

東京大学大学院新領域創成科学研究科
社会文化環境学専攻

2019 年度
修 士 論 文

A Study on Railway Passenger Monitoring with Mobile
Big Data
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

2019 年 7 月 12 日提出
指導教員 柴崎 亮介 教授

黄 殊哲
Huang, Shuzhe

Abstract

As railway is one of the most important part in the transportation system of Japan, it occupies very high rate in the traffic mode selection of residents. Thus, monitoring railway passengers as well as their living patterns is of great significance in building 'Mobility as a Service' system, which means provide useful information to passengers via mobile apps, based on their mobility. With the development of Location Based Services(LBS) and big data mining technology, studying and analyzing the mobility of passengers via big mobile data becomes a popular issue to research in recent days. Under this background, in this research a new source of mobile GPS big data was utilized to detect the mobility and estimate the movement of railway passengers. Meanwhile, the study also provides methodologies to process the large and heterogeneous raw GPS big data and then make it available to extract information. In the end, a comparison with real census data and the similar estimation results from former GPS data sources is also provided in this research. With all of those mentioned works, this research will provide a multi-aspect perspective to understand the mobility of railway passengers.

Keywords: GPS big data mining, traffic analysis, mobility analysis, machine learning

Acknowledgements

At First, I'd like to appreciate my adviser Prof.Shibasaki and co-adviser Prof.Kobayashi and Prof.Song. Prof.Shibasaki gave me too much support for my research and Prof.Kobayashi and Prof. Song gave me many valuable suggestions, without their help, I'm afraid that I can not complete the thesis smoothly. At the same time, as well as my advisers, Prof.Song also provided me many assists for my research. He taught me lots of knowledge and methodologies about mobile big data mining and experiment constructing. Those advisers of my thesis taught me a lot, and supported me a lot to get through problems which I faced in my research.

In addition, many other members in Shibasaki Lab also gave me some help and supports. I'd like to express my thanks to some researchers, especially Haoran Zhang, Renhe Jiang, Zipei Fan and Qunjun Chen. They taught me a lot on my research and they were always glad to answer my questions. Besides, there are many senior and junior students who supported and encouraged me a lot. I appreciate Tianqi Xia, Dou Huang, Zekun Cai and Jinyu Chen, they are doctoral and master students who have strong ability of research and gave me many advice on my study and research. And thank Yuxuan Wang, Zhiling Guo, Xiaodan Shi, Xiaoya Song and other members who works in room 435 of Kashiwa Research Complex for their encouragement and assist in my daily life. Thanks the group members of data infrastructure, Mr. Witayangkurn, Mr.Kanasugi and Mr.Matsubara gave me many supports to help me to access the data, which is necessary for my research.

Finally, thank my parents as well as my family, my girl friend and my friends who always give me various help in my daily life. They are very solid backing to me.

Contents

Abstract	iii
Acknowledgements	iv
List of Figures	vii
List of Tables	ix
1 Research Background	1
2 Past Studies and Research Works	5
2.1 Research on Public Traffic System	5
2.2 Research on Traffic Monitoring and Evaluating Public Traffic System via GPS big data	6
3 Framework and Methodology	9
3.1 Framework	9
3.2 Life Pattern Classification	10
3.2.1 Data Filtering	10
3.2.2 Life Pattern Clustering	11
3.2.3 X-means Clustering	11
3.3 Mobility Analysis	12
3.3.1 Map Matching	12
3.3.1.1 Introduction of Map Matching Methodologies	12
3.3.1.2 HMM Based Map Matching	13
3.3.2 Transportation Mode Estimation	16
3.3.2.1 Stay Point Recognition	16
3.3.2.2 Transport Point Recognition	18
3.3.2.3 Transportation Mode Classification	19
3.3.3 Analysis of Passengers	20
4 Data Process and Experiment	23
4.1 Introduction of data sources	23
4.1.1 Navigation GPS Data	23
4.1.2 Collected GIS Data	24
4.1.3 Target Ads GPS Data	25
4.2 Data Filtering	26

4.3	Mobility Analysis	28
4.3.1	Life Pattern Observation	28
4.3.2	Preliminary of Railway Passenger Estimation	31
4.3.2.1	Map Matching	31
4.3.2.2	Data Interpolation	31
4.3.3	Analysis of Rail Transit Passengers	32
4.3.3.1	Preference Analysis via GIS Open Data	32
4.3.4	Railway Passenger Estimation via Mobile Big Data	33
4.3.5	Comparison and Accuracy Assessment of Passenger Estimation	37
4.3.5.1	Comparison with Census Data	37
4.3.5.2	Comparison with the Result of Navigation GPS Data	38
4.4	Review and Discussion	39
5	Conclusions and Future Work	41
5.1	Conclusions	41
5.2	Future Works	42
	Bibliography	43

List of Figures

1.1	Chart of traffic mode share of common time and commuting time in Tokyo . . .	1
1.2	Annual traffic accident counts	2
1.3	Share of railway usage in people with different gender and age	3
3.1	Framework:Estimate Human Mobility and Movement via big mobile GPS data	10
3.2	An example of Hidden Markov Model	14
3.3	An simple example of HMM map matching: output a best line with highest probability	15
3.4	An flowchart of procedures before traffic mode detection	17
3.5	Screen of the web-based labeling tool	19
3.6	Methodology of Extracting Information of Railway Passengers	20
4.1	An Example of Visualization of Target Ads GPS Data in Different Time of One Day	26
4.2	Total Days and Points Distribution of Target Ads GPS Raw Data	27
4.3	Hourly Statistics of Mean Points	28
4.4	Total Days and Points Distribution After Filtering	28
4.5	BIC variation of each amount of clusters	29
4.6	Groups of Activity Patterns	30
4.7	User Number of Each Clusters During Active Periods	30
4.8	The Rate of Rail Transit Usage in Different Districts of Tokyo Metropolitan Area	32
4.9	The Rate of Other Transit Mode Usage in Different Districts of Tokyo Metropolitan Area	33
4.10	The Favourite Transit Mode in Different Districts of Tokyo Metropolitan Area .	34
4.11	The Hourly Estimation of Passenger Flow by Stations	34
4.12	The Transfer Status of passengers in each stations	35
4.13	The Passengers of Commuting Time and Common Time in Each Stations . . .	36
4.14	Passenger Estimation of Shinjuku Station with Target Ads GPS data and Navigation GPS data	39

List of Tables

4.1	Dictionary of Navigation GPS Trajectory Data	23
4.2	Dictionary of Railway Lines Data	24
4.3	Dictionary of Railway Stations Data	25
4.4	Dictionary of Traffic Flow GIS Data	25
4.5	Dictionary of Target Ads GPS Data	25
4.6	Basic Statistics of Raw Data	27
4.7	Data Amounts Before and After Filtering	27
4.8	Table of Passenger Share in Commuting Time	36
4.9	A Comparison with Real Census Data	37
4.10	Table of Comparison of Passenger Occupation in Commuting Time	38

Chapter 1

Research Background

Nowadays, study of rail transit 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. According to statistic data from Ministry of Land, Infrastructure, Transport and Tourism(MLIT), Figure 1.1 shows that railway occupies 48 percent of traffic preference of residents in Tokyo, and for those commuting time, it increases to 79 percent – a very high percentage which shows people in Tokyo always prefer to choose railway to go to work. Hence, it comes very meaningful to study the performance of public transportation in Tokyo and it is a meaningful and even necessary work for traffic planning and building intelligent transportation system in Tokyo.

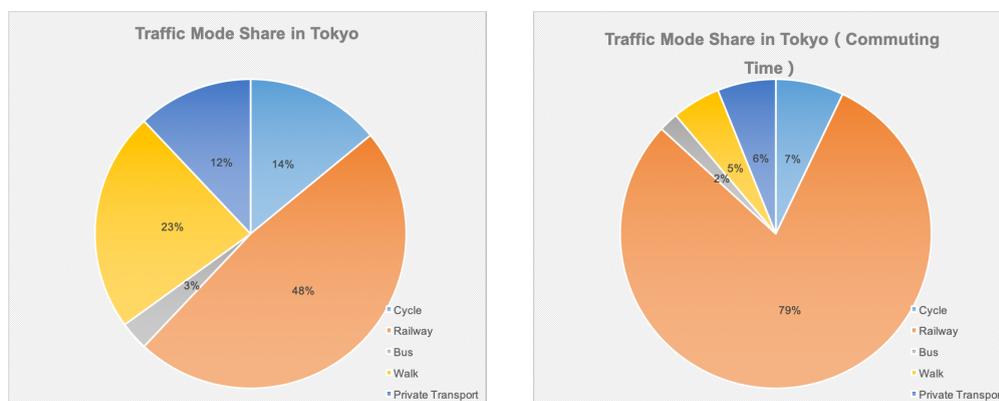


FIGURE 1.1: Chart of traffic mode share of common time and commuting time in Tokyo

But at the same time, though both of those two charts can show the importance of railway transportation in the whole traffic system in Tokyo, they are quite distinguishing for the different rate of traffic modes. In that sense, detect the performance of traffic system in different situation

should be a serious issue. On the one hand, there are always some accidents happening which can influence the traffic system. Figure 1.2 informs that the total trend of accidents is increasing in Tokyo(Data Source: MLIT). Those accidents always cause delay of public transportation service.

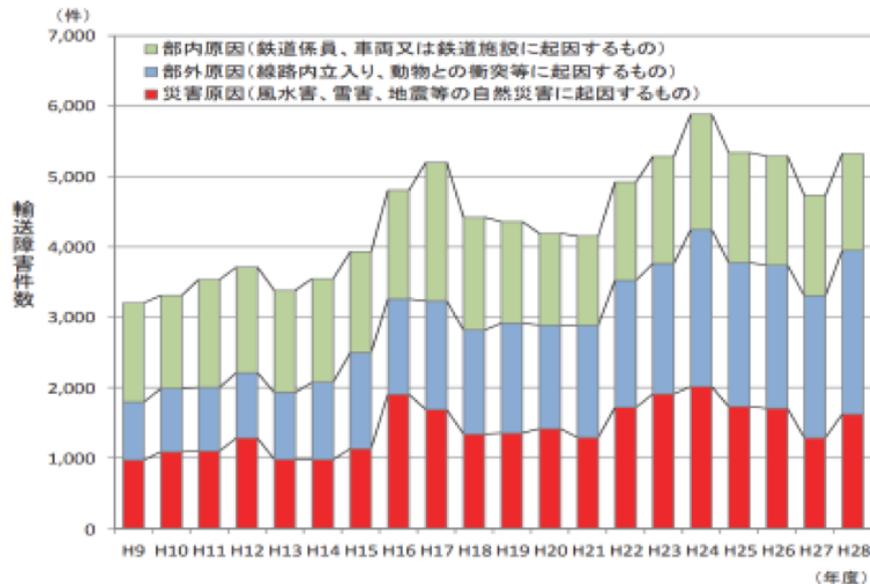


FIGURE 1.2: Annual traffic accident counts

On the other hand, the preference of different kinds and groups of people should also be considered in the research of evaluating performance of traffic system. For, example, people with different gender and age always have different habit of moving. Figure 1.3 shows the preference of using railway in people of different gender and age is apparently different. At the same time, there are also other conditions influencing the traffic preference such as occupation, finance condition and distance between home and workplace. And the traffic flow will also change in different time of one day. All in all, studying the traffic preference of different kinds of people and the performance of traffic system in different time is also important to evaluate the performance of railway transportation system.

Traditionally, statistics on how people use public transportation especially in larger cities are generated by 'transportation census of major cities' and Passenger Survey by individual transportation services providers. But for the development of Location Based Services(LBS), the traditional statistic data becomes not enough. Because this kind of data is with very low frequency. In Japan, the transportation census was only a 'one-day' survey in every five years, it is definitely not enough to monitor the performance of the public transportation comprehensively.

