携帯ビッグデータを利用した人々の鉄道利用のモニタリングに関する研究

A study on railway passenger monitoring with mobile big data

学籍番号　47-176826

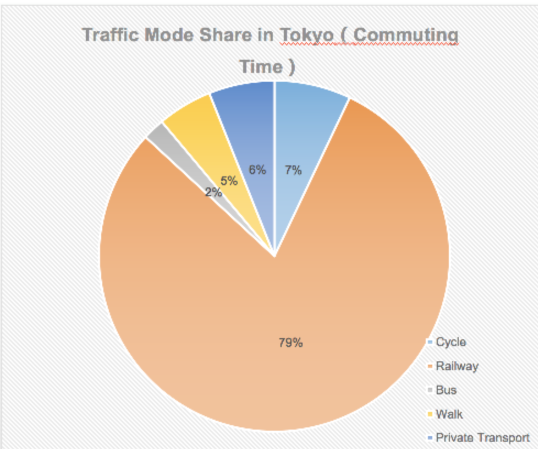
氏　名　黄 殊哲（Huang Shuzhe）

指導教員　柴崎 亮介 教授

小林 博樹 准教授

Nowadays, study of public transportation of Tokyo becomes a very popular and interesting topic, as public transportation, especially railway, occupies the highest rate in all of the traffic modes chosen by Tokyo residents.

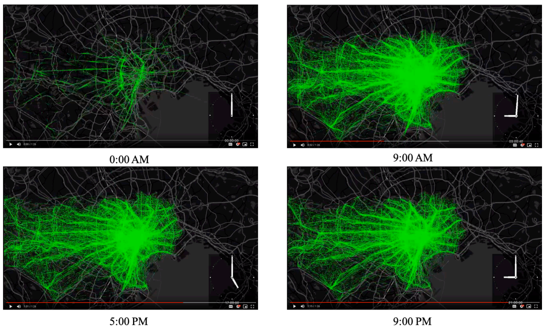
In this picture of traffic mode share in Tokyo, we can see that railway has the highest rate, about 48%, which is much higher than other traffic modes. And in the commuting time, the rate will increase to 79%, which is markedly high.



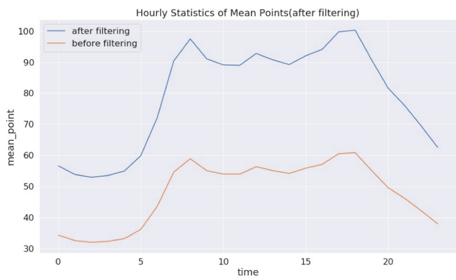
In this research, we mainly use Target Ads GPS data for analysis. As Target Ads GPS data is a new data source, many former researches focused on using navigation GPS data for studying human mobility as well as transportation system. Thus, we will use the navigation GPS data for comparison and provide another aspect to evaluate our result. Moreover, some collected GIS data will also be utilized to support this research.

As the backbone of all of the data used in this research, Target Ads GPS Data is a data collected by Location Based Services (LBS) from a private company. The data aims to track daily activities of each user, analyze their behavior and feedback some useful information (advertisements) for them. The location information is collected by users' smartphone, and it is utilized to satisfy the 'Mobility as a Service' requirement from diverse users.

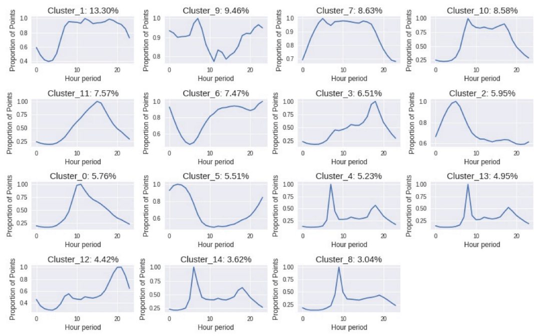
An example of visualization of Target Ads GPS Data is shown in following figure. It is clearly displayed that the dynamic population flow change in different point-in-time of one day.



