

Master's Thesis (2013)

Using Large-Scale, Long-Term GPS Data from Mobile Phones to
Identify Transportation Modes and Analyze Mobility in the
Tokyo Metropolitan Area

携帯電話による大規模・長期間の GPS データを用いた、
東京都市圏における交通モードの推定およびモビリティの分析

Ono, Natsumi

大野 夏海

The University of Tokyo, Graduate School of Frontier Sciences

Department of Socio-Cultural Environmental Studies

TABLE OF CONTENTS

1. INTRODUCTION.....	4
1.1 Background.....	4
1.2 Objective.....	6
1.3 Related works	6
2. METHODOLOGY	9
2.1 Overall process	9
2.2 About our data	11
2.2.1 Main dataset	11
2.2.2 Ground truth dataset.....	13
2.2.3 Validation dataset	14
3. EXTRACTION OF TRIP DATA	16
3.1 Stay points	16
3.1.1 Identifying stay points.....	17
3.1.2 Clustering stay points	18
3.2 Trips.....	20
3.2.1 Extracting individual trips	21
3.2.2 Grouping individual trips	23
3.3 Results of test data.....	24
4. IDENTIFICATION OF TRIPS	27
4.1 Classifier.....	27
4.1.1 Input: candidate features	27
4.1.2 Output: decision tree	30
4.2 Discussion.....	32
4.3 Results of test data.....	33
5. FINAL RESULTS	35
5.1 Processing results	35
5.2 Correlation analysis	37
5.2.1 Raw values for all trips.....	37

5.2.2	Weighted values for all trips.....	39
5.2.3	Weighted values for trips between different zones.....	41
5.3	Discussion.....	43
6.	MOBILITY ANALYSIS	45
6.1	Organization of data	45
6.2	Trip results	47
6.2.1	Distribution of users	47
6.2.2	Number of trips	48
6.2.3	Distance of trips.....	49
6.3	Transportation mode results	50
6.3.1	Modal share	50
6.3.2	Relation to railway station proximity	52
6.3.3	Relation to trip distances	54
7.	CONCLUSION.....	57
7.1	In summary.....	57
7.2	Future works.....	58
8.	REFERENCES.....	59

1. INTRODUCTION

1.1 Background

In recent years, information regarding the flow of people is becoming increasingly available, and considered valuable in various situations. Not only is such data essential in business fields such as marketing and public services, but it can also be used in the event of disasters. For example, the massive earthquake that occurred in eastern Japan on March 11th, 2011, disrupted transportation networks in the Tokyo Metropolitan Area and left millions of commuters stranded, drawing further attention to the importance of understanding daily travel patterns.

Until recently, such data was obtained primarily through household travel diary surveys, which are expensive and time-consuming to distribute, have filled out, collect, convert to data, and aggregate. They are also prone to human errors; for example, respondents may unintentionally forget to record a less significant trip, or simplify their departure or arrival times by rounding to the nearest hour or half hour. Most importantly, survey data provide information for a limited time period, such as the travel patterns of a single day.

On the other hand, the spread of mobile phones has made it possible to accumulate Global Positioning System (GPS) data in large scales (up to hundreds of thousands of people) for long periods of time (from several months to several years). Although data quality may not be optimal, such information is already proving quite useful in estimating population distribution. For example, online “congestion maps” are provided by ZENRIN DataCom Co., Ltd. (Figure 1-1). Fragmentary GPS data provided by mobile phone users are used to estimate and map the number of people in a certain area at any given time of a day. Grid cells of varying sizes (depending on how closely the user zooms in on the online map) are shaded according to the population density at that moment.

However, it is important to note that while this service examines the concentration of GPS data from multiple users, data from one individual are simply a sequence of points, as shown in Figure 1-2. Logs indicate when and where the user traveled, but does not indicate how the user traveled between these locations.



Figure 1-1 Image of online “congestion map”
 ZENRIN DataCom CO., Ltd., < <http://lab.its-mo.com/densitymap> >

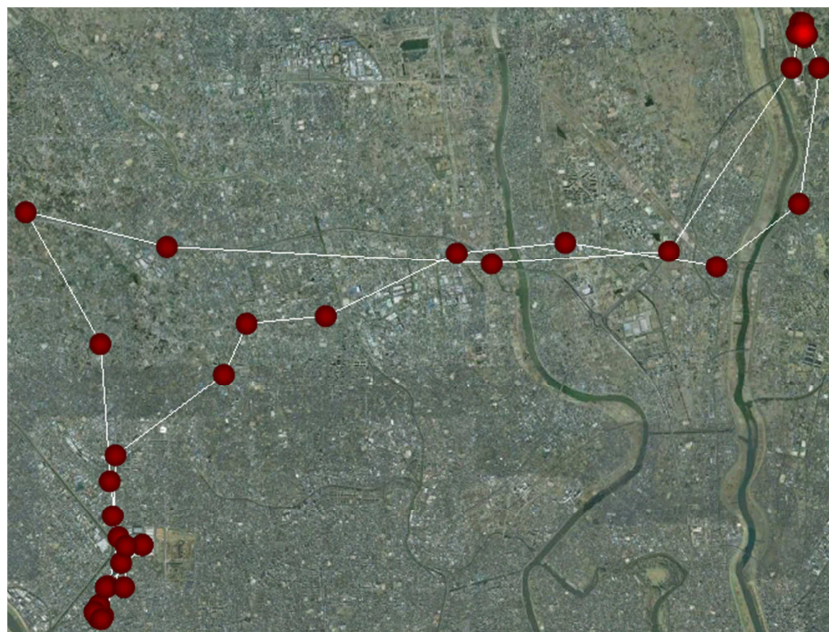


Figure 1-2 Example of GPS data from mobile phones, for one day from one person

1.2 Objective

Therefore, the objective of this study is to use large-scale, long-term GPS data from mobile phones to identify how users traveled from one place to another, or in other words, to identify the main mode of transportation that was used for each trip.

Our results can be used to conduct a mobility analysis, and examine how individuals depend on various on various transportation modes over a long period of time. Such an analysis is currently not possible with the aforementioned travel surveys, but is necessary for urban planning purposes as travel patterns differ by day.

For this study, we used data from the Tokyo Metropolitan Area, one of the world's most densely populated urban areas.

1.3 Related works

As the method of using GPS records to identify travel records has been gaining attention for some time, many similar studies exist. Most of these works begin with the segmentation of GPS logs into individual trips, usually when there is a significant drop in speed¹, or when GPS logs remain in one area for a certain amount of time². Each extracted trip is then identified into one of several transportation modes, including “walk”, “bicycle”, “car”, “bus”, or “rail”, using a variety of parameters.

De Jong et al. conducted fundamental studies by developing a set of rules to determine the trip ends and transportation modes used by an individual wearing a GPS logging device³, and Chung et al. developed a trip reconstruction tool that identified road links and transportation modes⁴. Both studies overlaid GPS data with GIS data, or geographical information such as road networks, bus routes and railways, and railway stations and bus stops. Stopher et al. adds a survey for acquiring personal information, including home location and vehicle accessibility, to help identify trip mode and also trip purpose⁵. Car or bicycle ownership can help confirm the identification of such

¹ Robert de Jong and Wytse Mensorides. (2003). Wearable GPS device as a data collection method for travel research. Working Paper, ITS-WP-03-02, Institute of Transport Studies, University of Sydney

² Hongmian Gong, Cynthia Chen, Evan Bialostozky, and Catherine Lawson. (2011). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*

³ Ibid.

⁴ Eui-Huan Chung, Amer Shalaby. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Journal of Transportation Planning and Technology* 28 (5), 381–401.

⁵ Peter Stopher, Eoin Clifford, Jun Zhang, and Camden FitzGerald. (2007). *Deducing Mode and Purpose*

modes by determining if use of such a transportation mode was possible.

While most studies conduct evaluation of their methods using census data, Bohte et al. developed an interactive online validation application where GPS users could confirm the transportation modes that were identified for their trips⁶. Furthermore, though most studies use a hierarchical labeling method, in recent years Shuessler et al.⁷ and Xu et al.⁸ have introduced a “fuzzy approach”, which uses ground truth data to determine the appropriate speed, acceleration, etc. ranges for each transportation mode, and classifies each GPS log accordingly.

Finally, as GPS accuracy can be heavily influenced in urban environments, Gong et al. conducted a study in New York City, the largest city in the United States, and used variables unique to the area, for example considering the average speed of traffic in that area⁹.

Basic information about each of these studies is shown in Table 1-1. However, most of these studies use GPS loggers, which tend to have short acquisition intervals, creating clear trajectories. Mobile phones, on the other hand, have longer intervals, making data sparser and providing less information. On the other hand, mobile phones have the advantage of being able to be collected for a large number of people over a long period of time.

from GPS Data, paper presented to the Transportation Planning Applications Conference of the Transportation Research Board, Daytona Beach, Florida, May

⁶ Wendy Bohte and Maat Kees. (2008). Deriving and validating trip destinations and modes for multiday GPS based travel surveys: An application in the Netherlands. Paper presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC

⁷ Nadine Schuessler and Kay W. Axhausen. (2009). Processing GPS Raw Data without Additional Information. In Paper presented at the 88th annual meeting of the transportation research board, Washington, DC

⁸ Chao Xu, Minhe Ji, Wen Chen, Zhihua Zhang. (2010). Identifying Travel Mode from GPS Trajectories through Fuzzy Pattern Recognition. In Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery

⁹ Ibid., 6.

Table 1-1 Related works

first author (year)	modes	location	number of individuals	duration	data source	GPS interval
de Jong (2003)	walk, bus, rail, car	-	-	-	-	-
Zenji (2005)	walk, car, rail	Akihabara, Kinshicho, Tsudanuma (Japan)	N/A	1 day	GPS loggers	10 seconds
Chung (2005)	bus, car, bicycle, walk	Toronto (Canada)	60	-	GPS loggers	-
Stopher (2008)	walk, rail, bus, bicycle, car	Sydney (Australia)	-	-	-	not constant
Bohte (2008)	walk, bicycle, car	Amersfoort, Veenendaal, Zeewolde (Netherlands)	1104	1 week	GPS loggers	6 seconds
Schuessler (2008)	walk, bicycle, car, bus, rail	Zurich, Winterthur, Geneva (Switzerland)	4882	6.65 days (average)	GPS loggers	-
Xu (2010)	walk, bicycle, bus, rail, rest	Shanghai (China)	32	142 days	GPS loggers	-
Gong (2011)	walk, rail, bus, car	New York City (U.S.)	63	5 days	GPS loggers	5 seconds

2. METHODOLOGY

2.1 Overall process

For this study, we assume that large-scale, long-term GPS data refers to logs of several hundred thousand people for approximately one year. We also assume that GPS acquisition intervals range from several minutes to several hours, and that the margin of error is between several meters to several hundreds of meters. To improve mode identification accuracy, as well as to identify specific travel patterns, we overlay the long-term data of individuals and identify recurring trips. Our overall process, as illustrated in a simple flowchart in Figure 2-1, consists of the following steps:

Identification and clustering of stay points

First, we segment trips and remove non-trip data by extracting GPS logs recorded when the user stopped moving, hereafter referred to in this study as “stay points”. Since we overlay long-term data, we then identify which stay points refer to the same locations, or places that the user visits multiple times throughout the year, by clustering stay points that are located close to one another.

Extraction and grouping of individual trips

Next, we extract sets of consecutive non-stay points as trip data, identifying each by the combination of stay point clusters used as its origin and destination. Again we group similar trips, defining two or more “individual trips” to be the same “distinct trip” if they use the same combination of origin and destination clusters. This process helps make sparse GPS data become denser, providing more information for our next process.

Identification of distinct trips

At this point, we identify each set of distinct trips by the transportation mode that was most likely to have been used. The necessary parameters and thresholds were determined using ground truth data, or GPS logs that have been labeled by the users. As our dataset is very sparse in quality, for the purpose of this study we focus only on identifying one of three modes: walk, rail, or car, with car trips including all trips on buses, in taxis, and on bicycles.

Validation and analysis of results

Finally, we validate our results by comparing the total number of individual trips, by transportation mode, from one area to another, with the results of Person Trip Surveys. We then use our results to calculate the long-term modal shares of individuals, to compare how people living in different regions of the Tokyo Metropolitan Area depend on different transportation modes throughout their daily lives.

