxi as one of the choices for fast, convenient and flexible transportation. Based on the data provided by one of the taxi service call companies, X, in Korea, this paper proposes a taxi passenger riding pattern analysis at Seoul city, Republic of Korea. Using a multidimensional analysis method, we analyze the data to obtain an area with high potential demands of taxi passengers, which we call "hotspot" area. The results show that "Teheran Valley" in Gangnam-gu and Dongdaemun market in Jung-gu and Jongno-gu had the biggest number of passengers compared to other areas. This information will be useful for the company to re-evaluate its current taxi fleet management, especially for the taxi's drivers in minimizing their idle time when searching for customers.**

Keywords—GPS coordinates; multidimensional analysis; riding pattern analysis; hotspot area

I. INTRODUCTION

In daily life, many people prefer to use taxi as one choice of fast, convenient and flexible transportation. In an urban area like Seoul city, taxi also becomes the favorite choice for traveling. With the technology development and modernization, nowadays almost every taxi is equipped with a Global Positioning System (GPS) which can be used in the taxiôs operating system, especially to manage the navigation problem. Using this technology the correct location can be acquired in time especially for novice taxi drivers with no previous experience. Significant locations and predicted movements across multiple users using GPS hardware were investigated to collect the location dataset which automatically clusters GPS data into meaningful locations at multiple scales [1]. Even though equipped with the GPS system, there are other problems faced by taxi drivers. Typically, the taxi drivers search for their customers based on their intuition and experience only, go to a crowded area where customers may be available besides relying on taxi calling service, which means they come to a specific place ordered by a customer via telephone call. As a result, sometimes they still have ineffective time with absence of passengers for a long time and just waste their time and energy to search the passenger.

As reported in a Taipei urban area, about 60-73 % of their operation hours, taxi drivers were driving without passengers [2]. Also, at Jeju Island, Korea, many taxi drivers reported that about 80 % of their activity was carrying no passengers [3].

To minimize the problem above, the main objective of this research is analyzing the taxi passenger riding behavior, especially in Seoul City, in order to obtain valuable information from the huge number of taxi databases. Using the multidimensional analysis method, the variables time, location, and frequency will be analyzed. The proper information of taxi passenger riding behavior is one of the important aspects in a taxi business area because, according to the Seoul Statistical Year Book, around 6.5 % people in Seoul use taxi as their transportation mode [4].

II. DATA AND METHODOLOGY

A. Data and research framework

Data for this research were provided by one of the taxi companies in Korea, which is defined later as Company X. The 400.000 raw data used for the analysis are those of a 24-hour taxi service.

TABLE I. THE EXAMPLE DATA

Y/M/D	Taxi ID	Get ON time	Get OFF time	GPS longitude (GET ON)	GPS latitude (GET ON)	GPS longitude (GET OFF)	GPS latitude (GET OFF)
20101003	1035922307	10:59:09	11:45:30	127.056959	37.591416	127.09276	37.613859
20101004	1037912431	22:58:00	23:15:09	127.050851	37.504816	126.974691	37.519674
20101006	1047802409	13:35:31	14:01:39	127.07382	37.546922	126.900754	37.530143

The example of processed data is shown in Table 1. The data consist of date (year, month, and day), taxi ID, time (get on and get off time), and GPS coordinates (get on and get off latitude and longitude coordinates). In this paper we use term of ñGet onò and ñGet offò as a pick-up and drop-off of taxi passengers respectively.

Fig. 1. General research framework

As is depicted in Fig.1, raw data from the taxi company database will be processed using multidimensional analysis to find the information about the passenger riding pattern. Pre-processing data is needed in order to prepare the raw data from the company into proper data for further analysis. The raw data from the company cannot be directly used for this analysis because it used ORACLE format databases which different from our analysis platform. Using an additional calculation formula, data can be read and transformed into our data analysis platform.

This research is limited only in the Seoul city area. According to the geographic condition and administration, Seoul area comprises around 605.25 km², with a radius of approximately 15 km from the north to the south, roughly halves by the Han River; it consists of 25 districts (-gu) and 522 sub-districts (-dong).

In this research, we used four (4) time span allocations to identify the distribution of taxi passengers by dividing the 24 hours into 4 time spans:

(1) Morning time (07.00-10.00 am)
(2) Afternoon time (10.00 am-18.00 pm)
(3) Night time (18.00-24.00 pm)
(4) Late night/early morning (00.00-07.00 am).