As Target Ads GPS data is a raw GPS data with tremendous amounts of points, data filtering is an indispensable step to delete those data with unsatisfactory accuracy and decrease the noise. In order to define the standard of filtering, it is necessary to make a basic statistic of the data previously. In this statistic work, we calculated the total amounts of IDs, total records of GPS points, Average records of each Ids per day, maximum and minimum points of each IDs. After the data filtering, the data become more dense and homogeneous.



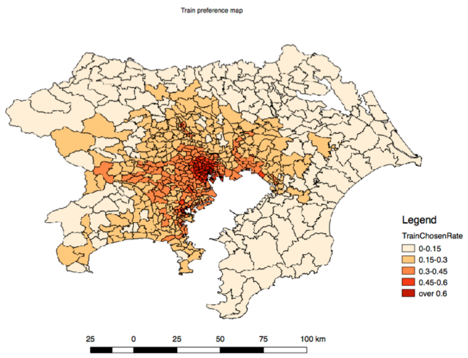
Understanding life patterns of residents is an important task to reflect an aspect of the life situation of residents. In this research, on the one hand, this work is regarded as a part of mobility analysis to show the life pattern of Target Ads GPS users. The most common life patterns of residents in Japan is performing active in the daytime, which indicates those people with a job or students. As Japan is a country with complex and convenient railway transportation system and about 80 percent of people prefer to utilize trains for commuting. Thus, studies on monitoring railway passengers have a great significance in Japan.



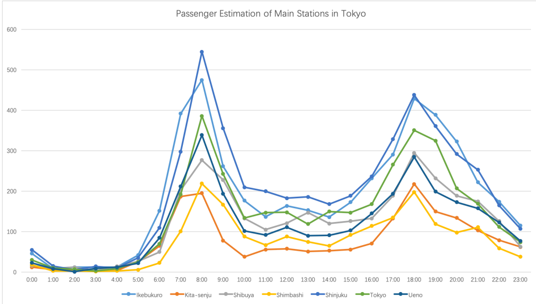
As the function of life pattern detecting is not only providing a result to help us understanding the residents behavior, it also makes sense in improving the accuracy of map matching. As we imported a web-based system to collect training data and evaluate the validity, the classification of a variety of life patterns can support us to recognize those trajectories which are difficult to identify its OD information as well as traffic mode. There are also some failures and noises in map matching. The main reason which causes those failures is the average time interval of Target Ads GPS data is larger than Navigation GPS data, thus the GPS points are easily far from each other, then decrease the accuracy. Besides of this reason, the confusion of different railway system and the error of topology information can also bring failures in the map matching work.

Linear interpolation is regarded as an another essential pre-processing work of GPS data mining. For there are always some signal loss in GPS dataset, to improve the accuracy of trajectory data, data interpolation is a methodology conducted to supplement and estimate the lost data. On the other hand, data interpolation can help us to arrange the data more regular and make it convenient to extract the points in a specified point-in-time. In this research, we chose 5 minutes as the fixed time interval. After data interpolation, the dataset become more dense and precise.

In this research, mobile big data is the principle data source for rail transit passenger analysis. However, a brief analysis via GIS Open Data can support us to understand the overall transfer preference of residents. We choose Tokyo city area as the study region, and try to analyze the preference the usage of rail transit in Tokyo. As it is shown, the residents of center part of Tokyo prefer to trip by train. Under this background, the analysis of railway passengers in this research will focus on the center part of Tokyo Metropolitan Area.

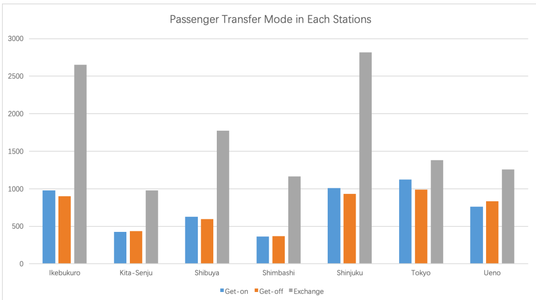


Thus, we chose seven main stations for this study: Shinjuku, Shibuya, Ikebukuro, Ueno, Tokyo, Shimbashi and Kita-senju. All of those stations are important transportation junctions in Tokyo. At first, a basic statistic was made to show the passenger flow of each time distribution in those station which is shown in following figure.