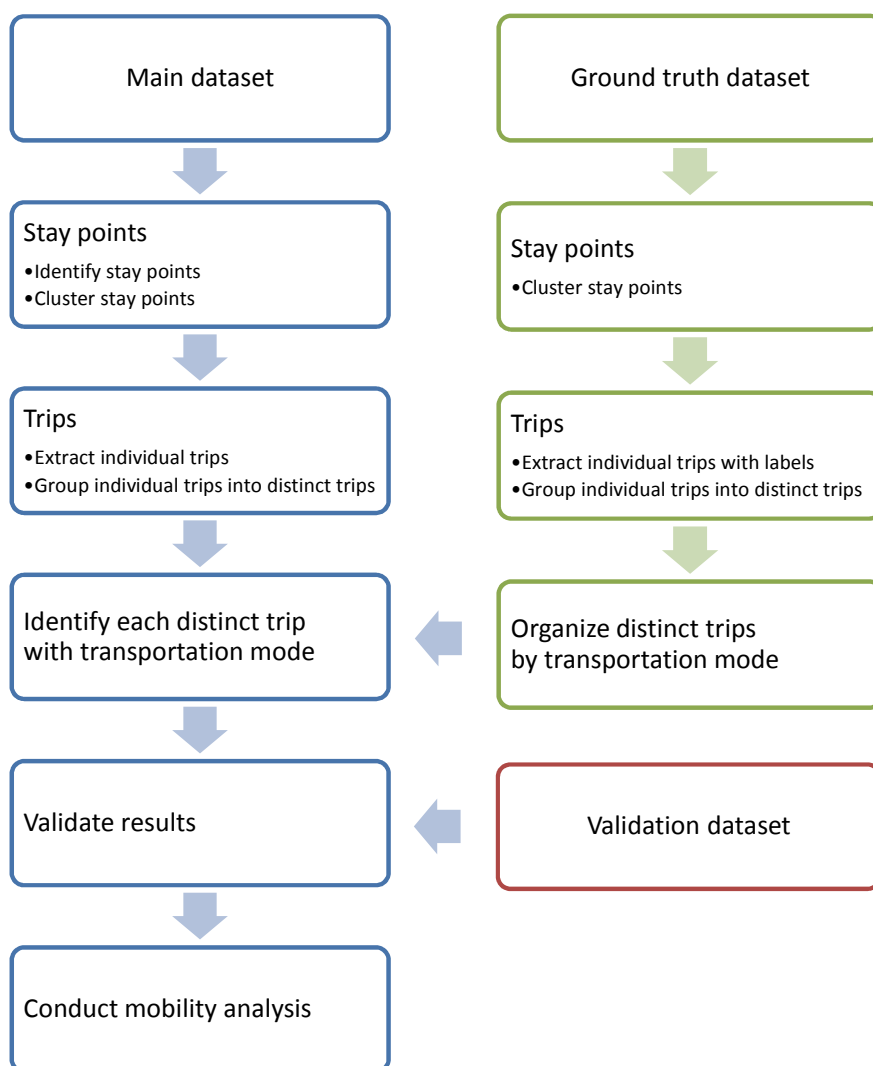


Figure 2-1 Flowchart of overall process

2.2 About our data

This study uses three different datasets: our main, large-scale and long-term GPS data; labeled GPS data as ground truth; and travel survey data for validation purposes.

2.2.1 *Main dataset*

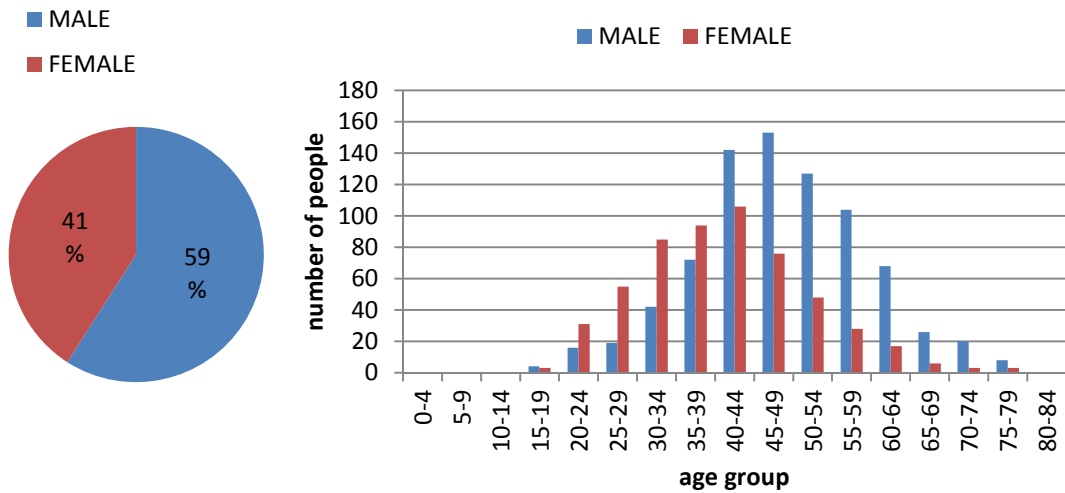
Our main dataset consists of statistical data provided by ZENRIN DataCom Co., Ltd. GPS logs were collected from the approximately 1.5 million users of a certain mobile phone service, provided by a leading mobile phone operator in Japan. Although the dataset includes short-term users of this service, there are estimated to be about 800,000 users' worth of year-long data. GPS data was recorded between August 2010 and July 2011, with intervals at a minimum of 5 minutes, but only when movement was detected, and only when GPS reception was available. Therefore, the average number of GPS logs collected was 37 points per individual per day. For each GPS log we use the following information: a user ID number, a timestamp of date and time (in seconds), and location coordinates represented as longitude and latitude.

For the purpose of this study, we use the approximately 221,100 individuals observed to have spent the majority of their time in the following four prefectures: Tokyo Prefecture, Kanagawa Prefecture, Chiba Prefecture, and Saitama Prefecture. From this dataset, we randomly select nine users to use as test data for experimentation. These users satisfied the following requirements: each had a minimum of 5,000 logs, with at least 95% located in the aforementioned four prefectures.

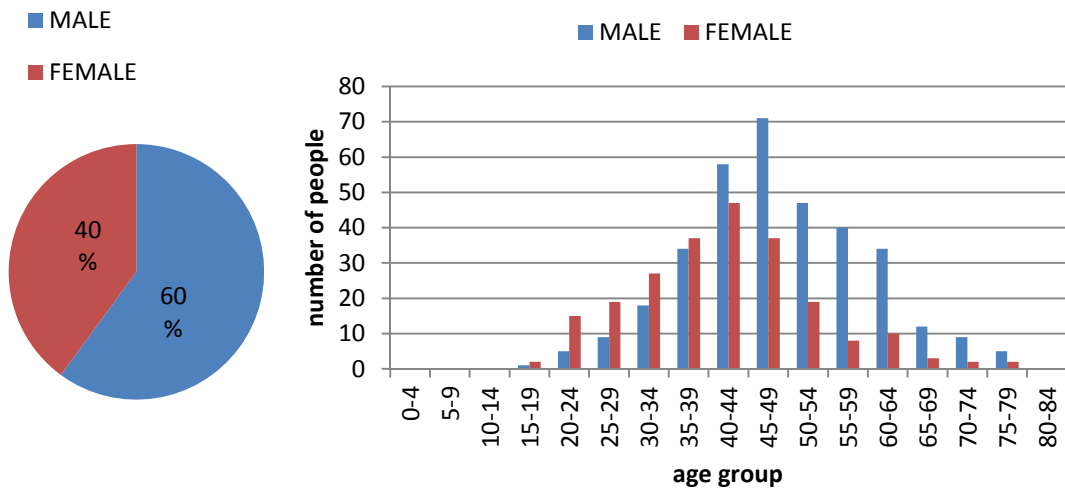
It is important to acknowledge that there is some selection bias in our dataset, as participants are limited to users of a specific mobile phone service. Therefore, we estimate the distribution of personal attributes for our dataset by using the results of a survey conducted a few years ago, about mobile phone-based navigation services and games. Respondents answered personal information, including gender, age, and home addresses, as well as which mobile phone services they had used within the past year.

Of the 50,000 respondents of this online survey, 1,356 (about 2.6%) replied to using this service. As there are estimated to be over 100 million mobile phone users in Japan, and our dataset of 1.5 million users comprises 1.5% of this number, we can infer that users

were slightly overrepresented in this survey¹⁰. Figure 2-2 (a) shows the gender ratio and age distribution of respondents who use this service. The bottom half (b) indicates the same information, but for the 571 respondents who also reside in the Tokyo Metropolitan Area (42.1% of total service users). Note that users in their twenties and thirties are more likely to be female and those forty and over are more likely to be male.



(a) All service users



(b) All service users who live in the target area

Figure 2-2 Gender ratio and age distribution of users from survey data

¹⁰ Telecommunications Carriers Association < <http://www.tca.or.jp/database/index.html> >

2.2.2 *Ground truth dataset*

Ground truth data is used to determine our algorithm for identifying transportation mode. In a separate study, 160 individuals used the GPS features in their mobile phones for approximately one month, from November 28th to December 22th in 2011. Stay points were extracted automatically, and participants used an online application to confirm the main transportation modes used for each extracted trip.

We process this dataset using the same methodology used in our main study. Stay points are clustered, and trips with the same combination of origin and destination clusters are grouped, creating distinct trips made up of multiple individual trips. By imitating this methodology, we prepare sets of trip data that are similar to the ones used in our main study. (It is important to note, however, that this dataset differs from our main dataset in that it has only been collected over one month, and therefore the numbers of individual trips for each distinct trip are fewer.)

In this dataset, each individual trip has been labeled with the actual transportation mode. As shown in Table 2-1, there were a variety of candidate modes for study participants to choose from, and therefore each label is converted into one of the three modes used in this study.

As transportation modes have been labeled by individual trip instead of distinct trip, it is possible for individual trips belonging to the same distinct trip to be labeled with different modes. Therefore, we label each distinct trip with the transportation mode that has the largest number of individual trips. (In the rare event that a tie occurs between two or more transportation modes, that distinct trip is considered faulty and is removed from our ground truth dataset.)

Table 2-1 Labeled modes and converted modes, for ground truth data

Labeled modes	Converted modes
Unclassified, Other	Unclassified
Rail	Rail
Bus, Car, Taxi, Motorcycle, Bicycle	Car
Walk	Walk

2.2.3 Validation dataset

The final results of our study are validated using aggregated results of Person Trip Surveys, questionnaires where respondents are asked to write down the details of every trip they took during a single day. These details include the time of departure and arrival, trip purpose, as well as the transportation mode used. This survey has been conducted in the Tokyo Metropolitan Area, in this case defined as Tokyo Prefecture, Kanagawa Prefecture, Chiba Prefecture, Saitama Prefecture, and the southern half of Ibaraki Prefecture, every ten years since 1960. Our study uses data from 2008, which consists of approximately 800,000 randomly chosen people (5 years or older), or roughly 2% of the 36 million residents, who live within the Tokyo Metropolitan Area. Respondents are asked to write about any weekday between October and November¹¹.

Aggregated data was provided by the National-Land Information Office website¹². Personal attributes of respondents are used to multiply trip count so that the dataset becomes representative of the actual population. The Tokyo Metropolitan Area is segmented into 601 zones (164 in Tokyo, 154 in Kanagawa, 113 in Chiba, 118 in Saitama, and 52 in the southern half of Ibaraki), and trips are represented as traveling from one zone to another, as shown in Figure 2-3. All trips with the same pair of origin zones and destination zones are combined together, according to transportation mode, and referred to in this paper as a set of “grouped trips”. Therefore, the dataset is organized so that each line of data represents a single grouped trip, including the following features: origin zone code, destination zone code, trip count for surveyed modes, and total trip count. See Table 2-2 for a list of surveyed modes and how we converted them for use in this study.

Table 2-2 Labeled modes and converted modes, for validation data

Surveyed modes	Converted modes
Rail	Rail
Bus, Car, Bicycle/Motorcycle	Car
Walk	Walk

¹¹ Tokyo Metropolitan Region Transportation Planning Commission <<http://www.tokyo-pt.jp/index.html>>

¹² National-Land Information Office, <<http://www.mlit.go.jp/kokudoseisaku/gis/index.html>>

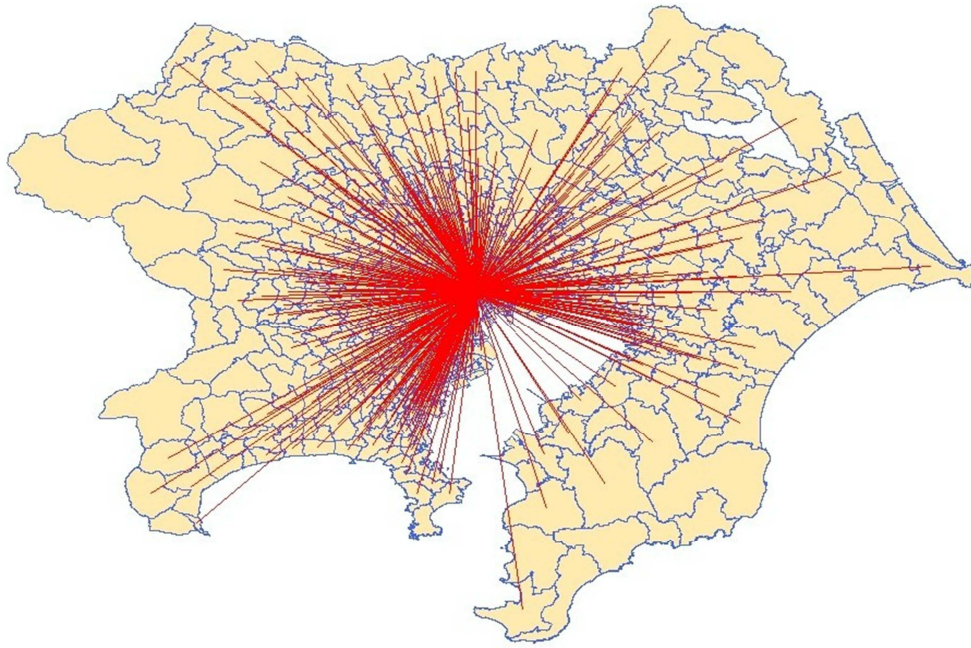


Figure 2-3 Image of aggregated trip data from Person Trip Surveys
National-Land Information Office, < <http://www.mlit.go.jp/kokudoseisaku/gis/index.html> >

3. EXTRACTION OF TRIP DATA

3.1 Stay points

The first half of the process for extracting trip data consists of separating trip data from non-trip data, or stay points. This process is illustrated in Figure 3-1, where we suppose an area for three days' data overlap. Stay points are identified, clustered, and labeled.