B. Pre-processing data

The most important thing for pre-processing the data in this research is converting the GPS coordinates data into a location or address. GPS coordinates consist of Latitude and Longitude coordinates. Latitude is the angular distance of the North or South locations of the Equator and usually denoted by the Greek letter phi (ű). Longitude is the angular distance of a point's meridian from the Prime (Greenwich) Meridian and denoted by lambda (ə). The latitude and longitude in general are used for a geographic coordinate system to specify any location on the globe1.

In this research, The Delphi Geo-code software is used to convert the GPS latitude and longitude coordinates into the location. This software program is connected with a Google location database which automatically searches the location. In this case the result of the location is in Korean character (Hangeul) This software can search the address in details, including the district and the sub district area (-gu) and (ï dong), for example : 대한민국 서울특별시 강남구 대치 4동 89). The results of Delphi Geo-code software are in a Shape file (.shp, .shx and .dbf) format, and using the DBF Viewer Plus software, the Shape file is converted into a more general format for the analysis.

C. Multidimensional Analysis

A multidimensional analysis is a data analysis process that groups data into two or more categories of data dimension and measurement. After the process of analyzing the data, the result of this research will help the company to optimize and increase their taxi driver working performance, for example: to re-evaluate their current taxi fleet management. The better the understanding of the passenger behavior, the higher the companyôs profit earned.

In this research the variables used for the multidimensional analysis method are time, location, and frequency, as shown in Figure 2. By this analysis, we will get the detailed information about not only where, when, and how many taxis passengers are distributed but also how long they ride the taxi.

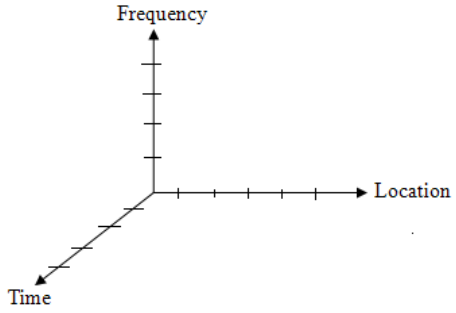

Fig. 2. Three variables for experiment

Using the SAS Enterprise Guide 4.3 as the analysis tool, the variable location, frequency and time data can be separated and grouped. From the multidimensional analysis we can obtain some information that will be useful for the company to optimize their fleet management and increase the company profit, for example: distribution of taxi passenger, passenger riding behavior and other information.

III. EXPERIMENT RESULT

A. General taxi passengers distribution

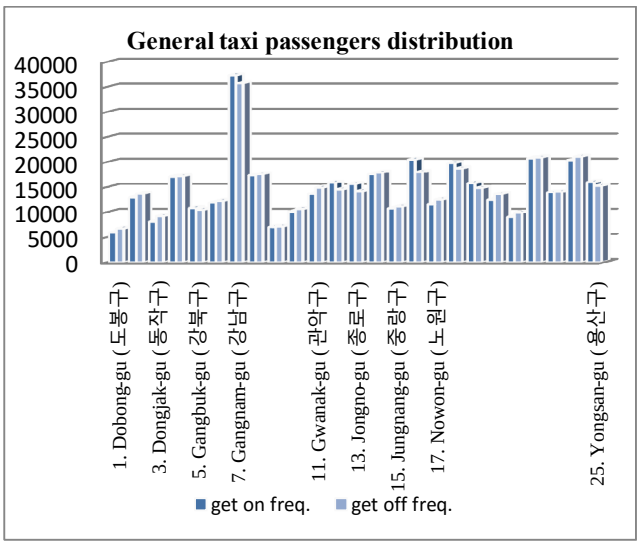

Fig. 3. General taxi passengers distribution

In As shown in Figure 3, out of 400.000 raw data, 93 % of the passengers (37.2357 passengers) ride taxi from Seoul city area, especially in Gangnam-gu, which has the highest frequency of passengers, compared to other districts. The remaining 7 % get on at outside of Seoul city or in the nearby city, for example: Incheon City or Gyonggi province.

It is noted that the analysis focuses on the get on distribution of passengers only. The get on passenger information give the information to determine the target area (hotspot) where many passengers probably need the taxi service.

B. Inside, adjacent and remains area analysis

The ñadjacentò district is defined as the district bordering or nearby the main district analyzed. By using this adjacent district analysis, we obtained the information about the taxi passenger riding behavior across other district areas.