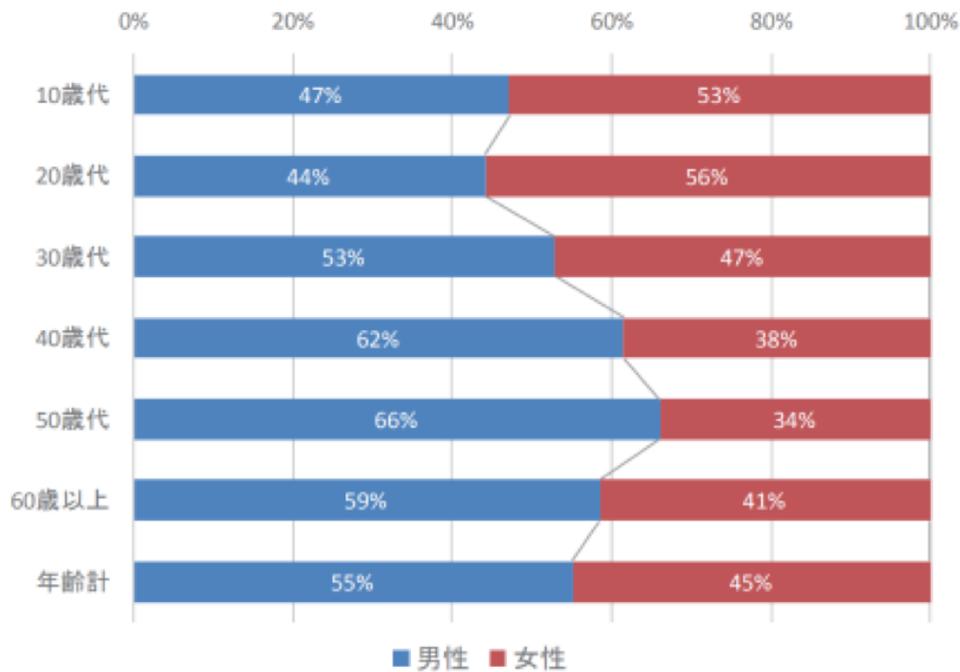


FIGURE 1.3: Share of railway usage in people with different gender and age

GPS data can solve this problem, as it can show the dynamic change of traffic flow [32] and even detect the anomalies such as bad weather and accidents [18]. Thus, in recent days, detecting human mobility and evaluating performance of transportation system via GPS big data become a very popular topic. In this study, we focus on both observing the human mobility and the performance of railway system in Japan via GPS big data. At the same time, there is a comparison of estimation result between the two different kinds of data as in this research, to make an all-direction discussion of monitoring passengers and their behaviour.

Under this background, the research has a specific purpose to monitor the human mobility and public traffic system in Tokyo. In this thesis, I will put some related works in the second chapter of past studies and research works. In the third chapter of framework and methodology, I will introduce the methodologies used in this research, and then present my results of experiments in the fourth chapter of data processing and experiment results. Finally, I will discuss and draw a conclusion in the final chapter.

Chapter 2

Past Studies and Research Works

2.1 Research on Public Traffic System

Studies of public traffic systems is a very popular issue in past several decades. Some of the research objects which focus on studying public traffic system use some statistic data like OD statistic data and data of traffic volume. Generally, those kinds of data are collected by methods as traffic census, questionnaires and even ticket counts in a period. Sanchez [28] paid attention to analyze the public transportation and its impact from the process of census data with GIS technology. And Fielding et al. [7] analyzed the reported statistic data for evaluating the performance of bus transit and identified seven performance indicators to assess the transit monitoring. As well as the statistic data and census data, many researchers found that analyzing the entrance-exit data of each station can improve the precise of the information of origin and destination. Lehtonen et al. [17] proposed the idea and stated the feasibility on the utilization of transportation smart card data in studies of transport planning. Chu and Chapleau [3] analyzed and estimated arrival time of buses based on spatial-temporal methodology using the record of smart card. As far as the development of data mining technology, more and more researchers prefer to study not only simply using the statistic record data. Morency et al. [20] indicated the performance of transit via card data mining methodology, and Kusakabe et al. [16] proposed a method to estimate passengers' behavior by using smart card data and proved the method is adaptable for estimating patterns of passenger usage.

Meanwhile, many researchers chose to study public transportation via data obtained from the sensors. Ishihara et al. [13] developed an algorithm to track the movement of pedestrians using

laser scanner. Filip [8][9] focused on estimating the GNSS sensors' performance on railway transportation. Those researches showed potential and possibility to remotely detect the residents' mobility utilizing different varieties of sensors.

2.2 Research on Traffic Monitoring and Evaluating Public Traffic System via GPS big data

As far as the development of location based system(LBS), nowadays, GPS data is extensively used in the researches of detecting and recognizing human behavior and monitoring public transportation system. More and more researchers started utilizing GPS big data to analyze and even predict traffic conditions.

Many studies focused on analyzing and estimating human mobility via GPS big data. Researchers always establish different models to recognize the patterns of human mobility. Sudo et al. [31] researched human behavior in disasters based on a methodology of real-time estimation. Jiang et al. [14] used deep learning methodology and proposed a modeling approach based on ROI to detect and predict human mobility with high efficiency. Song et al. [29] also utilized a deep learning model for simulating human mobility as well as transportation mode when a disaster happens in a citywide level.

Besides of researches on human mobility, traffic analysis and prediction is also a very important and popular research field. Some of studies focus on researching the performance of public traffic system using the taxi GPS data. Zhou et al. [39] proposed an online system to detect anomalous trajectories of taxi for real-time monitoring. Qian and Ukkusuri [24] developed a geographical regression model for monitoring spatial variation of taxi in New York City via GPS big data. Luo et al. [19] put their concentration on the emission of taxis and analyzed their spatial information in Shanghai. In many countries, taxi plays an very important role in traffic system and it is always easier to have its tracking data than railway and bus, many researchers prefer to do researches based on GPS trajectory data of taxi. But in Japan, as I mentioned in the first chapter, railway occupies the highest rate of residents' traffic modes and much higher than any other modes. It seems more valuable to study the performance of railway system in Japan. Wang et al. [33] detected the situations of train delays in Tokyo via records from Twitter. And Xia et al. [35] established a deep learning model to forecast the railway traffic system in a city-wide scale.

In this research, we will deal with a new kind of mobile big data and process it with a fresh start. There are two main purpose of this research, on the one hand, we can extract information from a new variety of big mobile data just like the life patterns of residents and estimation result of passengers. This work will provide some new aspects to understand human mobility because not only the data source but also the date of data collection is different. It can help us to update our cognition of those issues to a 'new version'. On the other hand, the pre-processing work will hugely contribute to the research works of following researchers to study on Target Ads GPS data, then make the work of data mining more smooth and efficient.

Chapter 3

Framework and Methodology

3.1 Framework

In this research, the objective is evaluating the performance of public transportation system, especially railway using GPS big data. To detect the performance of traffic system, at first, we need to filter those data which have a mere handful of points or too much noise and generate GPS trajectories by the GPS dataset. After that, it is necessary to collect training data for map matching and traffic mode detection. And then process the data to detect and classify the life patterns and cluster the residents with different varieties of life patterns. The work after life pattern cluster is mobility analysis, including pre-process such as collecting training data and map matching, transportation mode estimation, and estimation and analysis of railway passengers. Finally, a comparison between the result of estimation from two different GPS datasets will be proposed, and then make an assessment on the result.

For a specific user, the trajectory data can be defined as $traj = \{traj_1, traj_2, \dots, traj_n\}$. In this dataset, $traj_i$ represents each trajectories of this user. In this research, two kinds of GPS data will be processed, navigation GPS data and Target Ad GPS data. For the raw data of navigation GPS data, a trajectory can be indicated as $uid, tid, pid, longitude, latitude, timestamp$ in which uid means the id of a passenger, tid means the id code of a trajectory and pid refers to the number of points which compose to this trajectory. In the dataset of Target Ad data, the users' OS types are also indicated.

From the theory of graph, from one node to another, the shortest path should be calculated by shortest path algorithm. In our research, we calculate the path basing on Dijkstra Shortest Path

Algorithm[5]. On the other hand, for a GPS point P_i and a link L_i , the projection of P_i to L_i can be defined as a point p_i on L_i . The distance from P_i to p_i needs to be the shortest one within all of the points on L_i . The distance from P_i to L_i can be defined as the euclidean distance from P_i to p_i .

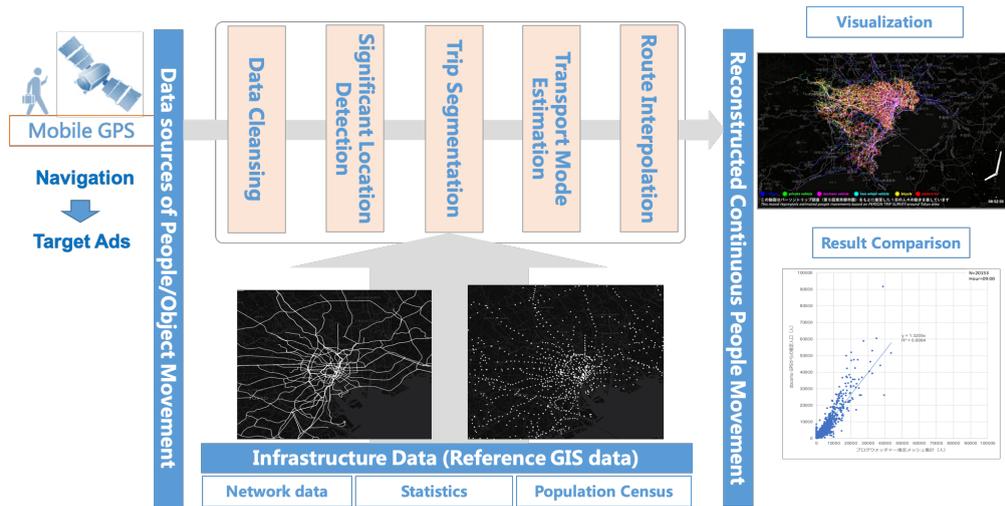


FIGURE 3.1: Framework: Estimate Human Mobility and Movement via big mobile GPS data

The framework of estimating human mobility is demonstrated in figure 3.1. including modules of data filtering, trip segmentation and mode detection, interpolation and people movement reconstruction. Examples of GPS trajectory visualization and real population estimation are also showed in this figure. Based on those modules, the technology of data filtering, life pattern classification, mobility analysis and passenger estimation will be introduced in following sections.

3.2 Life Pattern Classification

3.2.1 Data Filtering

For the preparation of life pattern classification and other following works, The work of data filtering is necessary. It is mandatory to prepare long-term data for life pattern discovery, otherwise the accuracy of the result will become contingency. The first step of data filtering is to make a basic statistic on GPS data to know the main interval of data amount of each users. Under the premise of reserving most of users, those data which have less points or valid dates should be deleted. From this work, the data for processing will become more precise and dense.

3.2.2 Life Pattern Clustering

For detecting human mobility, understanding the life pattern of residents is an essential work[?]. In this research, life pattern clustering is a very important work in pre-processing and data processing module. On the one hand, collecting training data from those different clusters can make it possible to detect traffic mode preference in different groups of people with disparate custom of activity. On the other hand, life pattern clustering itself is an effective methodology to analyze human mobility. For indicating people's activity at a specific period, the number of GPS points can be regarded as a significant index. For instance, the points of people who are commuting or travelling is always more than those who are inactive. The work of life pattern clustering can be divided into three steps.

Step 1 (Point Calculating): To count number of points of each user in every hour period(24hours).

Step 2 (Normalization): For each feature, using max-min norm methodology for normalization. Max-min normalization can be expressed as equation 3.1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3.1)$$

In this equation, X_{norm} is normalized data, X is original data. X_{max} and X_{min} are data with the maximum and minimum value.