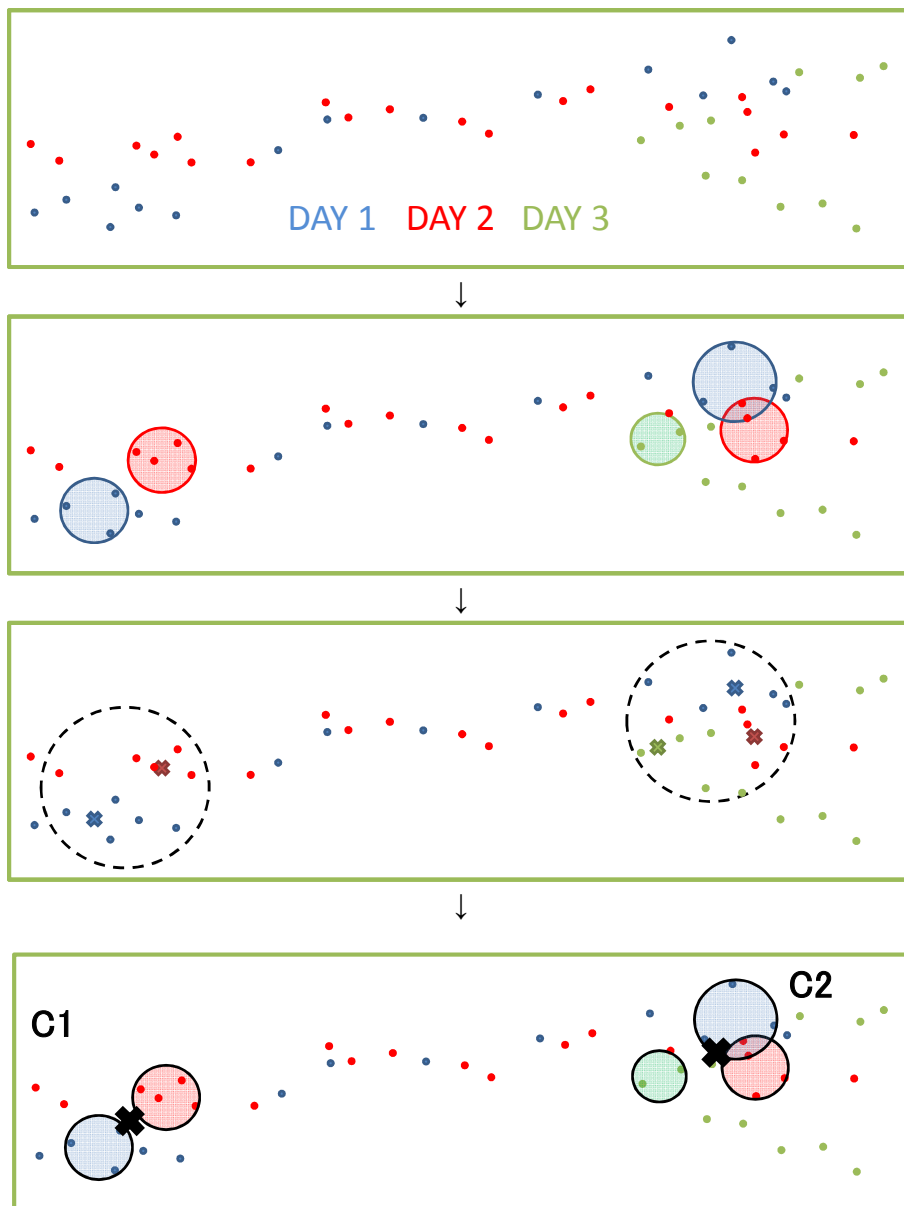


Figure 3-1 Illustration of identifying, clustering, and labeling stay points

3.1.1 Identifying stay points

First, we determine if each log was recorded while traveling or not. Previously published studies, which used GPS loggers, extracted logs with a significantly slow speed¹³, or logs concentrated in a certain area¹⁴, to indicate when users stopped moving. But as our study uses mobile phones, which record GPS data at longer and inconsistent intervals, speeds cannot be accurately calculated, and the time spent in an area is not necessarily proportional to the number of logs recorded there.

Therefore, we extract stay points by selecting GPS logs that are located within a certain distance of each other for longer than a specific length of time. As in the work of Witayangkurn et al.¹⁵, we assume that a set of GPS points for one individual are represented as $P = (p_1, p_2, \dots, p_n)$ where $p = (\text{id}, \text{time}, \text{lat}, \text{lon})$ and $n =$ the total number of points. We can then apply the following equation¹⁶:

$$\text{Distance}(p_{\text{start}}, p_{\text{end}}) < D_{\text{threh}} \text{ and } \text{TimeDiff}(p_{\text{start}}, p_{\text{end}}) < T_{\text{threh}}$$

where D_{threh} and T_{threh} are adjustable parameters. D_{threh} is the maximum diameter of an area considered as a stay point, whereas T_{threh} is the minimum time spent in that area. After some experimentation, we decided to set $D_{\text{threh}} = 150$ meters and $T_{\text{threh}} = 20$ minutes. This time frame ensures that stops for transferring from one transportation mode to another, such as rail transfers which are common in Tokyo, are not extracted. Separate stay points are identified by numbering them in the order that they are extracted.

¹³ Ibid., 6.

¹⁴ Daniel Ashbrook and Thad Startner. (2003). Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, v.7 n.5, p.275-286.

¹⁵ Apichon Witayangkurn, Teerayut Horanont, Yoshihide Sekimoto, and Ryosuke Shibasaki. (2010). Large Scale Mobility Analysis: Extracting Significant Places using Hadoop/Hive and Spatial Processing. Technical Study.

¹⁶ Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. (2012). Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, 1-29.

3.1.2 Clustering stay points

In our next step, we cluster stay points to identify which ones refer to the same locations, and therefore may be considered the same trip origin or destination. Note that during this process, stay points are represented as a single point using the coordinates of their centroids. For this process, we considered various clustering algorithms.

First we considered OPTICS (Ordering Points To Identify the Clustering Structure) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), both well-known algorithms for finding density-based clusters in spatial data. This method is convenient for extracting frequently visited places, such as home or work areas, but as the purpose of this study was to extract all trips and therefore all possible trip nodes, we determined it to be inadequate.

Another common algorithm is the K-means clustering method, which groups data into a predetermined k number of clusters where each point belongs to the cluster with the nearest mean¹⁷. An n number of points (x_1, x_2, \dots, x_n) is divided into k clusters ($k \leq n$) $S = \{S_1, S_2, \dots, S_k\}$ with the aim of minimizing the following squared error function:

$$J = \sum_{j=1}^k \sum_{i=1}^n \left\| x_i^{(j)} - c_j \right\|^2$$

where c_j is the mean of points in S_j . As a result, data space is partitioned into Voronoi cells. However, in this study the number of clusters k represents the number of different locations visited by a single person. As this value is sure to vary for each individual, yet difficult to predetermine, application of this algorithm seemed difficult.

We also considered a variant of the K-means algorithm which was proposed by Ashbrook et al.¹⁸ for the same purpose of extracting significant locations from GPS data. Instead of the number of clusters, we determine a specific cluster radius. One randomly selected point from the dataset is positioned as the center of a cluster, and all other points within the radius are extracted. The mean of these points becomes the new center, and this

¹⁷ James MacQueen. (1967) Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J. (eds), *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. University of California Press.

¹⁸ Ibid., 17.

process is repeated until the center stops changing. At this point, all extracted points are grouped as a cluster, and removed from the dataset. This procedure is repeated until all points in the dataset have been assigned to clusters. However, since points that have been added to clusters are removed from further consideration, the outcome becomes dependent on the order that points are chosen.

We finally decided to use a combination of canopy clustering and the aforementioned K-means algorithm. Canopy clustering is similar to the K-means, except that it creates overlapping subsets by using two radii thresholds, where points within the larger radius and outside the smaller one are added to the cluster but not removed from the remaining dataset. After some experimentation, we set both thresholds at 500 meters to create a set of temporary clusters. We used the number of clusters generated from this algorithm to conduct the K-means algorithm.

Note that stay point clusters are formed solely for the purpose of identifying similar stay points. Therefore, non-stay points that may be located within the radii of these clusters will remain non-stay points, and will be used as trip data in the subsequent process. Again, we identify stay point clusters by numbering them in the order that they are extracted.

3.2 Trips

After identifying, clustering, and labeling stay points, we use these markers to prepare sets of trip data. This process is illustrated in Figure 3-2, where non-stay points are extracted as individual trips and grouped into distinct trips.

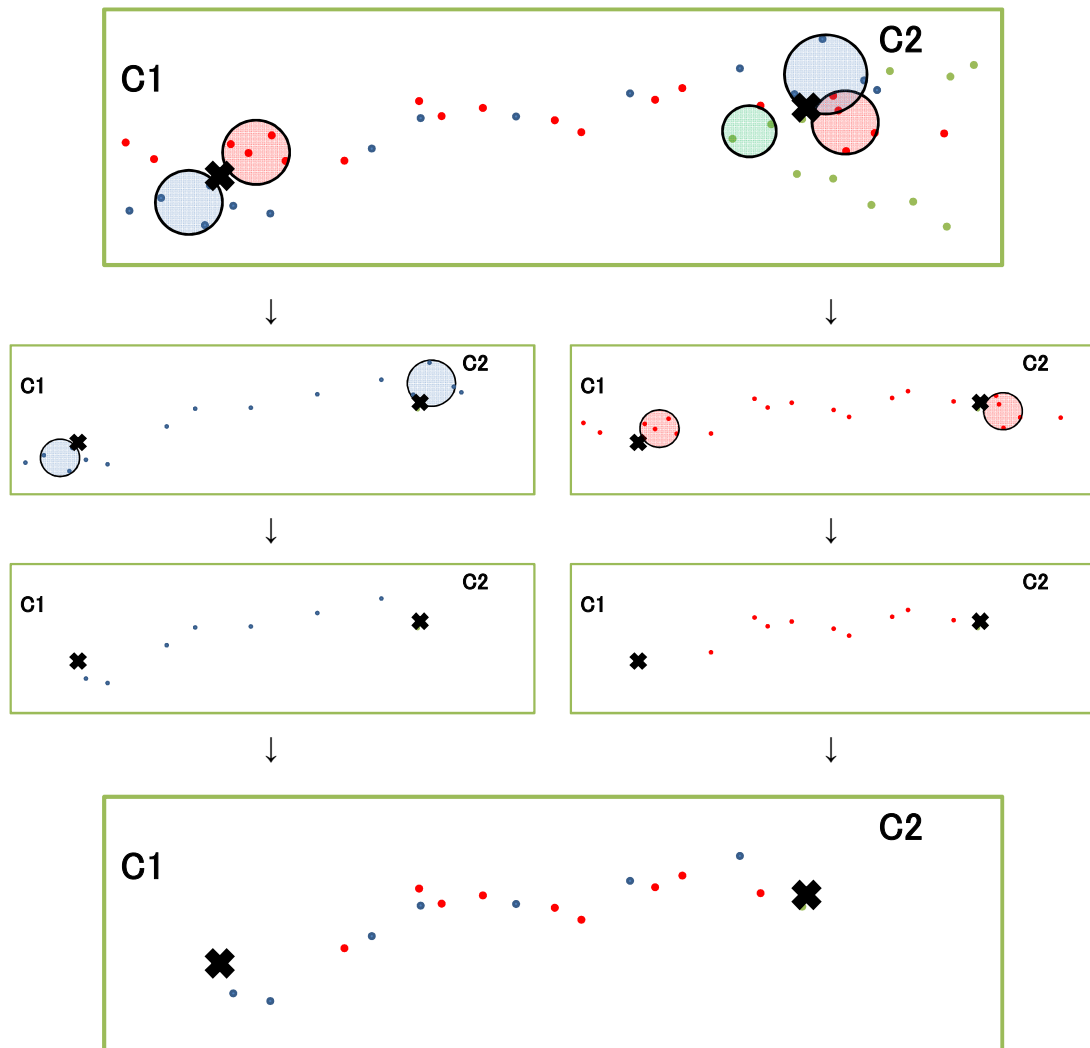


Figure 3-2 Illustration of extracting and grouping individual trips

3.2.1 Extracting individual trips

At this point, each GPS log has been labeled as either a stay or non-stay point, with a specific stay point number and cluster number. Therefore, we extract consecutive non-stay points as sets of trip data, as shown in Figure 3-3. This procedure is conducted throughout the entire dataset of each individual, with no segmentation of days. Each extracted individual trip is identified by the combination of its origin and destination, or the cluster numbers of the stay points immediately before and after it.