Figure 4 shows the example of the adjacent area. From this figure, it is shown that Jung-gu (fixed line) has 6 adjacent areas. The adjacent areas (dotted) are: Jongno-gu, Dongdaemun-gu, Seongdong-gu, Yongsan-gu, Mapo-gu and Seodamun-gu. The areas not included in the adjacent districts will be called the óremainsô area.

Fig. 4. The illustrations of inside and adjacent areas in Seoul City

TABLE II. JUNG-GU TAXI PASSENGER RIDING BEHAVIOR

Seoul city	TOTAL AVERAGE PERCENTAGES FROM 25 DISTRICT (%)		
	Inside	*Adjacent*	*Remains*
07.00-10.00 (Morning)	43.37	36.11	20.52
10.00-18.00 (Afternoon)	47.40	35.64	16.97
18.00-24.00 (Night)	35.98	32.78	33.21
24.00 -07.00 (Late night/early morning)	31.53	33.16	36.06

Table 2 gives the example of the analysis result in Jung-gu area. As shown in this table, the taxi passengers prefer to ride the taxi inside their district area in the morning. Meanwhile in the afternoon, the passengers prefer to ride the taxi across their district (outside area). In the night and late night/early morning the remains area have the higher percent

The analysis results of passenger riding behavior from 25 districts in Seoul city are tabulated in Table 3. As shown in this table, we can see that in the morning time (07.00-10.00) most of the passengers prefer to ride the taxi inside their district (43.37 %); in the afternoon time the result analysis is similar with that in the morning; most of the taxi passengers prefer to ride from their district (47.40 %). Also in the night time; the result of the analysis is similar with that of the morning and afternoon time when most passengers prefer to ride taxi inside their district (35.98 %). The different results of analysis show in the late night/early morning time, the remains area have higher percentages than in the inside and adjacent areas, in which the taxi passengers prefer to ride the taxi (36.06 %).

TABLE III. SEOUL TAXI PASSENGER RIDING BEHAVIOR

Seoul city	Total average percentages from 25 district (%)		
	Inside	*Adjacent*	*Remains*
07.00-10.00 (Morning)	43.37	36.11	20.52
10.00-18.00 (Afternoon)	47.40	35.64	16.97
18.00-24.00 (Night)	35.98	32.78	33.21
24.00 -07.00 (Late night/early morning)	31.53	33.16	36.06

From the tables above, we have obtained the information that most of the passengers ride taxi for a short distance, inside their district but rarely use taxi for the long distance, except at the late night/early morning. In general, the passengers will use the taxi in a long distance when they need are under an urgent condition or another situation when the service of public transportation such as bus or subway is not available anymore.

C. Taxi riding time analysis

Riding time is one of the important aspects of taxi passenger analysis behavior. Based on the get on and get off time data, the riding time can be obtained by alleviating the get off time by get on time. Basically, this result gives the information about ñhow long the timeò spent by passengers in the taxi.

The riding time analysis result shows that most of the passengers ride taxi in a short time, between 5-10 minutes (30 %) followed by 15-30 minutes (22%) and 1-5 minutes (20 %). Figure 5 shows the analysis result.

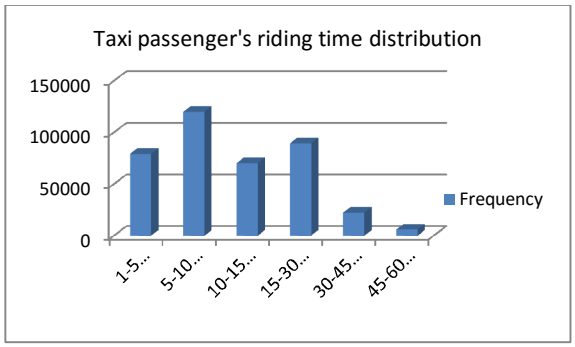

Fig. 5. Taxi passengers riding time distribution

It proves that the previous analysis result which concluded that most of the passengers used taxi inside their districts. The short riding time means the short riding distance.

D. The "hotspot" area analysis

From the whole analysis, finally we found which area has a potential demand for the taxi passengers or we call it

ñhotspotò area, an area that had higher numbers of passengers compared to the others. The first is Teheran Street in Gangnam-gu and the second is Dongdaemun Market in Jung-gu & Jongno-gu. Figure 6 shows the visualization of Teheran street or known as ñTeheran Valleyò using Google Earth software application.