Step 3 (Data Clustering): To detect and cluster people with different kinds of life patterns. In our research, we use X-means algorithm for data clustering.

3.2.3 X-means Clustering

X-means is an efficient algorithm for estimating number of clusters proposed by Pelleg and Moore [23]. It works behind every times run of K-means, to detect and select the better subset of current centroids should be splitted for suiting the data by calculating the Bayesian Information Criterion (BIC).

The first step of X-means algorithm is normally run K-means to convergence. The second step is to iterate two means in each cluster, to know whether and where those fresh centroids occur. Then make a decision of whether to do bisecting cluster according to the value of BIC score.

For the data D and different values of K , a set of alternative models M_j should combine with different cases of K . The score of those models can be defined by posterior probability $P_r[M_j|D]$. The formula proposed by Kass and Wasserman [15] can be used for approximating the probability until finish norm. The formula can be expressed as 3.2:

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \cdot \log R \quad (3.2)$$

For the j th model, $\hat{l}_j(D)$ is the log-likelihood of data. The amount of parameters of M_j is defined as p_j .

Based on the identical spherical Gaussian assumption, the maximum likelihood estimation can be expressed as equation 3.3:

$$\hat{\sigma}^2 = \frac{1}{R - K} \sum_i (x_i - \mu_{(i)})^2 \quad (3.3)$$

For centroid n , we can only focus on its attached set D_n then plug in maximum likelihood estimation yield as equation 3.4:

$$\hat{l}(D_n) = -\frac{R_n}{2} \log(2\pi) - \frac{R_n \cdot M}{2} \log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n \log R_n - R_n \log R \quad (3.4)$$

Iterating until $n_i K$, then finally choose the best model by X-means methodology.

3.3 Mobility Analysis

3.3.1 Map Matching

3.3.1.1 Introduction of Map Matching Methodologies

For mobility analysis, map matching is an essential part for detecting the information from users' real life. At the same time, for studying the performance of public transportation system, map matching is also important to know the information such as name of station or railway line. It is a key issue to match GPS points or trajectories to road networks.

There are four different kinds of traditional map matching methods based on a review from Qudus et al. [26]: topological, geometric, probability method and other advanced methodologies. Map matching based on topological is a methodology taking topological features for considering and this methodology performs well in detecting spatial patterns of region containment and road network[37]. Map matching based on Geometric uses geometry data and information for matching, but not use topological methodology[2]. probability map matching is a methodology which consider the errors of GPS point and those points should be matched to road networks with the maximum probability[22]. And other methodologies prefer to use some advanced theories and approaches like Particle Filter[10], Fuzzy Theories[25] and Kalman Filter[36]. Those map matching methodologies are developed and proposed to improve the precise and accuracy positioning result.

As Target Ads GPS data is mainly used in this research, the GPS data is rather heterogeneous than navigation GPS data. For instance, the time interval is very large, mainly about 5 minutes to 30 minutes. Thus, traditional methodologies such topological based methods and geometric based methods is no longer adaptable. In Tokyo, the railway network is very complex. It is a common situation that the distance from a GPS point to another railway line is nearer than the correct one as the precision of data is not very high and the time interval is quite large. For those reasons, we need to choose a map matching methodology which is more advanced and have higher accuracy. In our research, we use an advanced map matching methodology based on Hidden Markov Model(HMM) which can decrease the influence of the heterogeneous GPS dataset.

3.3.1.2 HMM Based Map Matching

Hidden Markov Model (HMM) is a statistical model to describe a Markov Process with unobservable or hidden states and extend the Markov Chain developed by Baum and Petrie [1]. The Markov Chain is a model which shows the probabilities of a sequence of random events. Figure 3.2 shows a simple example of Hidden Markov Model. It is widely used in fields such as Signal Processing, Pattern Recognition and Fault Diagnosis.

In Hidden Markov Model, there are two basic assumptions. Firstly, the observed data should be assumed to be dependent on the unobserved state at any stage with an output probability distribution. Secondly, the unobserved state at a specific stage should be determined by the former unobserved state by a transition probability. Thus, if we can know the initial probability

and transition probability for any unobserved state, the probability distribution of all of the observations can be calculated.

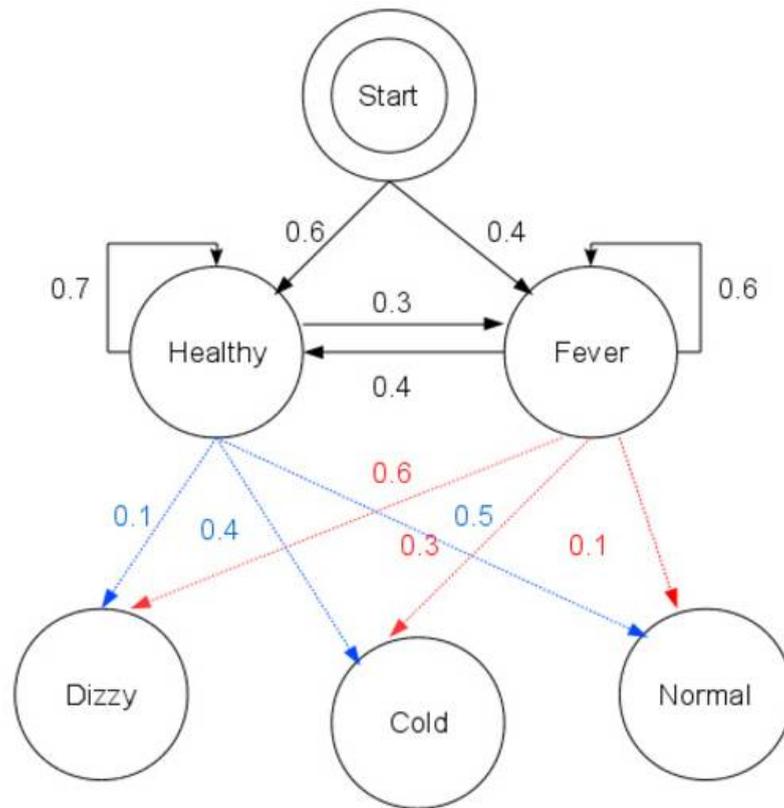


FIGURE 3.2: An example of Hidden Markov Model

Nowadays, Hidden Markov Model has been regarded as an efficient methodology for map matching with heterogeneous GPS data in order to improve the accuracy[21]. For the HMM map matching work, the objective that we want to output visibly is the trajectory points. In this model, the unobserved state is which road or railway that the point belongs to. By the Hidden Markov Model, the output probability can be calculated by the distance from the point to its projection in road or railway and the transition probability can be calculated by the distance of two neighbouring points. Based on the unobserved state sequence with maximum values of output probabilities and transition probabilities, the most likely route can be calculated. An simple example of HMM map matching is showed as figure 3.3.

The map matching work can be illustrated in following steps:

1. Search the railway network to discover the potential routes in a defined searching radius for each GPS point.
2. Calculate the probability of each line.

3. Calculate the transition probability of the unobserved stated between every two neighbouring GPS points, then output the total probability and find out the best path.

The output probability can be calculated by a Gaussian distribution method in equation 3.5:

$$p(p_i|L_j) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-0.5(\frac{d_{piL_j}}{\sigma})^2} & \text{when } 0 < d_{piL_j} \leq r \\ 0 & \text{when } d_{piL_j} > r \end{cases} \quad (3.5)$$

In this equation, p_i is the GPS trajectory point and L_i is the matched route. d_{piL_j} is the distance between the route and point, r is the searching radius and σ is the standard deviation in GPS measurement.

At the same time, the transition probability from two unobserved states can be calculated by following methodology:

- 1 The probability will be 1 between two states with the same route.
- 2 For the different routes, the probability can also be calculated as proportion to the times of transfer and distance.

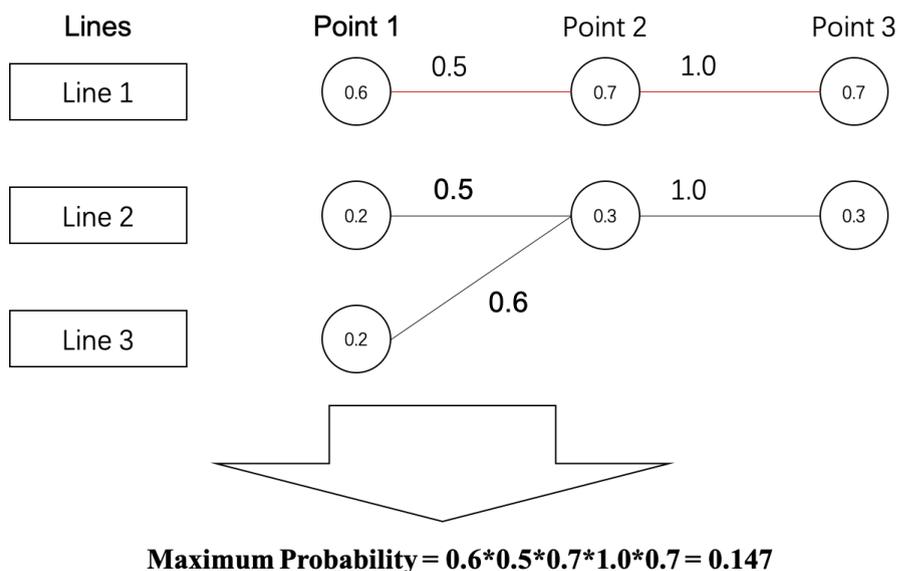


FIGURE 3.3: An simple example of HMM map matching: output a best line with highest probability

For the definition of output probability and transition probability, the following rules proposed by Newson and Krumm [21] should be considered:

Rule 1: The output probability and the distance between the routes should be inversely proportioned.

Rule 2: The transition Probability between two unobserved states and the accessibility of two routes should be proportioned.

Based on above-mentioned methodologies and rules, those two kinds of probability can be calculated and then make it possible to use Hidden Markov Model for map matching works.

3.3.2 Transportation Mode Estimation

In this research, an another core data processing work is transportation mode estimation. In order to detect human mobility and evaluate the performance of traffic system, it is essential and necessary to know the traffic mode of each GPS trajectories and then reconstruct the trip trajectories of users. There are several different researches for traffic mode estimation or trip reconstruction, based on tools or methodologies such as GIS information[30], GPS-based travel surveys[4] and GPS data with web application[38]. Those researches of traffic mode detection mainly concentrated on understanding human move pattern and then predict or estimate the movement of human beings. All in all, traffic mode detection plays an important role in the field of traffic planning.

In order to finish this work, there are several procedures need to be accomplished before traffic mode estimation. At first, we need to distinguish the status of a GPS trajectory in each period with the condition of stay and move. In one trajectory, it may have different status as some points shows the user is moving as well as other points keep staying in one or several areas. After the stay and move detection work, we can pick up those points with a status of moving and divide the points with walking and not-walking. For this detection work, we need to calculate the speed and find out those points which are taking some traffic tools. After this, the traffic mode detection work begins. The procedure before traffic mode detection can be showed as a figure [3.4](#).

3.3.2.1 Stay Point Recognition

For the first step of distinguishing stay points and move points, we utilize an algorithm based on spatio-temporal values of each point. The detecting is based on a rule that the distance of

several neighbouring points is less than a default value which is considered to be a reasonable range of staying. And the time interval from the first point to the last one is longer than another default value which users spend in a stationary place.