LINE	DATE	LON	LAT	ERROR	SPEED	STAY	SPNUM	CLUSTER
472	2010/8/18 12:31	139.7981	35.80481	3	0.4045	0	0	0
473	2010/8/18 12:37	139.7992	35.80555	1	2.6226	0	0	0
474	2010/8/18 12:51	139.8164	35.82092	1	0	1	26	2
475	2010/8/18 12:56	139.8164	35.82092	1	0	1	26	2
476	2010/8/18 13:21	139.8164	35.82092	1	7.6856	1	26	2
477	2010/8/18 13:26	139.7992	35.80555	1	0.5528	0	0	0
478	2010/8/18 14:36	139.8164	35.82092	1	0.6447	0	0	0
479	2010/8/18 15:36	139.7992	35.80555	1	0.0218	1	27	3
480	2010/8/18 17:41	139.8005	35.80625	1	0.3141	1	27	3
481	2010/8/18 17:56	139.7982	35.80483	3	3.0413	1	27	3
482	2010/8/18 18:01	139.7906	35.80138	2	4.4724	0	0	0

Figure 3-3 Process of extracting individual trips





During this process, we identified two different types of trips: those where the origin and destination were separate clusters, and those where they were the same cluster. We call these trips, respectively, types A and B, as shown in Table 3-1. Type B may indicate a situation where the user did not stop for longer than our threshold time (20 minutes) before returning to their former location. However, our algorithm may have failed to extract a stay point if GPS logs stopped recording for long periods of time, which is a high possibility in dense urban environments and especially when users step indoors. For the purpose of this study, we decided to acknowledge such type B trips as trip data.

On the other hand, there were also cases where GPS logs jumped from one stay point cluster to another, without any non-stay point logs in between them. This may occur when GPS logs stop being recorded temporarily, such as if the mobile phone is turned off or its battery is low, or if the user travels through an indoor or underground passageway, as when taking the subway. For this study, the purpose of which is to extract and identify all of the trips made by individuals, we count these occurrences as type C trips. As our next process involves grouping individual trips, type A trips would

provide data for type C trips if they are grouped together..

Conversely, similar situations where GPS logs jump from one stay point to a different stay point, but where both stay points belong to the same cluster, were not considered trips (type D). In all likelihood, GPS logs stopped recording when users stopped moving and generated a time gap, causing two different stay points when there should only be one.

Table 3-1 Organization of trip types

Trip types	Traveling trips (Origin \neq Destination)	Staying trips (Origin = Destination)
Extracted trips (Trip data exists)	Type A: trip data can be used to detect mode 	Type B: returning trip 
Detected trips (Trip data does not exist)	Type C: count as a trip, but no data to detect mode 	Type D: not a trip 

3.2.2 Grouping individual trips

For each individual, trips with the same combination of origin and destination are grouped as distinct trips. In this way, we assume that a user who travels from one specific location to another, multiple times, will always use the same route. Although we understand that this may not always be the case, this assumption allows us to use more GPS points, and thus more information, for each distinct trip; in particular, it provides data for trips categorized as type C in the previous section (3.2.1).

At the same time, this process allows us to calculate the number of different trips a user takes throughout a year. An OD matrix for each individual helps to organize this information, such as the one shown in Table 3-2. Rows represent origin clusters and columns represent destination clusters, and each cell indicates a distinct trip. (Due to space constraints, this matrix has been created using only stay point clusters with ten or more stay points.)

Table 3-2 Example of an OD matrix (partial)

	1	3	4	8	16	18	19	20	22	29	30	31	35	38	44	47	48
1	147	9	14	5	8	9	14	3	12	10	8	5	5	4	3	2	5
3	7	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	9	2	4	0	0	1	0	0	0	0	0	0	1	0	0	1	0
8	10	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
16	8	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	0
18	9	0	2	1	0	12	0	0	0	0	0	0	0	0	0	0	0
19	12	0	0	0	0	0	20	0	0	0	1	0	0	0	0	0	0
20	2	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
22	10	0	1	0	1	0	1	0	14	0	0	0	0	0	0	0	0
29	7	0	0	1	1	0	0	0	0	13	11	0	0	0	0	0	0
30	5	0	1	2	0	0	0	0	0	8	14	0	0	0	0	0	0
31	4	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	0
35	2	0	1	0	1	0	0	0	0	0	0	0	10	0	0	0	0
38	3	0	1	1	0	0	0	0	0	0	0	0	0	5	0	0	0
44	2	0	1	0	0	0	0	0	0	0	0	0	0	0	8	0	0
47	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
48	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	7

3.3 Results of test data

Results of the trip extraction process on our test data, labeled as IDs “A” through “I”, are as shown in Table 3-3. On average there were 724.3 individual trips and 205.2 distinct trips for a full year, and thus about 1.98 individual trips per day. In addition, there were an average of 793.0 stay points and 71.4 stay point clusters for a full year. Therefore, roughly 2.17 stay points are extracted per day.

In the following pages, we visualize some of our results for ID “A”. In Figure 3-4, small red circles indicate the locations of all 12,327 GPS logs and larger blue circles indicate the locations of the 864 extracted stay points. Each trajectory of GPS records appears to end in one or multiple stay points. Figure 3-5 is an enlarged view of the area marked by a white rectangle in Figure 3-4, and shows yellow markers that indicate the centroids of stay point clusters. As should be the case, these markers fall in areas where stay points appear to be concentrated. Figure 3-6 shows an individual trip (the same image as in Figure 1-2) and the distinct trip that it belongs to, which consists of 14 individual trips. Therefore we can confirm that, compared to a single day’s worth of data, logs collected over a long period of time tend to concentrate on the route that was most likely to have been taken. In this way, we can improve mode detection accuracy. Figure 3-7 shows multiple distinct trips (including the one shown in Figure 3-6) and confirms that trip trajectories can be extracted and grouped regardless of their distance.

Table 3-3 Results of trip extraction process for test data

	All Logs	Stay Points	Stay Pnt Clusters	Trips (A)	Trips (B)	Trips (C)	AllTrips (ABC)	Distinct Trips
A	12327	864	53	381	337	39	757	185
B	8031	731	52	199	384	14	597	110
C	15998	890	69	708	141	12	861	273
D	12102	516	65	370	98	2	470	165
E	10410	782	55	425	275	31	731	148
F	6764	646	34	456	86	5	547	85
G	12133	969	33	536	274	92	902	87
H	25400	679	53	522	126	8	656	153
I	25703	1060	229	873	104	21	998	641
Avg.	14318.7	793.0	71.4	496.6	202.8	24.9	724.3	205.2

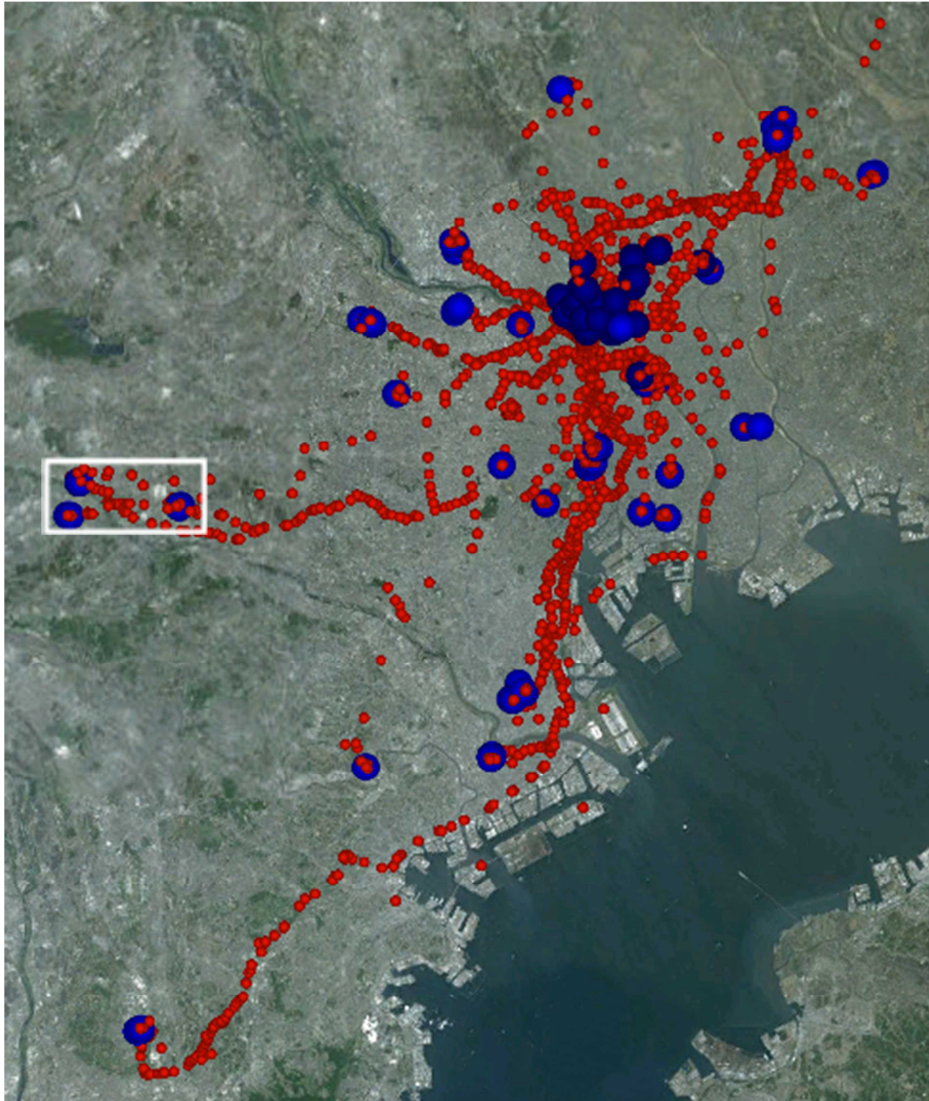


Figure 3-4 All GPS logs (small red circles) and stay points (larger blue circles) for “A”

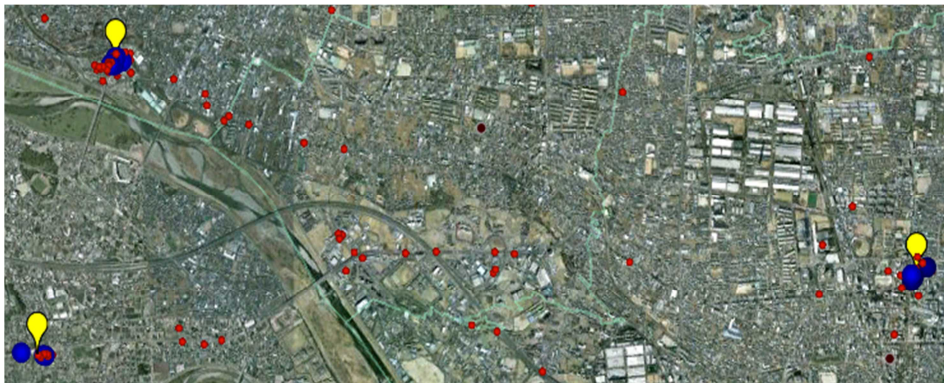
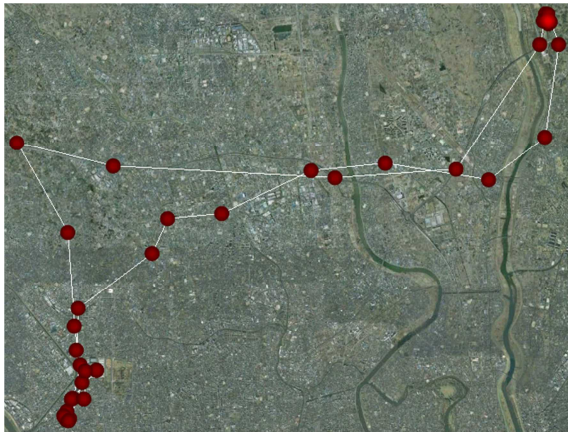


Figure 3-5 Enlarged view of white box in Figure 3-4, with stay point clusters (yellow markers)



(a) An individual trip



(b) A distinct trip

Figure 3-6 Examples of (a) individual trips (same as Figure 1-2) and (b) distinct trips, for “A”



Figure 3-7 Multiple examples of distinct trips

4. IDENTIFICATION OF TRIPS

4.1 Classifier

Finally, we identify the main transportation mode used in each distinct trip by running each group of trip data through a mode detection algorithm. Most previous studies used speed as a primary feature in separating walk from other modes, as in the works of Stopher et al.¹⁹ and Gong et al.²⁰ But through some experimentation with our ground truth data, where trips have been extracted and labeled with the correct transportation modes, we discovered that accurate speed values were difficult to determine from mobile phone-based GPS datasets, where logs are sparse and often inaccurate.