Fig. 6. View of ñTeheran Valleyò via Google Earth

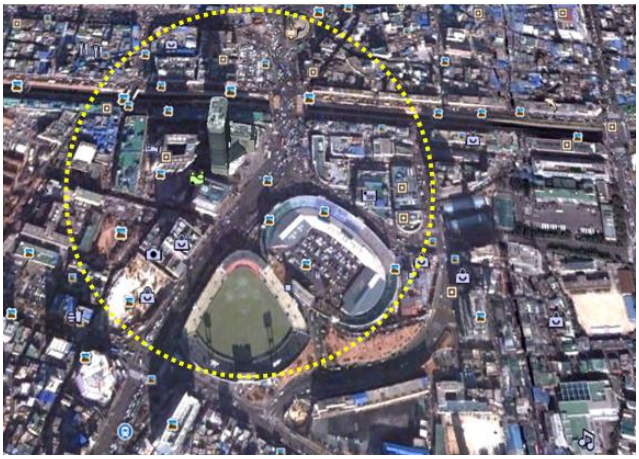

Fig. 7. View of Dongdaemun market via Google Earth

Figure 7 shows the visualization of Dongdaemun market area using Google Earth software application. It is noted that 35 % of the taxi passenger distribution in Gangnam-gu was found in ñTeheran valleyò and 25 % of taxi passenger distribution in Jung-gu & Jong-no-gu was that of the passengers in Dongdaemun market.

IV. CONCLUSION

Basically this paper analyzes the distribution of taxi passengers in Seoul city. Using the multidimensional analysis, the raw data can be transformed into valuable information. The variables used in this analysis are location, time, and frequency. Before analyzing the data, the raw data from one of Korean taxi companies which contains of GPS longitude and latitude coordinates are transformed into real locations/addresses. Using the Delphi Geocoding Software which is connected to Google location database, the GPS coordinates can be converted into locations and the results are in Korean characters (Hangeul).

Even though the distributions of taxi passengers in Seoul city are relatively prevalent, some areas in certain districts show preeminent frequency. From the general distribution of taxi passengers, the Gangnam-gu area had the highest taxi passenger distribution (35%). Gangnam-gu which is known as a business area, had the highest potential of taxi passengers compared to the others. In this case the company can add the additional dispatching fleet to this area, especially in ñTeheran Valleyò where the traffic time are conducted at the morning rush hours (07.00 am-10.00 am) and in the late night/early morning (01.00 am-02.00 am).The distribution of the taxi drivers in shopping areas should also be managed well by company. The analysis result shows where Dongdaemun market gives a significant contribution and has the highest rate of taxi passengers (25 %), with the traffic time between 14.00 pm to 19.00 pm, when many people go out for shopping.

The inside, adjacent and remains area analysis results give more detailed information about taxi passengers behavior in Seoul city. The results show that most of the passengers ride taxi for a short distance, inside their district and rarely use taxi for the long distance, except in the late night/early morning time. From the riding time analysis we found the fact that most of the customers ride taxi in a short time (less than 30 minutes) with the highest interval time between 5-10 minutes (30 %). It proves that most of taxi passengers ride the taxi inside their district not to the adjacent or remains area.

In general, this research gives an illustration of the taxiôs passenger riding distribution and taxi passenger behavior, especially in Seoul city area. Based on this analysis, the company can get the valuable information about their customer. It enables them to re-evaluate their current fleet management in order to optimize and increase their taxiôs driver working performance.

References

[1] D. Ashbrook, and T. Starner, ñUsing GPS to learn significant locations and predict movement across multiple usersò, Personal and Ubiquitous Computing, vol. 7, no. 5, pp.275-286, 2003.

[2] H.W. Chang, Y.C. Tai, and J. Hsu, ñContext ï aware taxi demand hotspots predictionò International Journals of Business Intelligence and Data Mining, vol. 5, no. 1, pp. 3-18, 2010.

[3] J. Lee, I. Shin, and G.L. Park, Analysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation, Fourth International Conference on Networked Computing and Advanced Information Management, 2008, pp.199-204.

[4] SDI Internal Data, Seoul Metropolitan Government, The study of construction of complex transfer center, Seoul Statistical Year Book. 2007.