This rule can be expressed in following inequalities 3.6:

$$\begin{cases} D(p_{start}, p_{end}) < D_{max} \\ T(p_{start}, p_{end}) > T_{min} \end{cases} \quad (3.6)$$

Where $D(p_{start}, p_{end})$ and $T(p_{start}, p_{end})$ are the distance and time interval between the start point and the end point, D_{max} is the maximum default value of staying distance as well as T_{min} is the minimum default value of staying time. S

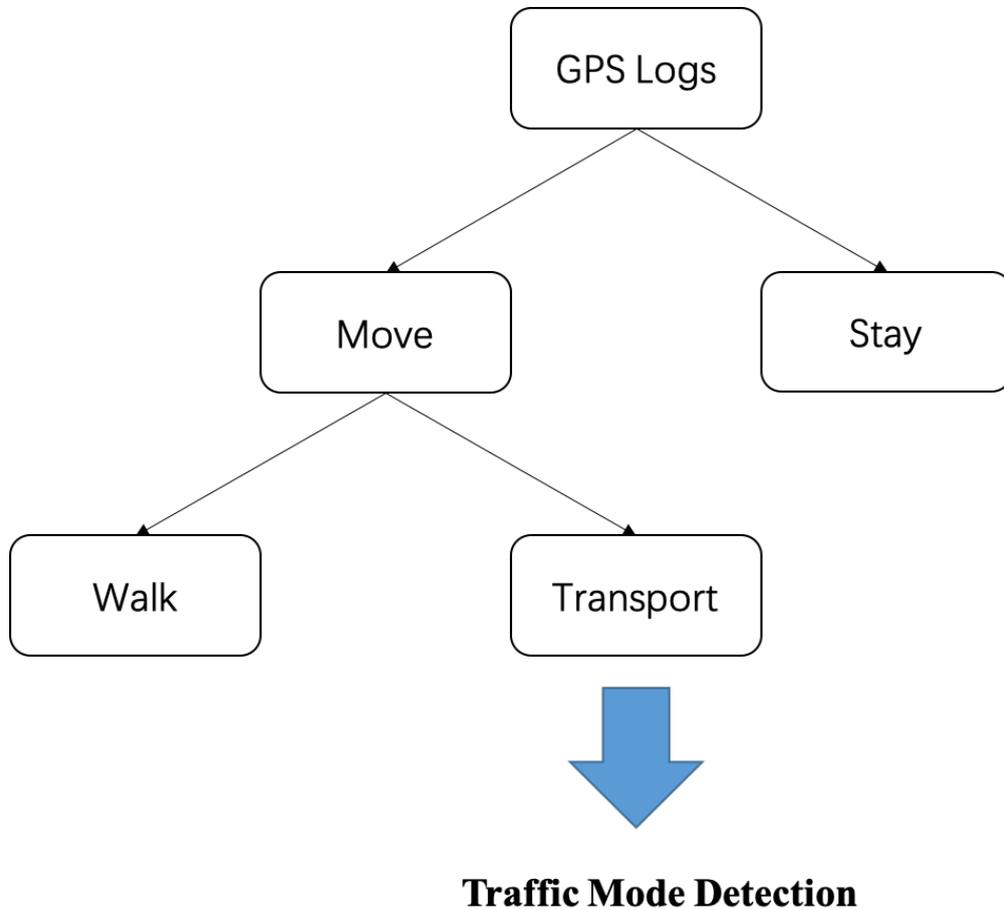


FIGURE 3.4: An flowchart of procedures before traffic mode detection

The set of detected stay points for one user can be defined as p_1, p_2, \dots, p_n , which includes information of latitude, longitude, start-time and end-time. The centroid of the latitude and longitude of all those stay points can be regarded as the coordinate of the stay point of the user

in this stay period. Moreover, it is also necessary to detect and delete those noise points. There are three types of noise points:

1. A latter point which is recognized as the first point of stay.
2. A point which is the start point but recognized as a latter point.
3. A point with a much farther distance with neighbouring points than others

In order to recognize those noise points, we calculate the mean value and standard deviation of the dataset. Then move those points with abnormally high value of standard deviation.

3.3.2.2 Transport Point Recognition

To detect the transition mode of one user, there are several tasks. At first, it is necessary to detect trip segments from each trajectories. The segment means that the user uses the same transportation mode in one continuous period. On the other hand, the change point, which means the user changes his transportation status or mode should also be detected. In addition, for those segments of 'non-walk', we are going to classify the transportation mode of those segments.

In our research, we calculate the speed so as to define the traffic mode of each segment. The acceleration is another important index to detect the traffic mode. But because the data is quite heterogeneous, acceleration might be hard to calculate because the time interval is too large (sometimes over tens of minutes). Thus, we use velocity change rate (VCR) instead of acceleration for another index to recognize the traffic mode of a segment. It can be calculated as following equation 3.7:

$$VCR = \frac{|v_{average} - v_{current}|}{v_{average}} \quad (3.7)$$

In this equation, $v_{average}$ means the average speed of the segment and $v_{current}$ means the speed of current point. If the value of VCR is over or less than a threshold, the point will be defined as a change point.

3.3.2.3 Transportation Mode Classification

After those above-mentioned procedures, we finally achieved the step of transportation mode classification. In our research, we utilize a supervised learning methodology to classify transportation mode of each segment. Thus, we need to collect training data. The tool of data collection is developed by Dr. Witayangkurn[34], an assistant professor in Center of Spatial Information Science, the University of Tokyo. It is a web-based activity analysis and labeling tool with Google Map, and can make it possible to use ground truth data to support us to classify. We can label the transportation modes of each segments and make them to be our training data, the screen of this tool is shown as figure 3.5:

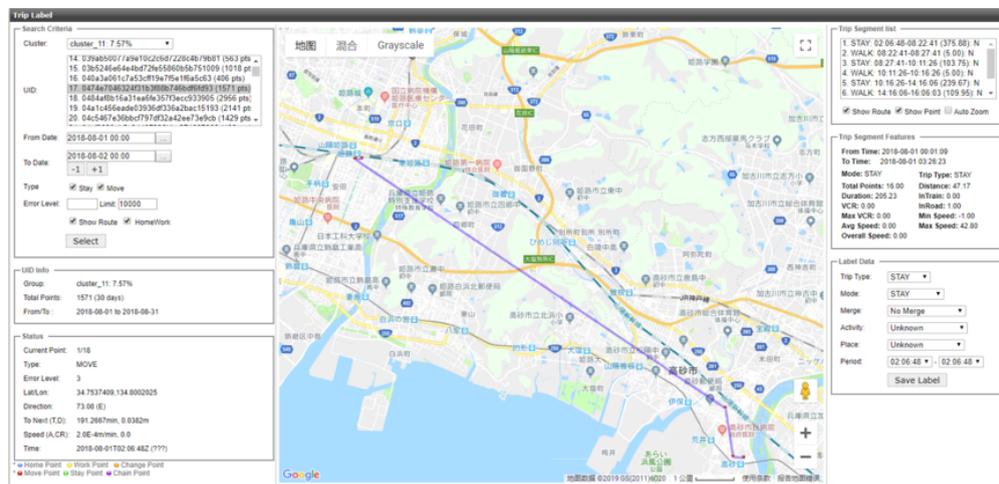


FIGURE 3.5: Screen of the web-based labeling tool

After collecting the training data, we can start the classification work. In this stage, we choose to use a Random Forest based methodology. Based on some former researches, it is showed that Decision Tree is a efficient and accurate model for mode detection[27]. But Stenneth et al. [30] urged that Random Forest model has even a better performance than Decision Tree. Random forest is formed by a variety of Decision Trees and export the output from those Decision Trees[6]. We input some features extracted from those segments to the model and estimate the transportation mode of each segment. The features includes:

Time Duration(min): The time duration is calculated by the difference between the start point and the last point.

Total Distance(m): The distance is calculated by summing all of the distances of each neighbouring points in a segment.

Velocity: The speed features including maximum and minimum speed, average speed, overall average speed and velocity change rate(VCR). As we have introduced VCR and we don't need to calculate the maximum and minimum speed, we only introduce those two types of average speed. The average speed is the average speed of each points in a segment as well as the overall average speed all over the segment.

Occupation percentage of points in road and train network: This feature will contribute to the classification of traffic modes such as train, bus and car. We use a methodology of spatial query to detect whether the point is in the network and combine with GIS buffers to finally define the percentage that we need to calculate.

After the classification work finished, we can achieve our aim to detect and estimate the traffic mode of users and continue our research work of analysis of passengers.

3.3.3 Analysis of Passengers

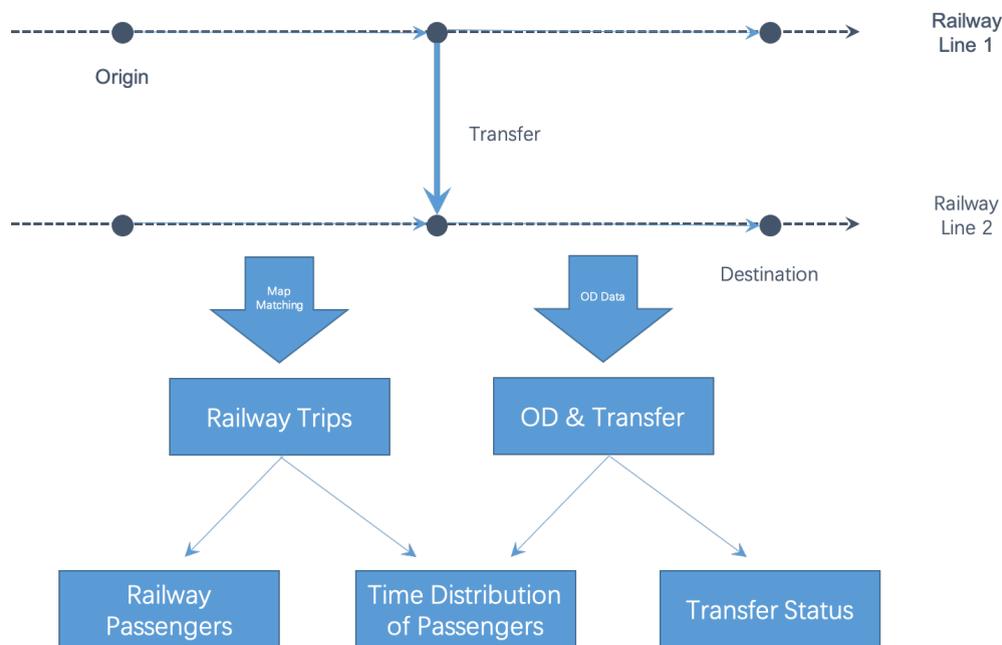


FIGURE 3.6: Methodology of Extracting Information of Railway Passengers

In order to estimate the movement of railway passengers after all of the preliminary work finished, we have two steps to complete[12]. On the one hand, map matching is applied to recognize the trajectories of rail transit, which has been introduced in former sections. On the other hand, it is necessary to aggregate the data to recognize the passengers of each station and the

status of transfer (get-on, get-off and exchange). For each trajectory, we can extract the information of its origin and destination. Based on the OD information, it is possible to count the passengers of each station in a time period. Then combining the OD information with the result of map matching, the status of transfer as well as the time distribution of each railway user can also be identified. The methodology can be described in figure 3.6.

Based on this methodology, we can reconstruct the railway trajectories and analyze the movement of passengers.

Chapter 4

Data Process and Experiment

4.1 Introduction of data sources

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.

GPS data is anonymised for the analysis. In addition, all data processing was conducted in the premise of an Ad Tech company who has the original data by using the software I developed.

4.1.1 Navigation GPS Data

Navigation GPS data is collected by a private company and a mobile operator. The total data size is over 1.5TB with over 30 billions GPS records. The raw GPS data had been fully processed to generate trajectories and the traffic mode has been detected by the work of Witayangkurn et al. [34]. The data dictionary is shown as table 4.1:

4.1.2 Collected GIS Data

In this research, though big mobile GPS data is mainly utilized, many sources of GIS data are collected to support this research. On the one hand, some data such as railway network and

TABLE 4.1: Dictionary of Navigation GPS Trajectory Data

Name	Specification
user_id	The code to identify different users.
trajectory_id	The trajectory id of different users.
date	The date of the trajectory.
traffic_mode	The traffic mode of each segments.
trajectory	The GPS points in a trajectory.

railway stations is necessary in this research. It functions in many parts such as map matching, traffic mode detection and visualization. Those data are indispensable part to support this research. On the other hand, some of the open GIS data such as Traffic flow census data is also utilized, this kind of data can be visualized by maps to intuitively show some status such as traffic mode preference of residents.