Therefore, we test several different features of ground truth data to determine what other parameters and thresholds are necessary. Our dataset consists of 160 individuals over a month-long period, resulting in a total of 7021 trips. Of this total, 5329 trips (75.9%) were labeled with transportation modes, and 5954 (84.8%) contained GPS logs that we could calculate various features from. For this study we used the 4733 trips (67.4%) that were both labeled and that contained GPS logs.

4.1.1 *Input: candidate features*

For each of these labeled trips, we calculate values for various features, and then use a software called RapidMiner²¹ to process these values and create a decision tree with the appropriate parameters and thresholds. As a result, the following features were used for classifying trips into transportation modes.

Average speed (meters per second)

Speeds were calculated for each GPS point by using the distance and time difference between it and the following point. We define the average speed of each distinct trip to be the average speed value of trip data.

As GPS points collected from mobile phones tend to include errors, especially in dense urban environments, we also tried calculating the average speed of distinct trips by

¹⁹ Ibid., 6.

²⁰ Ibid., 6.

²¹ RapidMiner < <http://www.rapidminer.com/> >

using the distance and time difference between the first and last logs of each trip. However, our decision tree results did not include this feature, judging it to be less relevant. This may be because GPS acquisition intervals are longer than 5 minutes, and because logs failed to be recorded when underground or indoors (as most users will be at the start or end of a trip), and therefore the starting and ending times of each trip were difficult to determine.

Trip distance (meters)

We calculate the direct distance between the centroids of the origin cluster and the destination cluster. Centroids of clusters are used instead of centroids of stay points because they seem to provide a more accurate representation of the visited location.

Proximity to railway network (%)

First, we conduct a simple process for removing GPS points that, supposing rail was used, may have been recorded between the origin or destination and the used railway stations. For this study, we assume that railway users board trains from stations that are closest to their origin or destination (although we acknowledge this may not necessarily be the case), as shown in Figure 4-1. We use a list of stations and their coordinates to find the one that is closest in distance to the origin or destination centroid. We then form a circle that is centered on the midway point between that station and the origin or destination centroid, and that passes through both locations. All GPS points found within this circle are removed. Next, we use railway data to determine which of the remaining GPS logs fall within 100 meters of the network, and calculate that number as a percentage of all logs in that distinct trip.

Both station data and railway network data were provided by the National-Land Information Office, the same source as our validation data²². We used the most recent data available, which was prepared on July 31, 2011.

To confirm if our decision to remove certain GPS logs was effective, we also tried preparing a set of values without conducting the initial removal process. As our decision tree results did not include this feature, we can be certain that this preprocessing helps improve detection accuracy.

²² Ibid., 14.

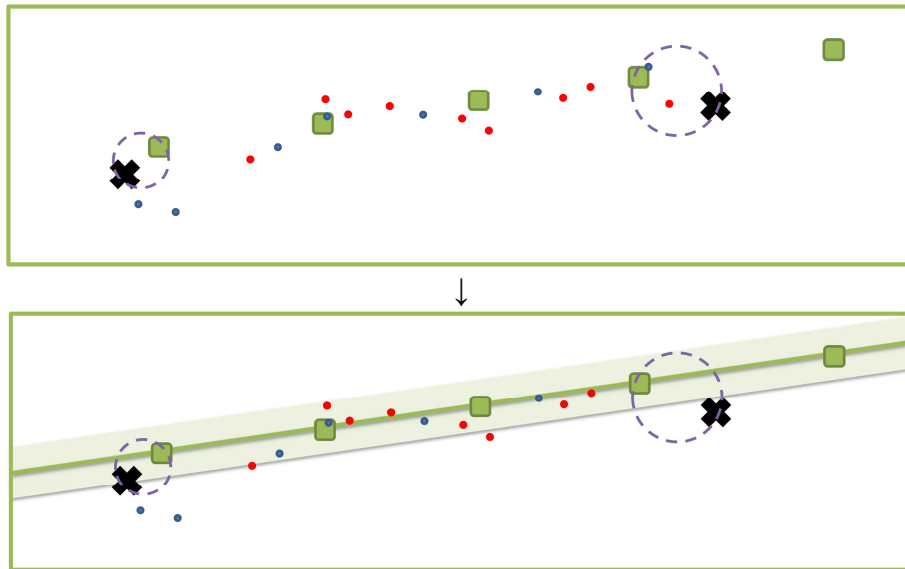


Figure 4-1 Removal of logs prior to calculating railway network proximity

Values for the following features were also calculated and processed, but did not appear in our decision tree results. Therefore, we can determine that they were not relevant enough for classifying transportation modes.

Trip type (A, B, or C)

We labeled each trip with one of the three trip types described in the previous chapter (3.2.1), on the assumption that different types of trips may need different methods of mode identification. However, it seems that trip type does not affect the classification process.

Density of data (number)

We calculated the number of individual trips, as well as the number of GPS logs, for each distinct trip, as we assumed that different data densities may require different methods. Again, this feature was determined as irrelevant.

Accessibility to railway station (meters)

We assumed that trips where origin or destination locations were closer to railway stations were more likely to have used rail as the main mode of transportation. Therefore, we calculated the distances between the origin or destination centroids and their closest stations. However, this feature, too, was less significant than we had assumed.

4.1.2 Output: decision tree

Decision tree results are as shown in Figure 4-2. Proximity to railway network seems to work best as the primary feature for classifying transportation modes, in particular for separating rail and car trips. Trip distance helped separate walk from rail trips, and average speed walk from car trips.

We find that trips with over 55.7% of GPS logs within 100 meters of railway networks are more likely to be classified as either rail or walk, and those with less as either car or walk. This threshold seems to take into consideration erroneous GPS logs found further away from the network. We experimented with increasing the maximum distance value as well, but found this did not improve accuracy.

Walk trips seem slightly more difficult to classify, as they are not limited to specific transportation networks. For trips found located near railway networks, those with a travel distance of less than 1402 meters are more likely to be walk trips. If we assume that the average walking speed is 6 kilometers per hour, we can state that most walk trips took place under 14 minutes.

For trips found located further away from railway networks, those with an average speed of less than 1.3 meters per second, or about 4.8 kilometers per hour, are more likely to be walk trips. This threshold was slightly lower than we expected, but is most likely due to the fact that bicycle trips—which travel more slowly than car trips—are included as car trips as well.

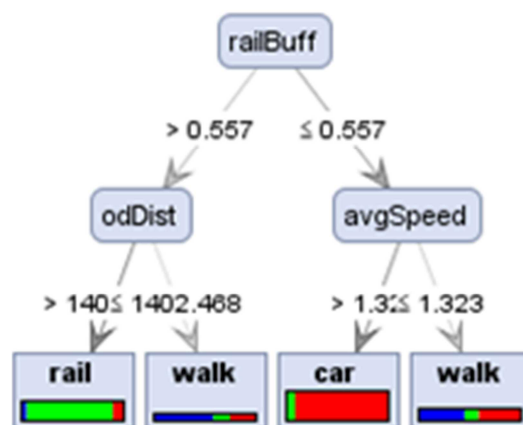


Figure 4-2 Decision tree result

Accuracy results for this decision tree are as shown in Table 4-1. Overall accuracy is calculated by dividing the sum of trips that were labeled and identified as the same mode, or “accurate trips”, by the total number of trips. In addition, precision and recall are calculated for each mode. Precision is defined as the fraction of predicted data that is accurate, and recall is defined as the fraction of labeled data that is accurate (Figure 4-3).

Table 4-1 Accuracy results for decision tree

Accuracy: 80.63%					
		True			Class precision
		Walk	Rail	Car	
Pred.	Walk	427	129	312	49.19%
	Rail	83	1160	138	84.00%
	Car	67	188	2229	89.73%
Class recall		74.00%	78.54%	83.20%	

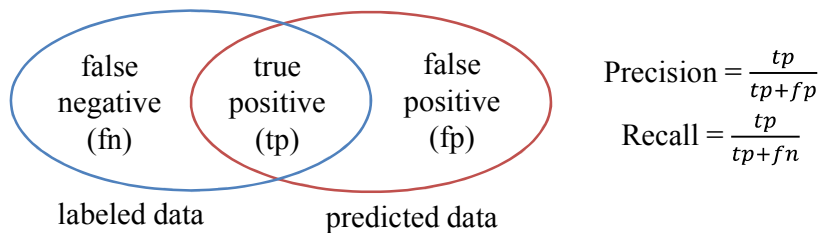


Figure 4-3 Basis for precision and recall

Furthermore, in order to analyze the characteristics of our decision tree, we calculated the actual breakdown of trips labeled as car. Figure 4-4 indicates that nearly three-fourths were actually car trips, but 16% were bicycle and 6% were bus trips.

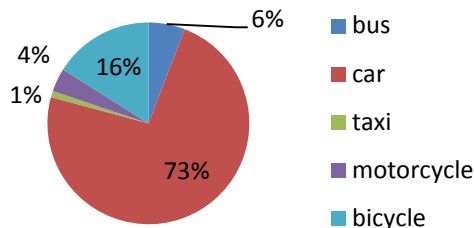


Figure 4-4 Actual breakdown of trips labeled as car

4.2 Discussion

Unlike past studies with GPS data, our results determined that railway proximity works best as the primary feature for identifying transportation modes. This feature separates rail from car trips, returning fairly accurate results for both modes. Most car trips, especially those that cover long distances, use main roads, which are usually located at some distance from railway lines. This is understandable as the majority of people accessing railway stations are pedestrians, and thus there is little need for main roads and railways to be located near each other.

It is interesting to note that different features were used to separate walk trips from rail trips and from car trips. One explanation may be that trip distance is a more definite feature in identifying rail trips, because it is highly unlikely for the trip to be shorter than the distance between two railway stations. In contrast, car trips may be used for relatively short distances, especially if they are actually bicycle trips.

The accuracy table indicates that precision for walk is particularly low, and the decision tree shows that most of these errors occur when separating walk from car trips, and that the trips labeled walk here contain the most mixed results. We assume two reasons why this might occur.

First, as a classification feature, average speed seems less reliable than trip distance. Distance is based on origin and destination clusters, which are calculated using multiple stay points, which in turn have been calculated from multiple GPS points. On the other hand, speed is calculated directly from GPS log coordinates. Therefore it is more likely to be distorted by inaccurate GPS logs affected by urban environments.

Second, let us suppose that average speed values have been calculated correctly. In this case, our labeling method may have created confusion. One example is that trips were labeled with the main transportation mode. As a result, trips with multiple modes including walk (for example, walk, car (such as bus), walk) may include low speed values and thus be predicted as walk trips, but will be labeled as either car or rail trips. Another example is that 16% of trips labeled as car were bicycle trips. Such trips, which may include situations where users get off and push their bicycles may, again, include low speed values and be predicted as walk trips, but be labeled as car trips.

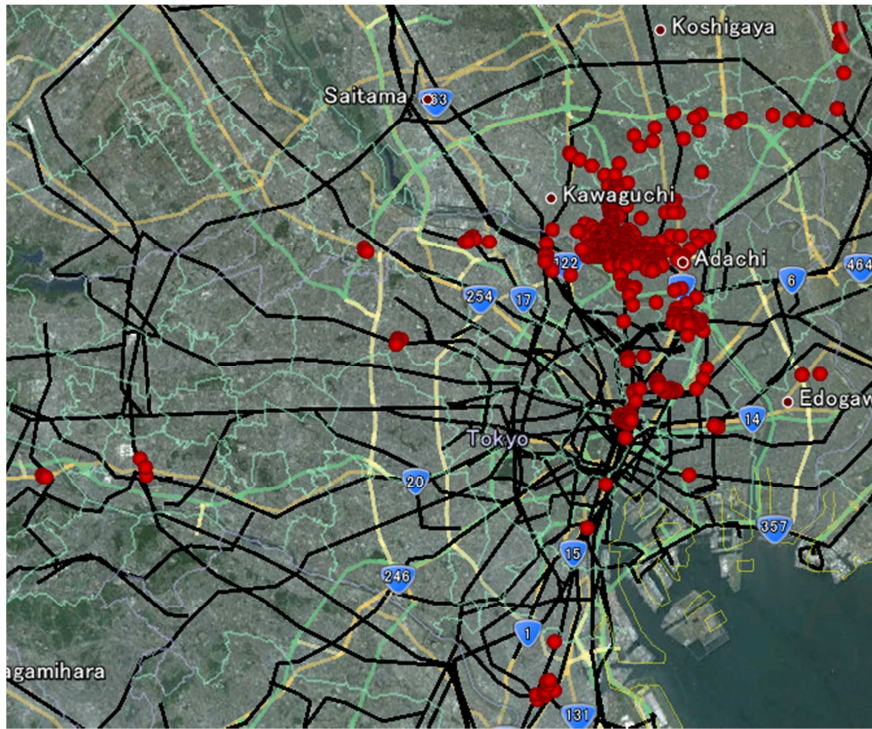
4.3 Results of test data

The features and thresholds in the aforementioned decision tree are used to identify transportation modes of distinct trips in our main dataset. Note that trips without trip data, meaning that they consist only of trips categorized as type C, are not identified and are labeled as “unclassified”.