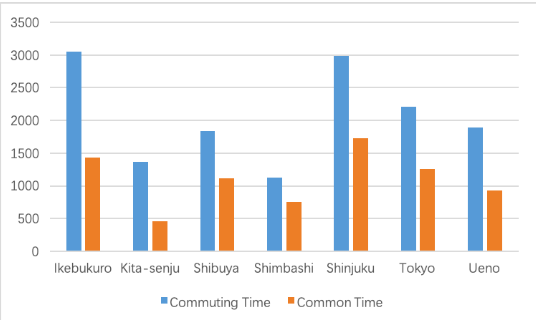


In this graph, we can see that the hourly variation tendency of passenger flow is similar in different stations. The peaks of passenger amounts occur in the time period of 8:00-9:00 and the time period of 18:00-19:00, which replaces the ‘rush hour' of commute to work place and home place. In the period from 10:00 to 17:00, the line is steady, for people don't assemble for commuting during this period.

Then, it is important to extract further information from those stations. As the transfer status has been recognized, the distribution of different transfer modes includes get-on, get-off and exchange can be shown in following figure.



To understand the performance of each station, an analysis of stations' utilization in commuting time is important and even essential, as about 80 percent of Japanese residents use train for commuting.The Figure shows the utilization of passengers in commuting time and common time of each station. Ikebukuro station and Ueno station has relatively higher rate of passenger volume in commuting time, while the rate of kita-Senju station is prominently high, reaches 74.6¥%. Meanwhile, the occupation of passenger volume in commuting time is similar in remain stations, values around 60¥%. This results indicates that Kita-Senju station plays a more important role in commuting time rather than common time, the passenger flow of Kita-Senju station in common time is much lower than other stations. Within the selected stations in our research, according to the result of passenger estimation, some stations are mostly utilized by passengers as an interchange station. Moreover, the percentage of passenger utilization in commuting time is different in each station. Kita-Senju station is a typical example, most of passengers utilize it for commuting and exchange.



Based on those analysis, we can have a direct-viewing impression of the result of railway transportation monitoring in Tokyo. In the next section, we will make a comparison between the estimation from Target Ads GPS data with the result of real census data and Navigation GPS data, to provide an another view of railway transportation monitoring as well as assess the accuracy.

The estimation result of regression analysis in Shibuya Station, Shinjuku Station and Tokyo Station is close to the real data, meanwhile the result of Ikebukuro Station, Kita-Senju Station and Shimbashi Station is larger than the census data. Moreover, the result of Ueno station is much larger -- about three times than the real census data. In the result of comparing commuting passenger occupation, the only error still happens in Ueno station, whose value of GPS estimation is much larger than the real rate. It may be caused by the extremely complex placement of railway transportation stations in Ueno area, and it is important to find ways to increase the accuracy in this kind of stations in the future.

Station	Commuting Passenger Rate of GPS Estimation	Commuting Passenger Rate of census data
Ikebukuro	68.0%	70.1%
Kita-Senju	74.6%	73.9%
Shibuya	62.3%	63.7%
Shimbashi	59.8%	61.6%
Shinjuku	63.3%	65.5%
Tokyo	63.7%	61.0%
Ueno	**67.1%**	**57.4%**

	Census Data	GPS Passenger Counts	Predicted Passenger Count
Ikebukuro	**3156867**	**4528**	**4519641**
Kita-Senju	**1399698**	**1842**	**1838698**
Shibuya	3576364	2994	2988539
Shimbashi	**1144719**	**1892**	**1888604**
Shinjuku	4663466	4762	4753200
Tokyo	3098325	3491	3484593
Ueno	**1073411**	**2848**	**2842754**

Finally, we made a comparison with the result of Navigation GPS data, the total accuracy of Target Ads GPS data is similar but slightly lower. And the estimated passenger counts of Target Ads GPS data in smaller, contrary to its larger amounts of users. This phenomenon points out that the map matching methodology is necessary to be optimized in the future.