The railway network data is collected and adapted by Kanasugi et.al[11], from National Land Numerical Information(NLNI). The railway network data had been simplified and some topological information had also been added to this data. The basic information of railway information data is shown as table 4.2.

The railway stations data is also collected in this research. The data can show the information of each stations, the dictionary is also shown as table 4.3.

The traffic flow open GIS data are collected from NLNI, the same as railway network data and railway stations data. This data is produced by traffic survey, and collected from three different main Metropolitan area: Tokyo Metropolitan Area (including the area around Tokyo), Chukyo Metropolitan Area (including the area around Nagoya) and Kinki Metropolitan Area (including the area around Osaka and Kyoto). In this research, as we chiefly discuss the transit performance in Tokyo in the part of evaluating the traffic performance, so we just choose the data of Tokyo Metropolitan Area for visualization, and adapted the data that deleted some unused information. Thus, the dictionary of traffic flow data is shown as table 4.4:

4.1.3 Target Ads GPS Data

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)

TABLE 4.2: Dictionary of Railway Lines Data

Name	Specification
linkid	The unique code to identify the railway network table .
comp_code	The company code of the railway lines.
comp_name	The name of railway companies.
line_code	The code of different railway lines.
line_name	The name of different railway lines.
source_station_code	The origin station of each link.
target_station_code	The target station of each link.
length	The length of each link.
geom	The geometry information of railway lines .

TABLE 4.3: Dictionary of Railway Stations Data

Name	Specification
station_code	The unique code to identify the railway stations.
comp_code	The company code of railway stations.
line_code	The code of the railway line.
station_name	The name of the railway station.
station_group	The column created for storing topology information.
geom	The geometry information of railway stations.

TABLE 4.4: Dictionary of Traffic Flow GIS Data

Name	Specification
Metropolitan_id	The code of Metropolitan Area.
survey_year	The Year of survey.
zone_code	The code of zones in the Metropolitan Area.
railway_trip	The amount of railway trip happened(origin and Destination).
bus_trip	The amount of bus trip happened(origin and Destination).
car_trip	The amount of car trip happened(origin and Destination).
bike_trip	The amount of bike trip happened(origin and Destination).

TABLE 4.5: Dictionary of Target Ads GPS Data

Name	Specification
user_id	The code to identify different users
timestamp	The time information of the GPS point.
longitude	The longitude of the GPS point.
latitude	The latitude of the GPS point.
accuracy	The accuracy of GPS point.
OS_type	The OS type of user's smartphone.

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.

Comparing with Navigation GPS Data, Target Ads GPS Data is a new data with much bigger users, which reaches over 20 millions. However, the quality is quite limited and the time interval is averagely bigger than Navigation GPS Data. In our research, as the Target Ads GPS Data was a completely raw data, it is necessary to process the data from the first step. As the total size of Target Ads GPS Data is extremely large, we chose one-month-data of Aug.2018 for processing and experiment. The size of chosen data is approximately 1TB in csv files. The attributes of the data is shown as Table 4.5. In Target Ads GPS data, the OS type of users is also collected, which is not included in the data attributes of Navigation GPS Data.

An example of visualization of Target Ads GPS Data is shown in figure 4.1. It is clearly displayed that the dynamic population flow change in different point-in-time of one day.

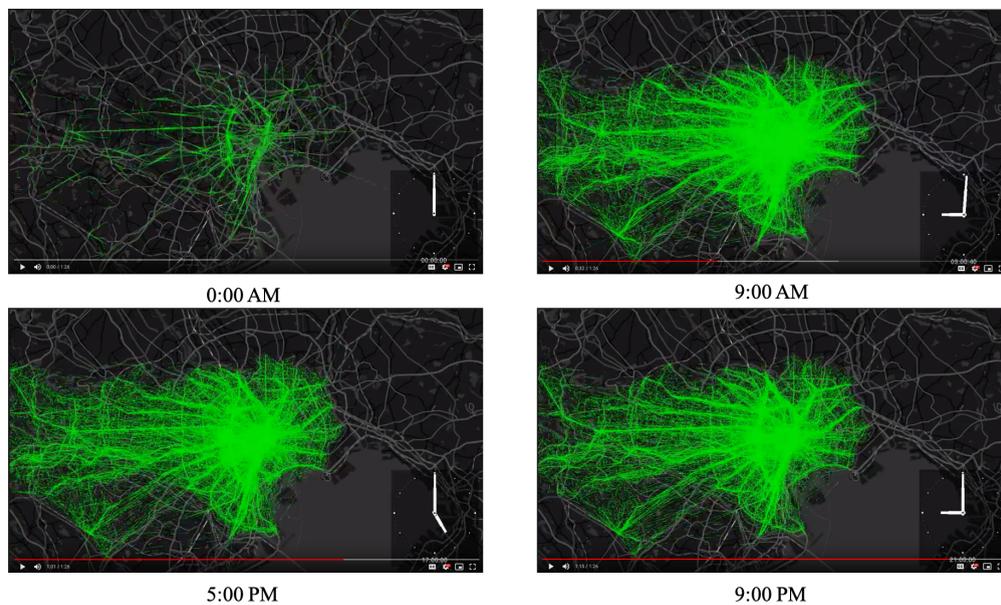


FIGURE 4.1: An Example of Visualization of Target Ads GPS Data in Different Time of One Day

4.2 Data Filtering

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

TABLE 4.6: Basic Statistics of Raw Data

Attribute	Amounts
Total IDs	Over 5 million IDs
Total Records	7.7 billions
Average Records Per ID	45/Day
Maximum Points	1.7 millions
Minmum Points	1
Data Period	2018/08/01-2018/08/31

points, Average records of each Ids per day, maximum and minimum points of each IDs. The result of basic statistics is shown as table 4.6.

For the data amount distribution of each days, we also made a statistic. The result is shown in figure 4.2. The distribution of total days of each users is polarized to both ends of x-axis, this result shows that many IDs have large amount of data covering most of days in the month, but there are also considerable amounts of IDs which just have data with very limited period. And the distribution of total points shows that IDs with less points occupies a high rate in all of the IDs. In this research, the life pattern analysis requires IDs with long-term data and relatively high quality, otherwise the results will become contingency. Thus, the IDs with below ten days period or below 100 total points are removed in this research.

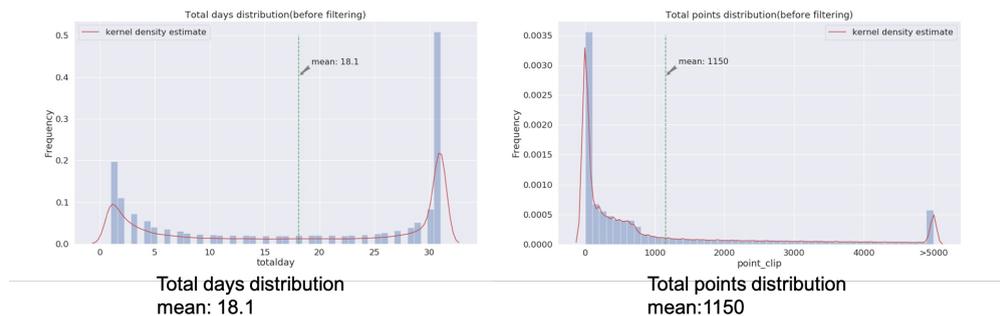


FIGURE 4.2: Total Days and Points Distribution of Target Ads GPS Raw Data

After the data filtering, the total days distribution and total points distribution is shown in figure 4.3. And the comparison of the data amounts before and after filtering is shown in table 4.7. The variation of mean point of each ID per each hour in one day is shown in figure 4.4. It clearly displays that the amounts of total IDs become much less meanwhile the total records does not have large change. At the same time the average records per ID obviously increases after filtering work. In other words, the data become more dense and homogeneous. It is foreseeable that the quality of following data analysis will be much better thanks to the data filtering work.

TABLE 4.7: Data Amounts Before and After Filtering

Attribute	Before Filtering	After Filtering
Total IDs	Over 5 million IDs	About 4 million IDs
Total Records	7.7 billions	7.5 billions
Average Records Per ID	45/Day	67/Day (+49%)

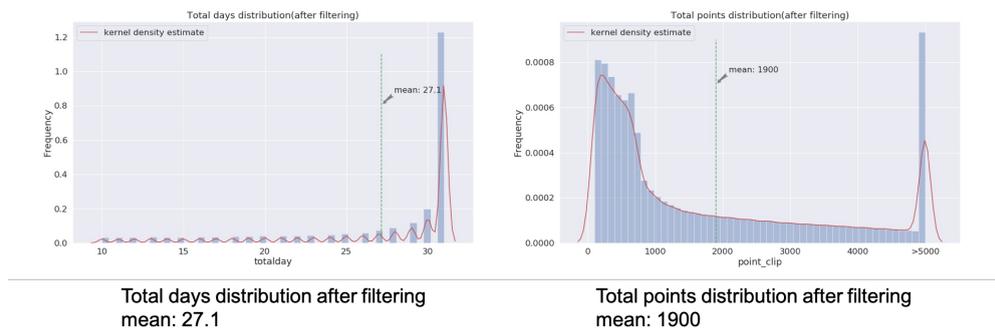


FIGURE 4.3: Hourly Statistics of Mean Points

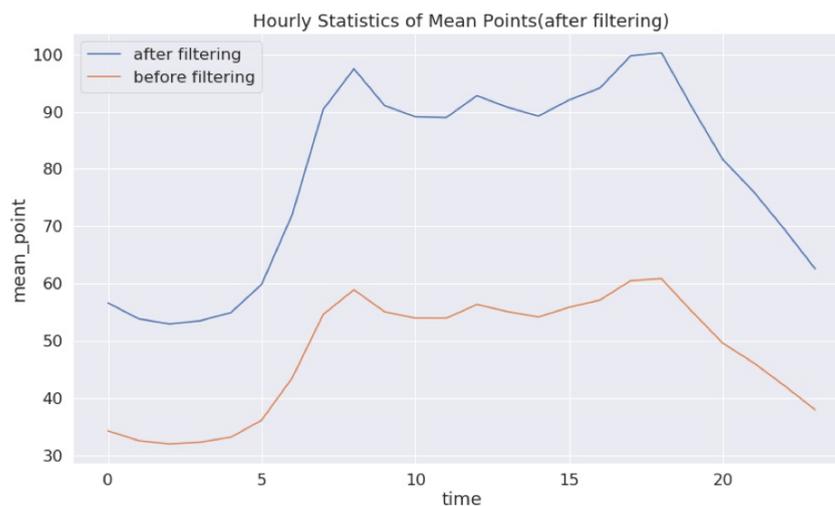


FIGURE 4.4: Total Days and Points Distribution After Filtering

4.3 Mobility Analysis

4.3.1 Life Pattern Observation

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. We will cluster and classify several kinds of main sorts of life mode of users to help us to recognize and understand the behavior

of users. On the other hand, it is also an important part in pre-processing work of Target Ads GPS data processing and analysis. For instance, if we detect the preferred transportation mode of different groups of users with various types of daily life, it will benefit not only researchers to have a more comprehensive analysis but also the company to know the different requirement of each user, and then provide more targeted information to users and improve the quality of their products.

To classify the life pattern, at first we need to know how many groups should we divide. In order to define optimum number of clusters, we introduced Bayesian Information Criterion (BIC) for evaluation and identification. The output of life pattern observation includes the result of BIC, cluster result of active periods and clusters of life patterns.