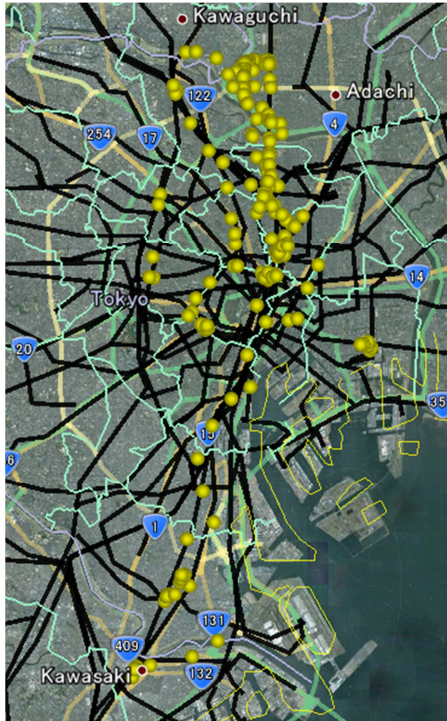
The results of this identification process on one user, mentioned in the previous chapter (3) as “A”, are as shown in Table 4-2. We identified the 182 distinct trips for this user, and multiplied each distinct trip with its number of individual trips to calculate the total number of trips by transportation mode. In addition, the ratio of modes for all trips returns what we define as the modal share of this individual. All trip data has been visualized according to transportation mode in Figure 4-5.

Table 4-2 Example of mode detection results

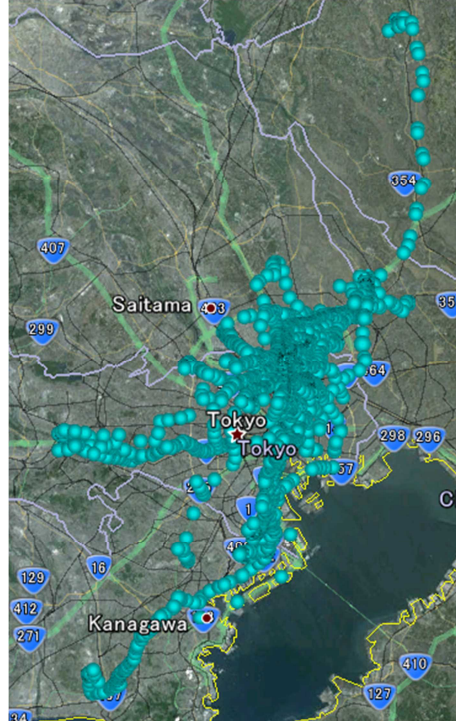
	Walk	Rail	Car	Unclassified	Total
Distinct trips	47	9	123	3	182
Individual trips	348	15	388	8	759
Modal share	45.80%	2.00%	51.10%	1.10%	100.00%



(a) Walk



(b) Rail



(c) Car

Figure 4-5 Trip data visualized according to transportation mode for “A”

5. FINAL RESULTS

5.1 Processing results

Finally, we ran the aforementioned process on our entire dataset for the Tokyo Metropolitan Area. From the 221,100 users, we extracted 26,837,942 distinct trips, or approximately 121.38 per user. Each distinct trip included the following data: a user ID number to distinguish individuals, origin coordinates (longitude, latitude), destination coordinates (longitude, latitude), the number of individual trips, and the identified transportation mode.

In order to validate our results, trips are grouped in the same way as our Person Trip (PT) Survey dataset. For each distinct trip, origin and destination coordinates are converted to the polygon zone codes they are located in. All distinct trips with the same pair of origin and destination zones are combined together, to form a single “grouped trip”.

Information about extracted trips for both datasets is organized in Table 5-1. Note that trip count comparisons cannot be made directly, due to differences in time frame and the number of represented individuals. For PT data, we extracted about 2.20 individual trips per person, for one day. For ZDC data, we extracted an average number of 402.89 individual trips per person. However, since many users only used this service for a short period of time, providing us with only a few months’ worth of data instead of a full year, it is not possible to estimate the average number of trips per day.

Table 5-1 Total numbers of extracted trips

Dataset		PT data		GPS data	
Time frame		1 day		1 day - 1 year	
Represented individuals		approx. 36 million		221,100	
Total trips	Grouped	116,667		309,979	
	Individual	79,038,534		89,077,507	
Individual trips, by transportation mode	Rail	23,984,945	30.3%	14,463,731	16.2%
	Walk	17,479,400	22.1%	6,576,126	7.4%
	Car	37,574,189	47.5%	67,637,798	75.9%
	Unclassified	0	0.0%	399,852	0.4%

Although our GPS data was collected from residents of the Tokyo Metropolitan Area, their logs include trips to or from outside of this area as well. Therefore, we limit trips from both datasets to those that occurred within our target area, which we define in this study as the following four prefectures: Tokyo, Kanagawa, Chiba, and Saitama. These datasets, the details of which are organized in Table 5-2, are used as the main dataset in our subsequent correlation analysis.

Table 5-2 Total numbers of extracted trips within target area

Dataset		PT data		GPS data	
Total trips	Grouped	111,591		271,454	
	Individual	75,587,788		82,835,555	
Individual trips, by transportation mode	Rail	23,641,874	31.3%	13,944,005	16.8%
	Walk	17,059,356	22.6%	6,448,940	7.8%
	Car	34,886,558	46.2%	62,105,171	75.0%
	Unclassified	0	0.0%	337,439	0.4%

From PT data, 95.6% of distinct trips and 95.6% of individual trips were within our target area. For GPS data, 87.6% of distinct trips and 93.0% of individual trips were extracted, which indicates that trips to or from outside of this area were occasional. Figure 5-1 draws comparisons between the numbers of all trips and trips within our target area, according to transportation mode. For both datasets, the majority of trips to or from outside this area were identified as car. This is logical as railway networks are sparse in areas outside of the Tokyo Metropolitan Area.

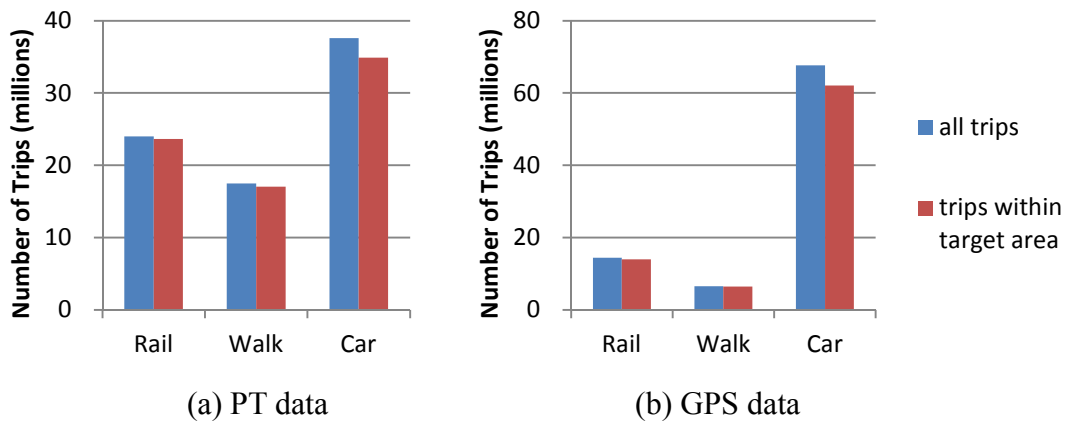


Figure 5-1 Comparisons between all trips and trips within our target area

5.2 Correlation analysis

Having compared the overall number of trips, we then confirm if our identified transportation modes are attributed to the appropriate groups of trips from one zone to another. We conduct a correlation analysis between the two datasets by creating scatter plots, with PT data plotted on the x axis and our GPS data plotted on the y axis. Each group of trips from one zone to another is represented as a point, where the x coordinate is the number of trips in the PT dataset, and the y-coordinate the number in our GPS dataset.

One scatter plot is prepared to compare the total number of trips, regardless of transportation mode, and evaluate the accuracy of our trip extraction process. Three separate plots are prepared to compare the total number of trips for each of the identified transportation modes, and evaluate the accuracy of our trip identification process.

5.2.1 *Raw values for all trips*

First, we extract all grouped trips that were found in both datasets, which we found a total of 110,640, as shown in Table 5-3. This value is 99.15% of the PT dataset and 40.76% of ZDC dataset. However, these distinct trips contained 99.9% of individual trips from PT data and 96.4% of individual trips from our GPS dataset. Note that at this point, we only draw ratio comparisons between raw values from both datasets.

Scatter plots are as shown in Figure 5-2. For the total number of trips from one zone to another, the coefficient of determination 0.69 suggests a fairly strong correlation, confirming that the trips were extracted fairly accurately. From the plot we can observe that certain grouped trips are overestimated in our GPS data, which will be discussed later in this chapter.

Of the three transportation modes, results for car show the strongest correlation and for walk the weakest. For walk in particular, we can observe that many grouped trips with a high value in PT data are not at all identified in GPS data, or have a value of close to zero. This tendency, too, will be discussed later in this chapter.

Table 5-3 Raw values of all trips for correlation purposes

Dataset		PT		GPS	
Total trips	Grouped	110,640		110,640	
	Individual	75,525,194		79,825,493	
Individual trips, by transportation mode	Rail	23,611,393	31.3%	13,173,186	16.5%
	Walk	17,058,553	22.6%	6,448,655	8.1%
	Car	34,855,248	46.2%	59,869,992	75.0%
	Unclassified	0	0.0%	333,660	0.4%

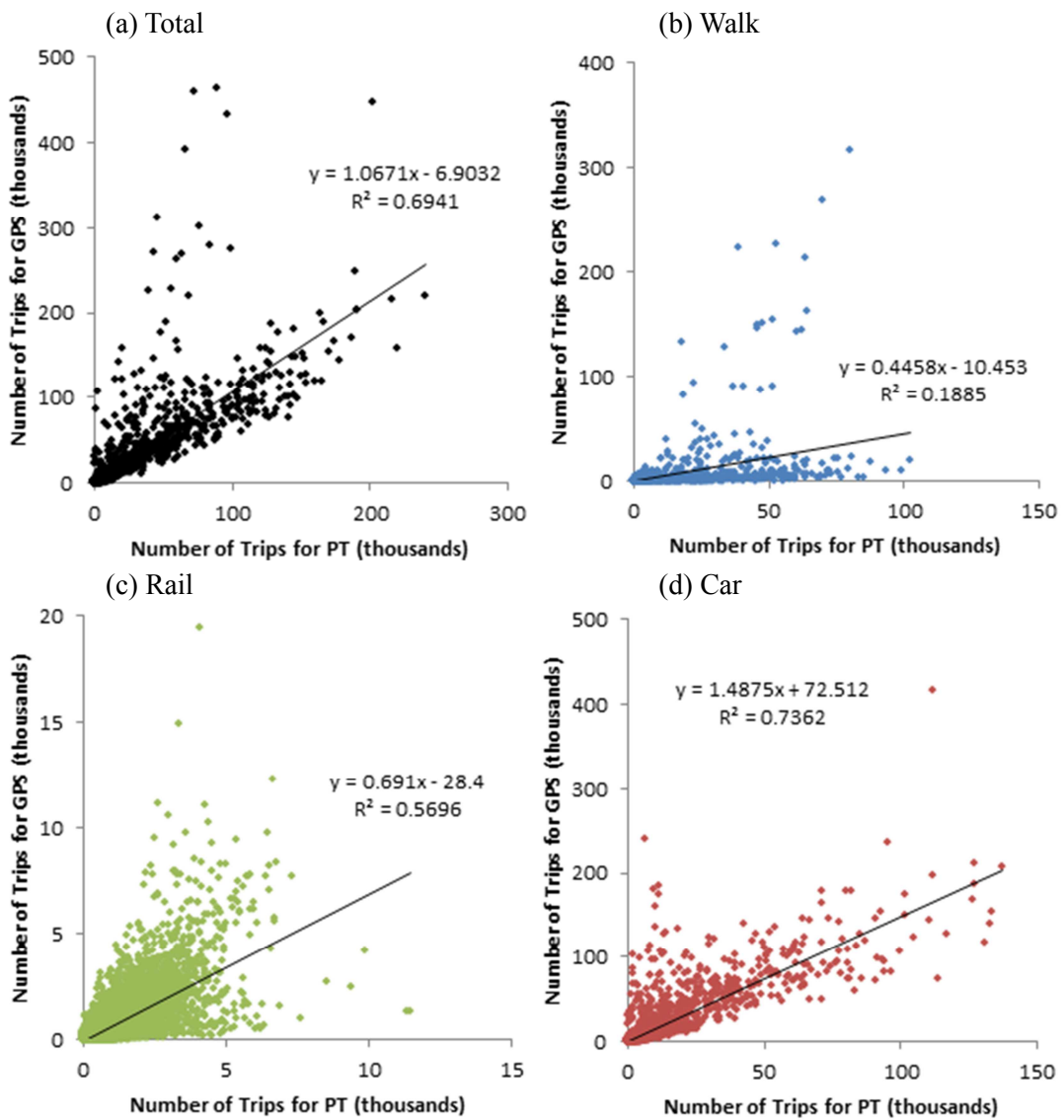


Figure 5-2 Scatter plots of raw values for all trips

5.2.2 Weighted values for all trips

Next, we recalculate trips to more accurately represent the actual population of the Tokyo Metropolitan Area. To do so, we multiplied the number of individual trips for each user by a specific weight W_m that was calculated for each mesh code m .

$$W_m = P_m / U_m$$

P_m : number of people who live in in mesh grid m

U_m : number of ZDC users who live in mesh grid m

where national census data was used for mesh-based population data and home locations of users were identified by Witayangkurn et al²³. The correlation between these two datasets is as shown in Figure 5-3.