Figure 4.5 shows the result of BIC evaluation. It is shown that when the amounts of clusters is 15, the value of BIC reaches the highest record, which indicates this amount of cluster is the most appropriate. Thus, we cluster 15 groups of life patterns in this research based on the result of BIC evaluation.

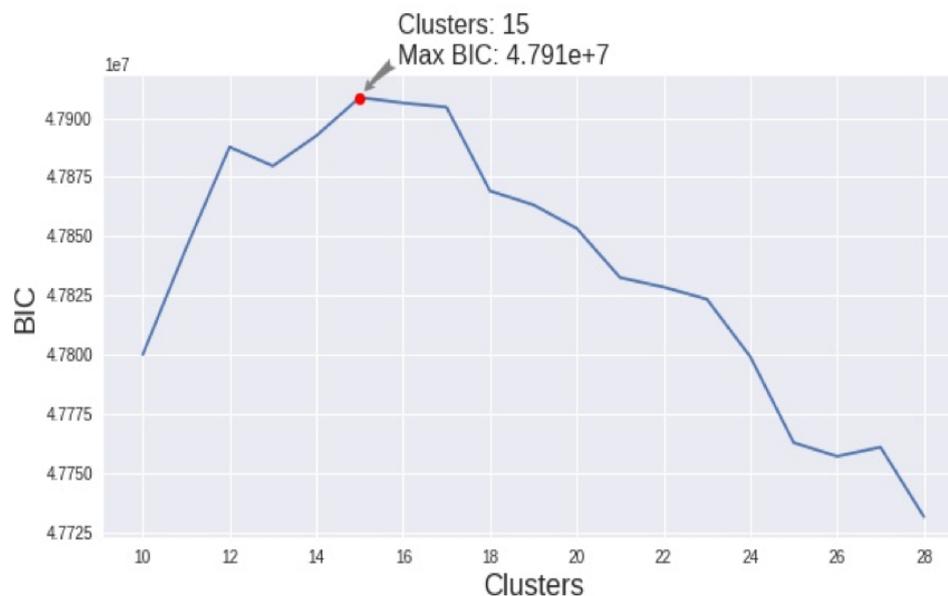


FIGURE 4.5: BIC variation of each amount of clusters

After defining the most adaptable amount of clusters, we use X-means methodology to estimate. The result is shown in figure 4.6 and figure 4.7. In figure 4.6, the mobility modes of 15 groups of residents are detected and visualized. Meanwhile the user amount of each patterns has been also estimated in figure 4.7. Those results are clearly shown that the most preferred life pattern of residents is the pattern of cluster one, about 13.3%, much more larger than other clusters.

This is a very common lifestyle that the person become active before 10am and have a rest after 8pm. There are some other similar life patterns such as cluster seven, cluster ten and cluster six which indicates this kind of lifestyle with a little difference in the time of activity or rest and also occupies a high rate in all of the life patterns. Those clusters mainly replace those jobholders or students with mainstream timetable.

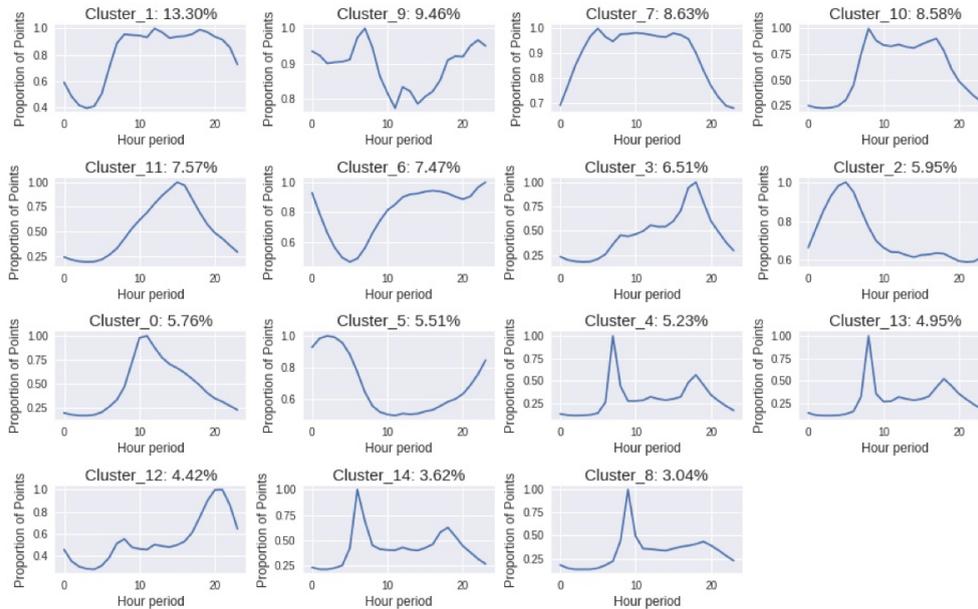


FIGURE 4.6: Groups of Activity Patterns



FIGURE 4.7: User Number of Each Clusters During Active Periods

On the other hand, there are also some other life patterns which is also popular. Cluster nine indicates a completely reverse life style, rest in the daytime but active at night. This kind of

clusters always indicate those people who stay up late at night. And cluster eleven replaces those residents who become active in the afternoon, this result may mostly reflex the lifestyle of elders, based on our common sense. There are some other clusters such as cluster four, cluster eight, cluster thirteen, cluster fourteen just have one or two peaks in the commuting time, one of the possible cause of those clusters might be the limited and heterogeneous quality of the data. This result might be an important issue to be considered and solved in the future.

4.3.2 Preliminary of Railway Passenger Estimation

4.3.2.1 Map Matching

As the function of life pattern detecting is not only providing a result to help us understanding the residents behaviour, 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. In this research, the output of map matching model includes:

- 1 The route of railway network.
- 2 The origin, destination and transfer information.
- 3 The new GPS coordinate matched to the railway network.

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.

4.3.2.2 Data Interpolation

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.

We chose 5 minutes as the fixed time interval. After data interpolation, the dataset become more dense and precise.

4.3.3 Analysis of Rail Transit Passengers

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.

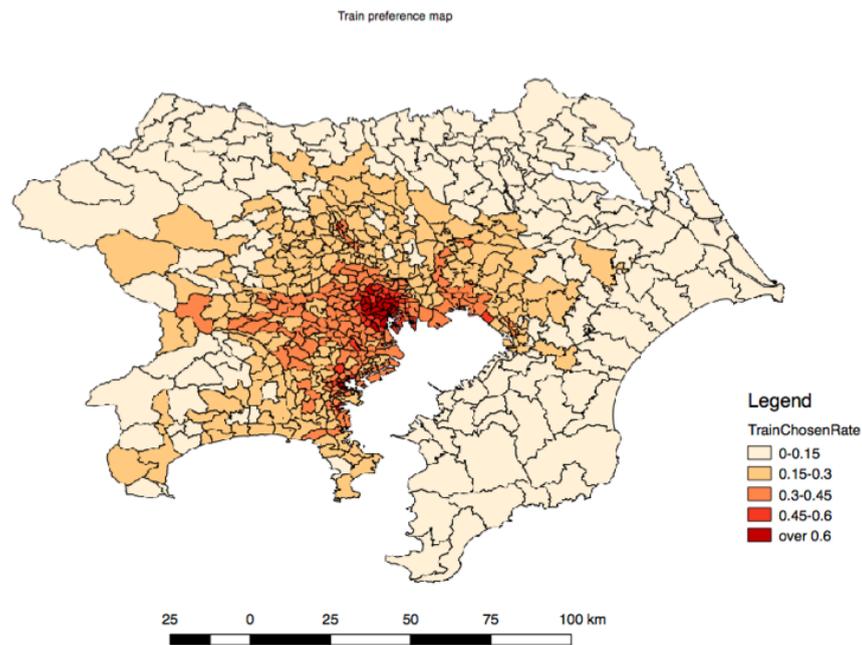


FIGURE 4.8: The Rate of Rail Transit Usage in Different Districts of Tokyo Metropolitan Area

4.3.3.1 Preference Analysis via GIS Open Data

For the analysis by GIS methodology, the data is collected from National Land Numerical Information. Figure 4.8 shows the various rate of train usage in different areas of Tokyo metropolitan area. It is obvious that people who live in the center part of Tokyo have higher preference to choose railway for their trip. In the meantime, most of those districts which are located in

outer regions of Tokyo metropolitan area have low preference of railway usage. The different development of railway lines may lead to this discrepancy.

On the contrary, figure 4.9 shows the usage rate of other transit mode including bus, car, bike and walk. The result of the usage of car is completely opposite from the result of train, and those maps can intuitively express the preference of each traffic mode in different districts. Figure 4.10 shows that people who live in the center region of Tokyo metropolitan area prefer to trip by train as well as those residents who live in the outer places prefer to trip by car. There are also a small amount of districts where people prefer to go out by bike or on foot, but with no district whose residents prefer to trip by bus than other traffic modes. In this research, we mainly focus on detecting human mobility and rail transit usage and performance via big mobile data, so we won't deeply explore the reason of those phenomenons.

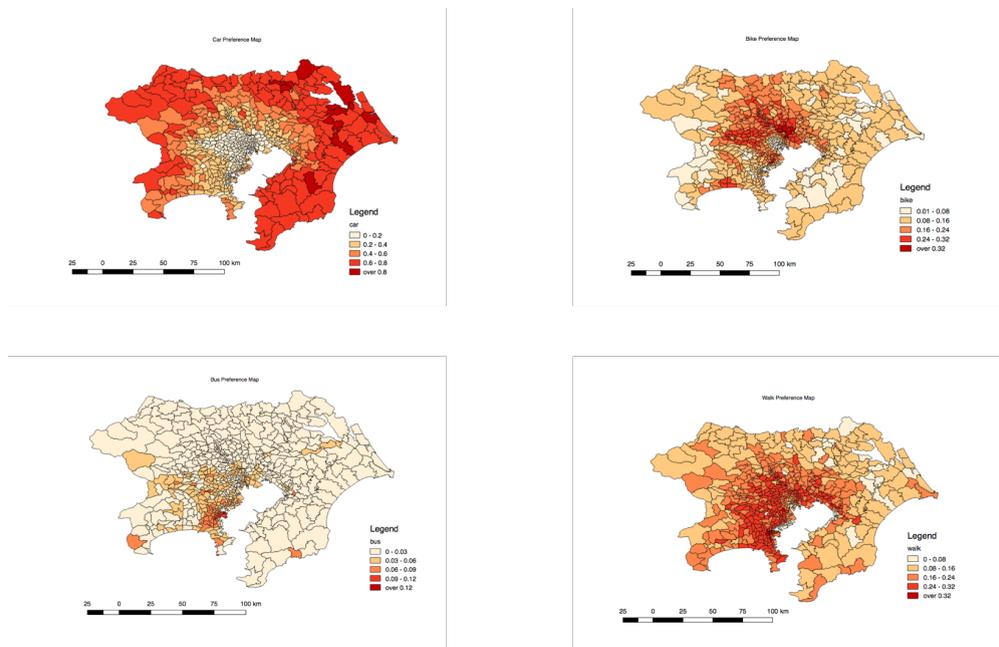


FIGURE 4.9: The Rate of Other Transit Mode Usage in Different Districts of Tokyo Metropolitan Area

4.3.4 Railway Passenger Estimation via Mobile Big Data

As it is shown in last section, 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.

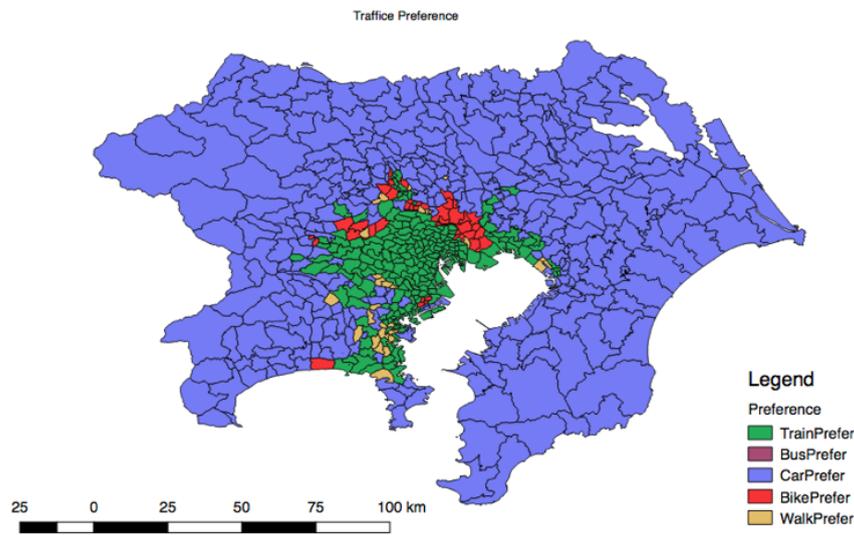


FIGURE 4.10: The Favourite Transit Mode in Different Districts of Tokyo Metropolitan Area

At first, a basic statistic was made to show the passenger flow of each time distribution in those station which is shown in figure 4.11.

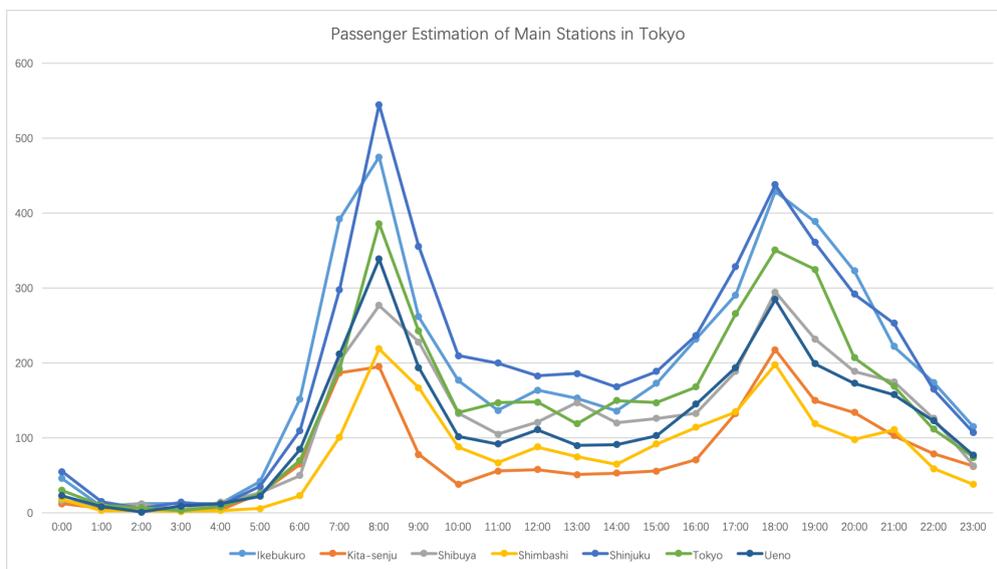


FIGURE 4.11: The Hourly Estimation of Passenger Flow by Stations

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 figure 4.12.

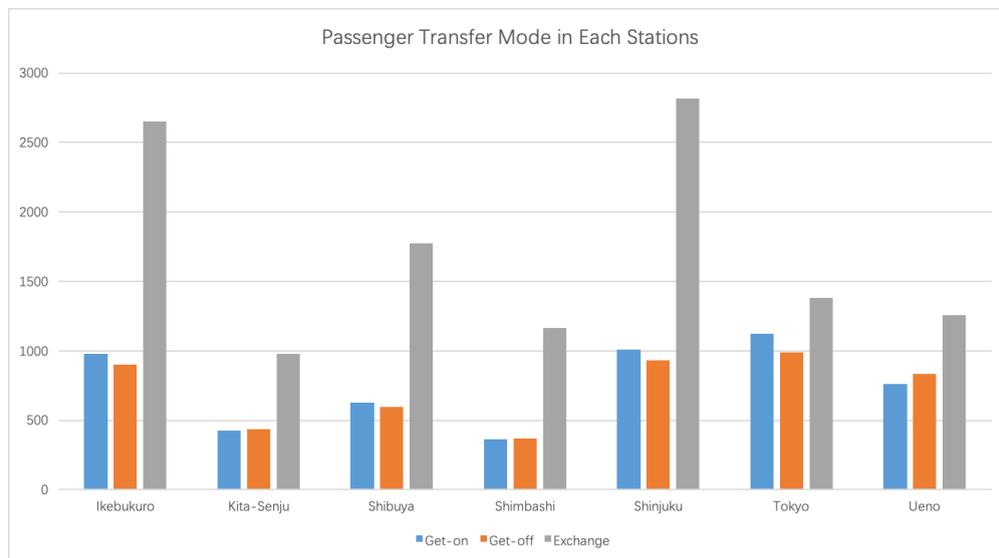


FIGURE 4.12: The Transfer Status of passengers in each stations

This result indicates the different transfer status distribution of each stations, as all of those stations are main transportation junctions in Tokyo, Shinjuku station and Ikebukuro station have the largest passenger volume. But on the other hand, most of the passengers use these two stations as an interchange station. The amount of passengers who get-on or get-off in these two stations is not such larger than Tokyo station and Ueno station, whose interchange passenger volume is much smaller than those station. Kita-Senju station, Shibuya Station and Shimbashi station are in similar situation. Passengers mostly utilize those stations to exchange rather than regard them as origin and destination. By contrast, the rate of exchange in Tokyo station and Ueno station is much lower, which indicates that passengers always get-on and get-off in those stations rather than exchange to other places.

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. Figure 4.13 and table 4.8 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,

TABLE 4.8: Table of Passenger Share in Commuting Time

Station	Occupation of passengers in Commuting time
Ikebukuro	68.0%
Kita-Senju	74.6%
Shibuya	62.3%
Shimbashi	59.8%
Shinjuku	63.3%
Tokyo	63.7%
ueno	67.1%

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. That might because Kita-Senju is a huge interchange station but located remoter than other stations from the downtown of Tokyo, so that the utilization of common time is not as high as those stations located in downtown of Tokyo.

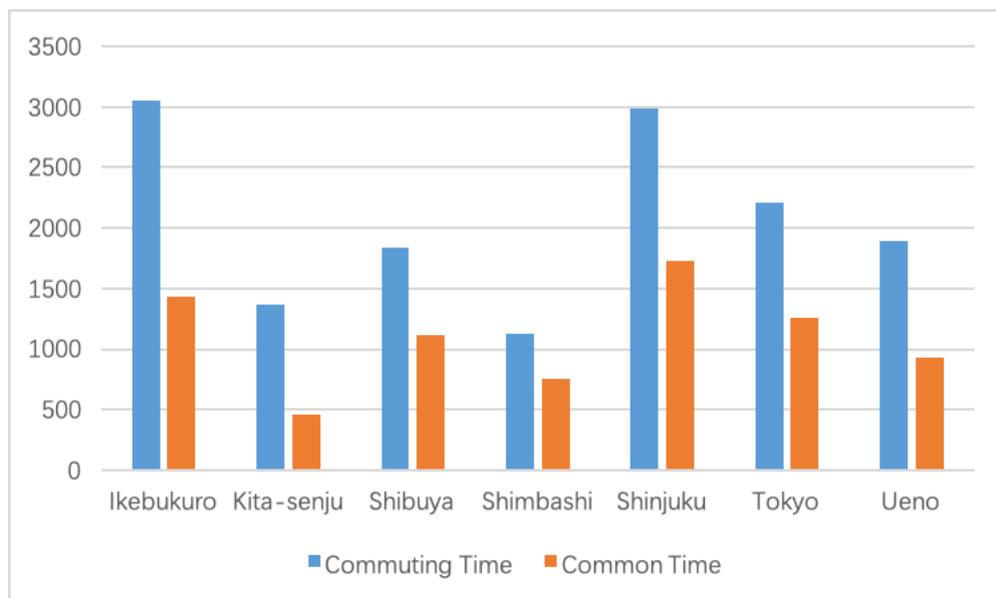


FIGURE 4.13: The Passengers of Commuting Time and Common Time in Each Stations

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.

TABLE 4.9: A Comparison with Real Census Data

	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

4.3.5 Comparison and Accuracy Assessment of Passenger Estimation

4.3.5.1 Comparison with Census Data

In this research, the utilized census data is collected from NLNI with the format of shapefile. The passenger volume in this data is counted by each station of a variety of companies.

First of all, in order to assess the accuracy of estimation, a least squares methodology can be imported to make a regression analysis between GPS estimation result and census data. In this research, we listed the result of the seven stations we selected, and make a comparison between the predicted real passenger volume via GPS passenger estimation and the real data. The result is shown in table 4.9:

In the regression analysis, the R^2 equals to 0.653 with the regression coefficient equals to 998.117. From table 4.9, we can see that the estimation result 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. One of the possible reason of this error is, there are many other stations is located very closed to Ueno Station, including Keisei-Ueno, Inaricho, Uguisudani, Ueno-Hirokoji, Okachimachi, Naka-Okachimachi, Shin-Okachimachi and Yushima, those stations are all within 500 meters from Ueno station. Thus, as the data is heterogeneous, the probability of detection failure is increased.

As the census data from NLNI also provides a classification of total passengers and commuting passengers. we will also provide a comparison of the commuting passenger rate between GPS data estimation and real data.

Table 4.10 shows the result, in general, the estimation result is quite close to the real data in most of the stations. The only error still happens in Ueno station, whose value of GPS estimation is

TABLE 4.10: Table of Comparison of Passenger Occupation in Commuting Time

Station	Rate of GPS Estimation(commuting)	Rate of census data(commuting)
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%

much larger than the real rate. The reason of the error might be similar to the error of passenger estimation. Thus, in the future, it will have a important topic on improving the estimation accuracy in the stations with complex networks and located in a area with multiple different stations.

4.3.5.2 Comparison with the Result of Navigation GPS Data

In the part of comparison with the navigation GPS data estimation, we imported the result from former researchers who research on navigation GPS data. At first, it is important to compare the accuracy of estimation by each data. According to the papers and thesis from former researchers, the correlation coefficient of navigation GPS data is 0.832, with the rank correlation coefficient is 0.908 and both of the p-values. But in our research with Target Ads GPS data, the correlation coefficient is 0.808, which is similar than the result of navigation GPS data, but the rank correlation coefficient is 0.661, which is obviously lower. Both of the p-values are less than 0.01, which shows the correlation is significant in Target Ads GPS data. From the comparison, we can find that the accuracy of Target Ads GPS data is slightly lower than the result of Navigation GPS data. As the quality of Target Ads GPS data is rather limited and heterogeneous than navigation GPS data, in the future, there is also a significant work to optimize the data processing methodology to make if more suitable for the Target Ads GPS data.

On the other hand, we provide another comparison of the passenger estimation in a specific station via two kinds of data. Ikezawa et al. [12] also chose Shinjuku station to estimate the passengers, which was included in our research. 4.14 is the comparison of both two results. The trend is similar in both results. The peaks, which replaces the 'rush hour' of commuting all occur in 8:00-9:00 and 18:00-19:00. But by contrast, the amounts of passengers counted by Target Ads GPS data is smaller than Navigation GPS data. According to my introduction, the users of Target Ads app is larger than the users of Navigation GPS app. There are several possible reasons to

explain the anomaly. One is the map matching methodology is not so applicable to the Target Ads GPS data so that the recognized train trajectories become fewer. The another is that the lower quality of Target Ads GPS data itself causes the abnormality. Anyway, this result urges latter researchers to improve the accuracy and make the quality of estimation higher and higher.