Results of this adjustment are shown in Table 5-4. Since the above weight is assigned to users regardless of the duration of their GPS logs, the total number of individual trips will not be representative of a full year's worth of data, and thus is not 365 times the number of PT data. The shares of identified transportation modes change slightly, but not significantly.

On the other hand, the scatter plots in Figure 5-4 indicate that the coefficients of determination increase in all four cases. Therefore, we can confirm that using weighted values instead of raw values helps, to some extent, to correct any bias in our GPS dataset.

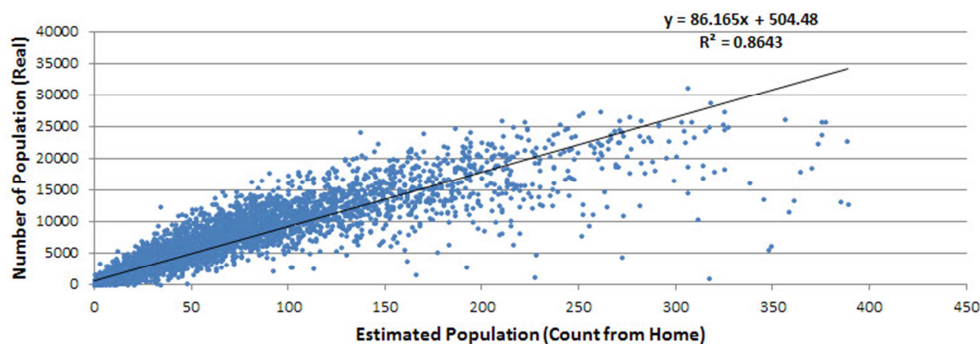


Figure 5-3 Correlation between ZDC user population and national census population

²³ Ibid., 17.

Table 5-4 Weighted values of all trips for correlation purposes

Dataset		PT		GPS	
Total trips	Grouped	110,640		110,640	
	Individual	75,525,194		8,209,160,983	
Individual trips, by transportation mode	Rail	23,611,393	31.3%	1,296,748,933	15.8%
	Walk	17,058,553	22.6%	618,362,774	7.5%
	Car	34,855,248	46.2%	6,259,244,321	76.2%
	Unclassified	0	0.0%	34,804,955	0.4%

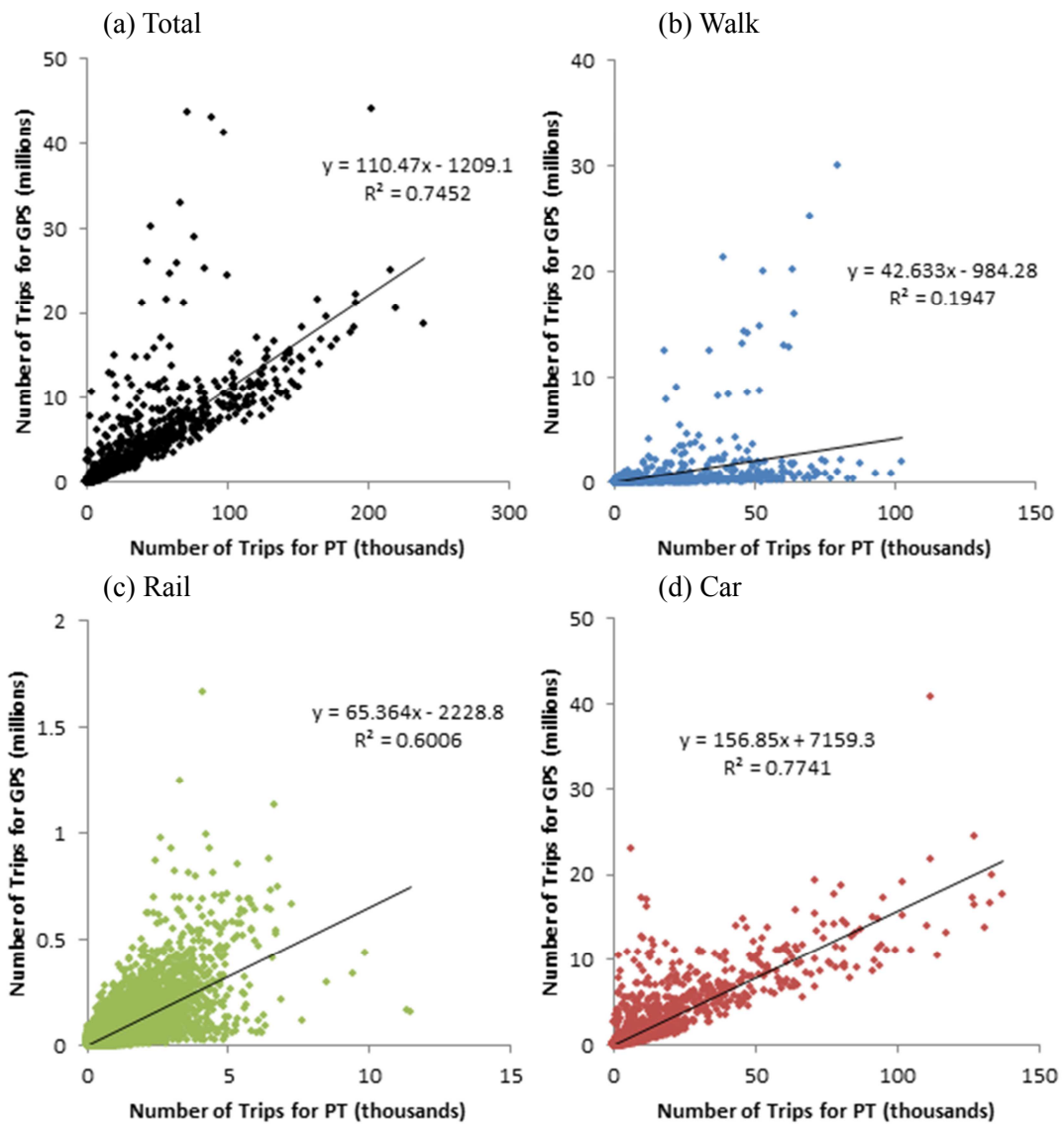


Figure 5-4 Scatter plots of weighted values for all trips

5.2.3 Weighted values for trips between different zones

In the scatter plots we can observe there to be many outliers, most of which had the same zone code as both their origin and their destination. These trips may be of short distances that took place within a single zone, or in some cases, trips where the origin cluster of stay points was the same as the destination cluster, or trip type B. Therefore, we prepared datasets where these trips were removed.

Results are as shown in Table 5-5 and the differences in data size are shown in Figure 5-5. The number of grouped trips has decreased by 547, and as there are 601 zones in the Tokyo Metropolitan Area, we can determine that 91.0% of zones include such trips. Since this process involved eliminating trips with shorter distances, walk trips decreased significantly during this process; for PT data to 16.2% and for GPS data to 11.3% of all trips. However, numbers for car trips decreased as well; for PT data to 54.0%, and for GPS data to 48.6%. Few rail trips were removed.

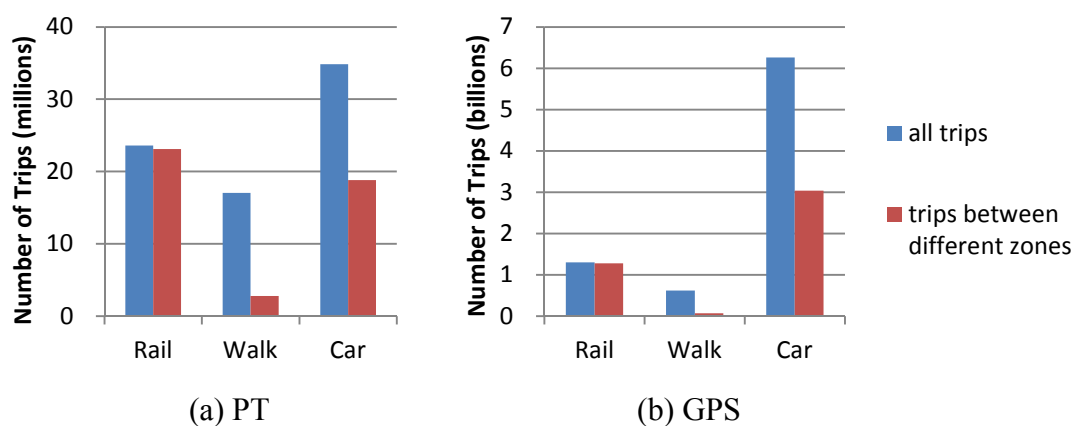


Figure 5-5 Comparisons between all trips and trips between different zones

As a result, the ratio of rail trips almost doubled, and the ratio of car and walk trips dropped significantly in both datasets. From this result, we can assume that nearly half of all car trips were used for relatively short distances, most of which may be bicycle trips. Our scatter plots, as shown in Figure 5-6, indicate that correlation becomes stronger when trips within the same zone are removed, especially for the numbers of trips identified as car and walk. We can assume that trips with longer distances are more likely to be extracted and/or identified accurately.

Table 5-5 Weighted values of trips between different zones for correlation purposes

Dataset		PT		GPS	
Total trips	Grouped	110,093		110,093	
	Individual	44,701,971		4,402,016,097	
Individual trips, by transportation mode	Rail	23,104,503	51.7%	1,279,393,281	29.1%
	Walk	2,764,378	6.2%	69,592,411	1.6%
	Car	18,833,090	42.1%	3,039,336,127	69.0%
	Unclassified	0	0.0%	13,694,278	0.3%

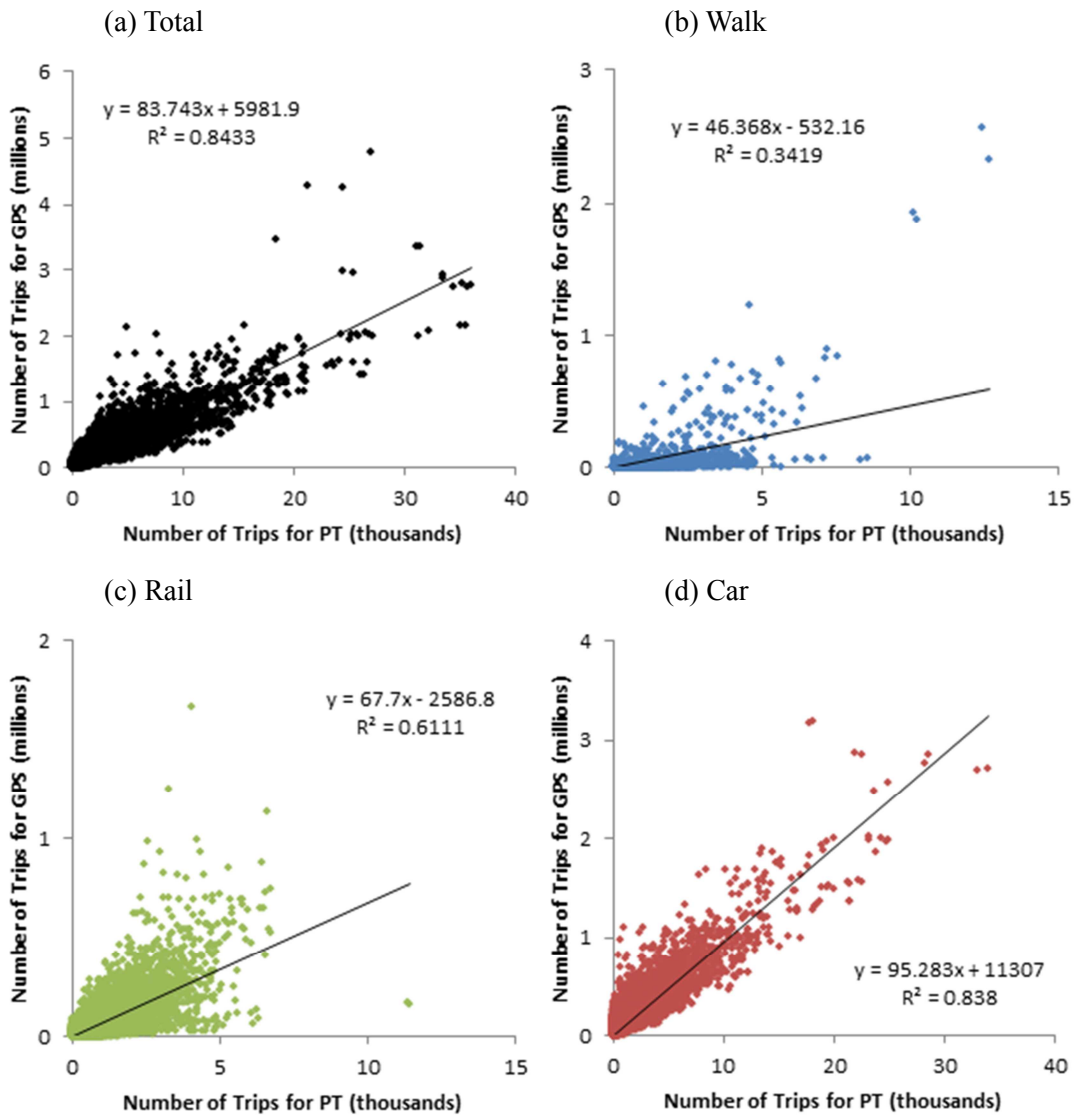


Figure 5-6 Scatter plots of weighted values for trips between different zones

5.3 Discussion

Throughout our comparison of the two datasets, the total number of trips extracted indicated strong correlation, and car trips indicate about the same coefficient of determination (in fact, for raw values the value for car trips is actually higher). Although not as high, rail trips indicate fairly strong correlation as well. In general, walk trips indicate weak correlation. We assume the following explanations for these results.