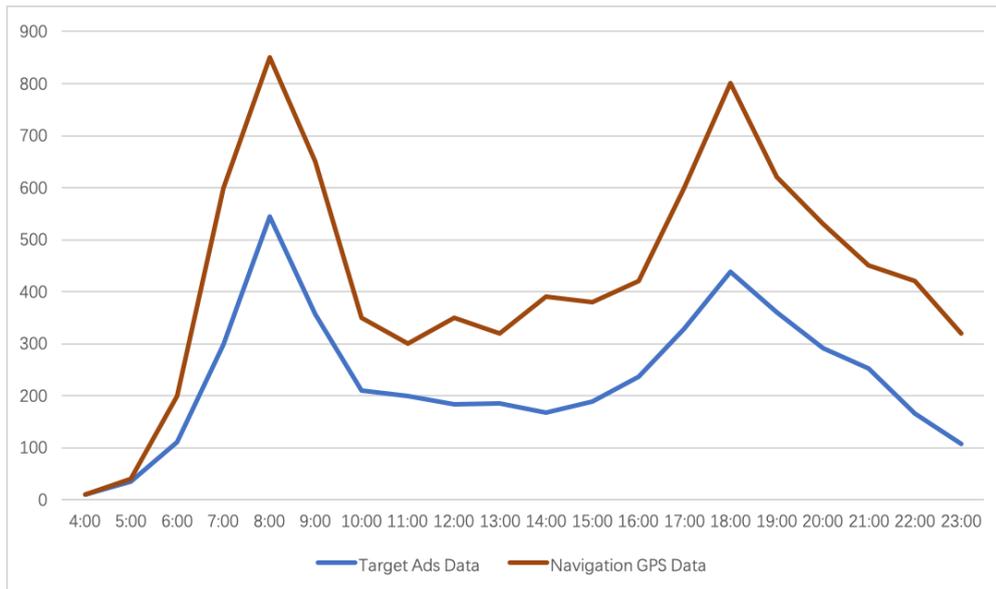


FIGURE 4.14: Passenger Estimation of Shinjuku Station with Target Ads GPS data and Navigation GPS data

4.4 Review and Discussion

In this research, we mainly utilized Target Ads GPS data. At first, we need to filter those IDs with low density of records to make the data become more dense, then improve the accuracy. After that, we classified the life patterns of residents. This work has double significance, firstly it is a part of pre-process which help us to have a better result of traffic mode detection. Secondly, as a part of mobility analysis, it indicates that the most common crowd of Japanese is those people who have a job or students, those people need to commute. As railway transportation is the most common method for commuting in Japan, the result expresses the significance of studying on monitoring railway passenger in Japan. Then we utilized traditional GIS methodology and found that the residents of center part of Tokyo Metropolitan Area have a preference to travel by train, thus we defined seven main stations in Tokyo for researching. After the work of passenger estimation, we found that some stations are mainly used as a interchange station while others are always regards as a destination. And the occupation of passenger volume in commuting time is also different, Kita-Senju is a station which is mostly used for commuting

and exchange. Finally, we compared the result with real census data and the result of navigation GPS data, then we found that the estimation result of Ueno is shown bad in both passenger volume estimation and passenger occupation estimation of commuting time. This result indicates that the recent data processing methodology needs to be optimized as the Target Ads GPS data is rather heterogeneous. When the traffic system network and layout is complex, the error will become large. Then based on the comparison with the result of navigation GPS data, though the users of Target Ads GPS data is much more than those of navigation GPS data, the counted railway passenger via Target Ads GPS data is lower than the former. This result shows that the map-matching methodology also needs to be updated, to decrease the failure of distinguishing the travel mode of train. So far, we monitored railway passenger in multi-aspects and indicated some possible directions of future work.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The conclusions of this research can be summarized as follows:

- 1 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 percents of people prefer to utilize trains for commuting. Thus, studies on monitoring railway passengers have a great significance in Japan.
- 2 Within the selected stations in our research, according to the result of passenger estimation, some stations are mostly utilized by passengers as a interchange station. Moreover, the percentage of passengers 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.
- 3 According to the comparison with real census data, the accuracy of estimation on each station via Target Ads GPS data is also various. The stations such as Tokyo and Shinjuku have results with high quality, but the noise in Ueno station is quite large. 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.
- 4 According to the 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.

Based on those conclusions, we can find an orientation of future works on Target Ads GPS data processing, which will be introduced in next section.

5.2 Future Works

As there are still some limitations and shortage in Target Ads Data in this stage, Initially, the following researches should focus on those problems. At first, as I mentioned in former chapters, the quality of Target Ads Data is rather heterogeneous than other GPS data source which have been fully processed such as Navigation GPS Data. But in this research, the methodology of pre-processing works such as interpolation and map matching is the same as the method of other GPS data. Thus, an optimized methodology of pre-processing should be developed in the following researches on Target Ads Data to create a better precision. On the other hand, only a short-term data was selected for this research due to the limitation from time cost of processing such a large-sized data. In the future, it is necessary to analysis the data with larger time-scale.

Besides of the deficiencies to be solved, some extensions of this research are also valuable to be considered. In this research we mainly focus on monitoring the life patterns of passengers and mobility of railway passengers. But we just considered those 'normal situation'. In the following works, some anomalies such as big event, accident and bad weathers should also be considered. Using some deep learning algorithms can make a valuable discussion of those issue. An another potential topic to be studied is other transportation methods in public transportation system, such as bus or taxi. For the limitation of technology, it is hard to distinguish the trajectories of bus or taxi from which of cars. To solve this problem, multiple types of data which focus on tracking bus or taxi should also be joined. Perhaps only in this way, we can make an all-around evaluation of the public transportation system in Tokyo or Japan, not only railway.

Bibliography

- [1] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.
- [2] Bernstein, D. and Kornhauser, A. (1998). An introduction to map matching for personal navigation assistants.
- [3] Chu, K. K. A. and Chapleau, R. (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation research record*, 2063(1):63–72.
- [4] Chung, E.-H. and Shalaby, A. (2005). A trip reconstruction tool for gps-based personal travel surveys. *Transportation Planning and Technology*, 28(5):381–401.
- [5] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.
- [6] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- [7] Fielding, G. J., Babitsky, T. T., and Brenner, M. E. (1985). Performance evaluation for bus transit. *Transportation Research Part A: General*, 19(1):73–82.
- [8] Filip, A. (2010). Which of eggnos navigation modes for railway signalling: Precision approach or en route? *Proc. CERGAL*.
- [9] Filip, A., Bažant, L., and Mocek, H. (2010). The experimental evaluation of the eggnos safety-of-life services for railway signalling. *WIT Transactions on The Built Environment*, 114:735–745.
- [10] Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., and Nordlund, P.-J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on signal processing*, 50(2):425–437.

- [11] Hiroshi, K., Yoshihide, S., and Takehiro, K. (2013). Development of open railway dataset towards people flow reconstruction. In *The 22th conference on GIS Association of Japan*.
- [12] Ikezawa, S., Kanasugi, H., Matsubara, G., Akiyama, Y., Adachi, R., and Shibasaki, R. (2016). Estimation of the number of railway passengers based on individual movement trajectories. In *6th INTERNATIONAL CONFERENCE ON CARTOGRAPHY AND GIS*, page 249.
- [13] Ishihara, N., Zhao, H., and Shibasaki, R. (2002). Tracking passenger movement with ground-based laser scanner. In *Proc. Japan Society of Photogrammetry and Remote Sensing (JSPRS) Annual Conference*, pages 305–308.
- [14] Jiang, R., Song, X., Fan, Z., Xia, T., Chen, Q., Chen, Q., and Shibasaki, R. (2018). Deep roi-based modeling for urban human mobility prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):14.
- [15] Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934.
- [16] Kusakabe, T., Iryo, T., and Asakura, Y. (2010). Estimation method for railway passengers train choice behavior with smart card transaction data. *Transportation*, 37(5):731–749.
- [17] Lehtonen, M., Rosenberg, M., Rasanen, J., and Sirkia, A. (2002). Utilization of the smart card payment system (scps) data in public transport planning and statistics. In *9th World Congress on Intelligent Transport Systems ITS America, ITS Japan, ERTICO (Intelligent Transport Systems and Services-Europe)*.
- [18] Liao, Z., Yu, Y., and Chen, B. (2010). Anomaly detection in gps data based on visual analytics. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 51–58. IEEE.
- [19] Luo, X., Dong, L., Dou, Y., Zhang, N., Ren, J., Li, Y., Sun, L., and Yao, S. (2017). Analysis on spatial-temporal features of taxis’ emissions from big data informed travel patterns: a case of shanghai, china. *Journal of cleaner production*, 142:926–935.
- [20] Morency, C., Trépanier, M., and Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203.

- [21] Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343. ACM.
- [22] Ochieng, W. Y., Quddus, M., and Noland, R. B. (2003). Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, 2(55).
- [23] Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734. Morgan Kaufmann Publishers Inc.
- [24] Qian, X. and Ukkusuri, S. V. (2015). Spatial variation of the urban taxi ridership using gps data. *Applied Geography*, 59:31–42.
- [25] Quddus, M. A., Noland, R. B., and Ochieng, W. Y. (2005). Validation of map matching algorithms using high precision positioning with gps. *The Journal of Navigation*, 58(2):257–271.
- [26] Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation research part c: Emerging technologies*, 15(5):312–328.
- [27] Reddy, S., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2008). Determining transportation mode on mobile phones. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 25–28. IEEE.
- [28] Sanchez, T. W. (1999). The connection between public transit and employment: the cases of portland and atlanta. *Journal of the American Planning Association*, 65(3):284–296.
- [29] Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R., Yuan, N. J., and Xie, X. (2016). Prediction and simulation of human mobility following natural disasters. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):29.
- [30] Stenneth, L., Wolfson, O., Yu, P. S., and Xu, B. (2011). Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 54–63. ACM.

- [31] Sudo, A., Kashiyama, T., Yabe, T., Kanasugi, H., Song, X., Higuchi, T., Nakano, S., Saito, M., and Sekimoto, Y. (2016). Particle filter for real-time human mobility prediction following unprecedented disaster. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 5. ACM.
- [32] Wang, X., Wang, H., and Dang, A. (2000). Research on large-scale dynamic monitoring of landuse with rs, gps and gis. In *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)*, volume 5, pages 2134–2136. IEEE.
- [33] Wang, Y., Nakaoka, Y., Siriaraya, P., Kawai, Y., and Akiyama, T. (2018). Detecting train delays using railway network topology in twitter. *iConference 2018 Proceedings*.
- [34] Witayangkurn, A., Horanont, T., Ono, N., Sekimoto, Y., and Shibasaki, R. (2013). Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone. In *Proceedings of the international conference on computers in urban planning and urban management (CUPUM 2013)*, pages 1–19.
- [35] Xia, T., Song, X., Fan, Z., Kanasugi, H., Chen, Q., Jiang, R., and Shibasaki, R. (2018). Deeprailway: A deep learning system for forecasting railway traffic. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 51–56. IEEE.
- [36] Yang, D., Cai, B., and Yuan, Y. (2003). An improved map-matching algorithm used in vehicle navigation system. In *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, volume 2, pages 1246–1250. IEEE.
- [37] Yu, M. (2006). *Improved positioning of land vehicle in ITS using digital map and other accessory information*. PhD thesis, The Hong Kong Polytechnic University.
- [38] Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. (2010). Understanding transportation modes based on gps data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1):1.
- [39] Zhou, Z., Dou, W., Jia, G., Hu, C., Xu, X., Wu, X., and Pan, J. (2016). A method for real-time trajectory monitoring to improve taxi service using gps big data. *Information & Management*, 53(8):964–977.