First of all, when tested on ground truth data our algorithm returned the highest recall rate for car trips, and the lowest rate for walk trips. Our results on the entire dataset may simply reflect this tendency.

We contemplated further explanations for this tendency in both datasets, and determined that it may be explained by a lack of, or inaccurate, GPS data. As described earlier, logs are recorded at a minimum of 5 minutes, and either do not record, or record a slightly different location, if reception quality is low. Walk trips are generally for shorter periods of time, and may not contain enough data to be identified; for example, trips with no data are labeled as “unclassified”. Furthermore, both walk and rail trips usually take place across relatively urbanized areas. Pedestrians will often use narrow streets, possibly lined with tall buildings, and subway users will travel underground. In contrast, most car trips will take place on wide roads, usually located some distance away from railway stations and the buildings that surround it. Therefore, car trips will record more and more accurate GPS data.

However, we noted a significant difference between the results of our ground truth data and our main dataset. In the former dataset there was an overestimation of walk trips, but in the latter dataset there is rather an overestimation of car trips.

One explanation for this may be that PT data is a survey taken during one weekday, whereas both of our GPS datasets includes trips on weekends. Many people living in the Tokyo Metropolitan Area may use the railway to commute to central Tokyo, where their workplace may be located, but use cars to travel to suburban areas on their days off. (Though there may be some sample bias, Figure 5-7 compares the actual breakdowns of trips labeled as car for both datasets, and indicates that is a higher percentage of actual car trips in the ground truth data.)

Also, by examining scatter plots for walk we realized that compared to other transportation modes, there were grouped trips for PT data that did not exist at all for our GPS dataset. When we identified some of these trips, we found that they included trips between zones that were located far apart from each other. Figure 5-8 shows one such example, from zone code 220, in central Tokyo, to zone code 4416, in Chiba Prefecture. This is a distance of approximately 40 kilometers, and while GPS data returned zero walk trips, PT data indicated 105. We assume this to be the result of an error that occurred somewhere during the process of collecting and aggregating Person Trip Survey data.

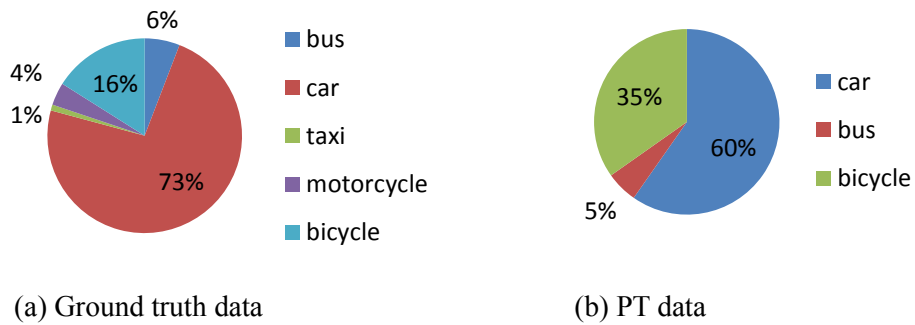


Figure 5-7 Comparison of actual breakdowns of trips labeled as car

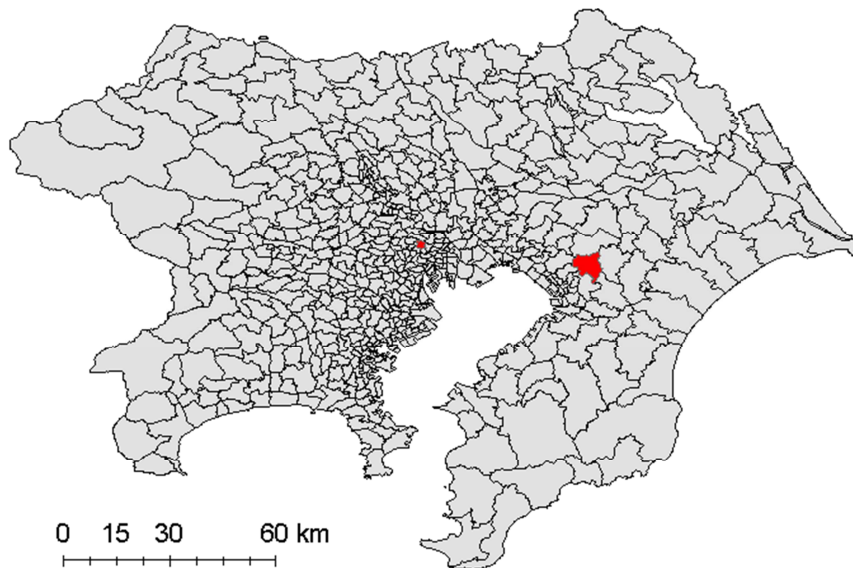


Figure 5-8 Example of a grouped trip that included walk trips for PT data

6. MOBILITY ANALYSIS

6.1 Organization of data

Finally, our identified trips are used to analyze the long-term mobility of individuals living in the Tokyo Metropolitan Area. Survey-based data, such as the Person Trip Survey information used in the previous chapter, can be acquired only for a limited period of time. GPS data, on the other hand, can be collected easily over a longer duration. Therefore, we can estimate how often each transportation mode is being used by each individual, and draw comparisons between people living in different areas.

First, we group all extracted and identified trips by user. Then, the long-term modal share of each user is calculated by dividing the total number of trips for each mode with the total number of trips for that person, as in the following equation:

$$S_m = \frac{T_m}{T_{rail} + T_{walk} + T_{car} + T_{unclassified}}$$

where T_m indicates the total number of individual trips for mode m . After determining the long-term modal share for each user (S_{rail} , S_{walk} , S_{car}), we group users according to the polygon zone where their home is located. Home locations are the same areas used in the magnification process in the previous section (5.2). Finally, we calculate the average modal shares for each group of users and map the results by polygon zone.

Note here that certain adjustments were made to our dataset. Our study and thus algorithm focuses on identifying transportation modes within our target area, the four main prefectures of the Tokyo Metropolitan Area. Therefore, we cannot be certain of trips to or from outside this area. For our validation process, we removed such trips, but for calculation of modal share, removal would reduce trip count for each individual and distort ratios. Instead, we re-label all trips to or from outside of this area as “unclassified”. Table 6-1 shows the numbers of trips within and outside of our target area, and the converted dataset that we used for the calculation of modal shares.

Table 6-1 Organization of number of re-labeled trips

Dataset	Trips within area		Trips outside of area		Converted dataset	
Total Trips	82,835,555		6,241,952		89,077,507	
Rail	13,944,005	16.8%	519,726	8.3%	13,944,005	15.7%
Walk	6,448,940	7.8%	127,186	2.0%	6,448,940	7.2%
Car	62,105,171	75.0%	5,532,627	88.6%	62,105,171	69.7%
Unclassified	337,439	0.4%	62,413	1.0%	6,579,391	7.4%

All of the following maps classify polygon zones into ten classes using the Jenks optimization method. This data classification method is a common one that aims to reduce the variance within classes and maximize the variance between classes. To focus on our target area, when we map our results we remove all users whose homes were located outside of this area. This left us with 219,364 of the 221,085 users in our dataset (99.2%), who resided in 548 of the total of 601 polygon zones (91.2%).

Population data for each polygon zone was provided by the Tokyo Metropolitan Region Transportation Planning Commission from their Person Trip Survey results in 2008²⁴.

²⁴ Ibid., 11

6.2 Trip results

6.2.1 Distribution of users

The distribution of users is shown in Figure 6-1. The first map plots the actual number of users, but as polygon sizes vary, the second map plots this value as a percentage of the actual residents in each zone. Each polygon is represented by a roughly equal percentage of the population, mostly below one percent. Lowest values are seen in the westernmost areas of Saitama, Tokyo, and Kanagawa, as well as the southernmost areas of Chiba, while certain polygons in central Tokyo are overrepresented.

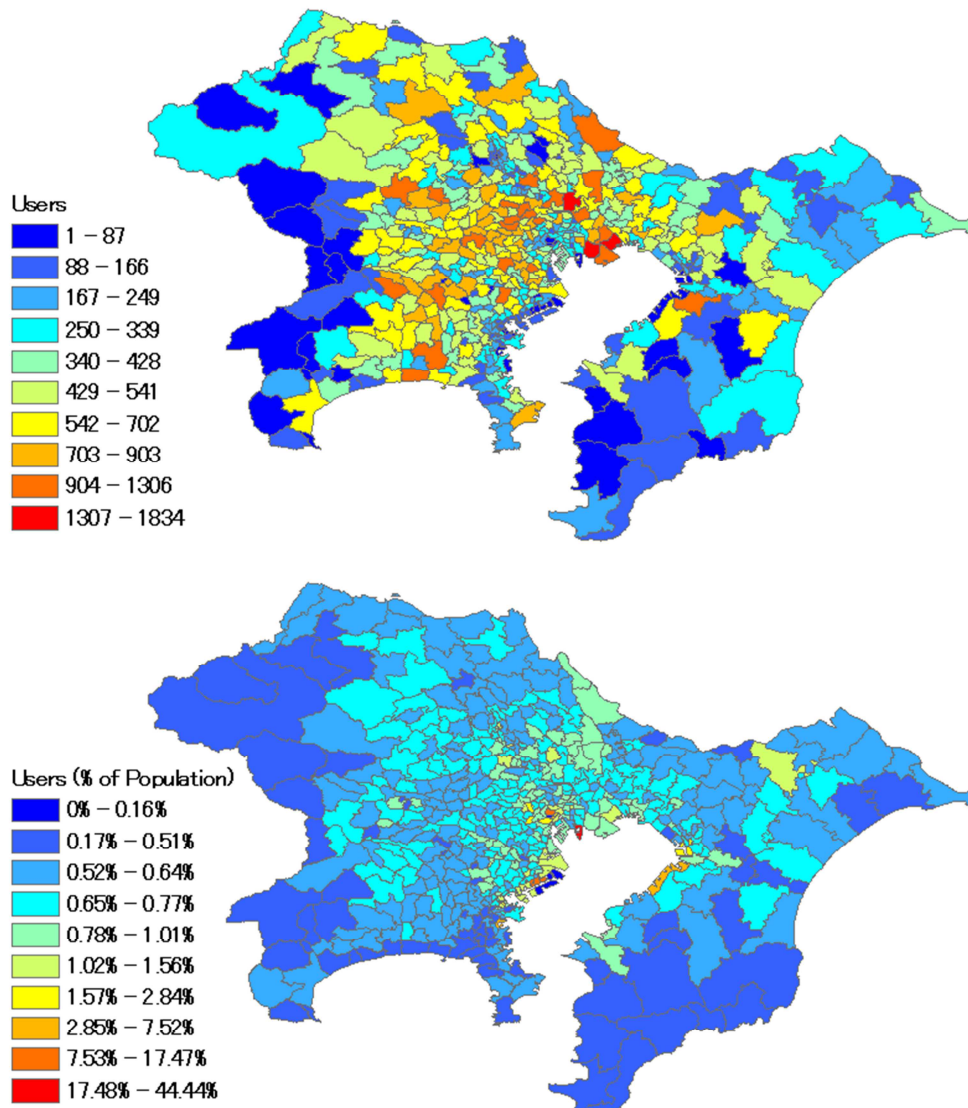


Figure 6-1 Number of users and percentage of actual population for each zone

6.2.2 Number of trips

Next, we examine the average number of trips for the residents of each polygon zone. The numbers of individual trips, as shown in the top map of Figure 6-2, range between 259 and 521, with lower values found in the outskirts of the Tokyo Metropolitan Area, such as southern Chiba and western Saitama. These values may be affected by the age of residents in each zone, as the proportion of elderly people is said to be higher in these areas. The numbers of distinct trips, on the other hand, range between 71 and 162, and indicate the variety of trips for residents of each zone.

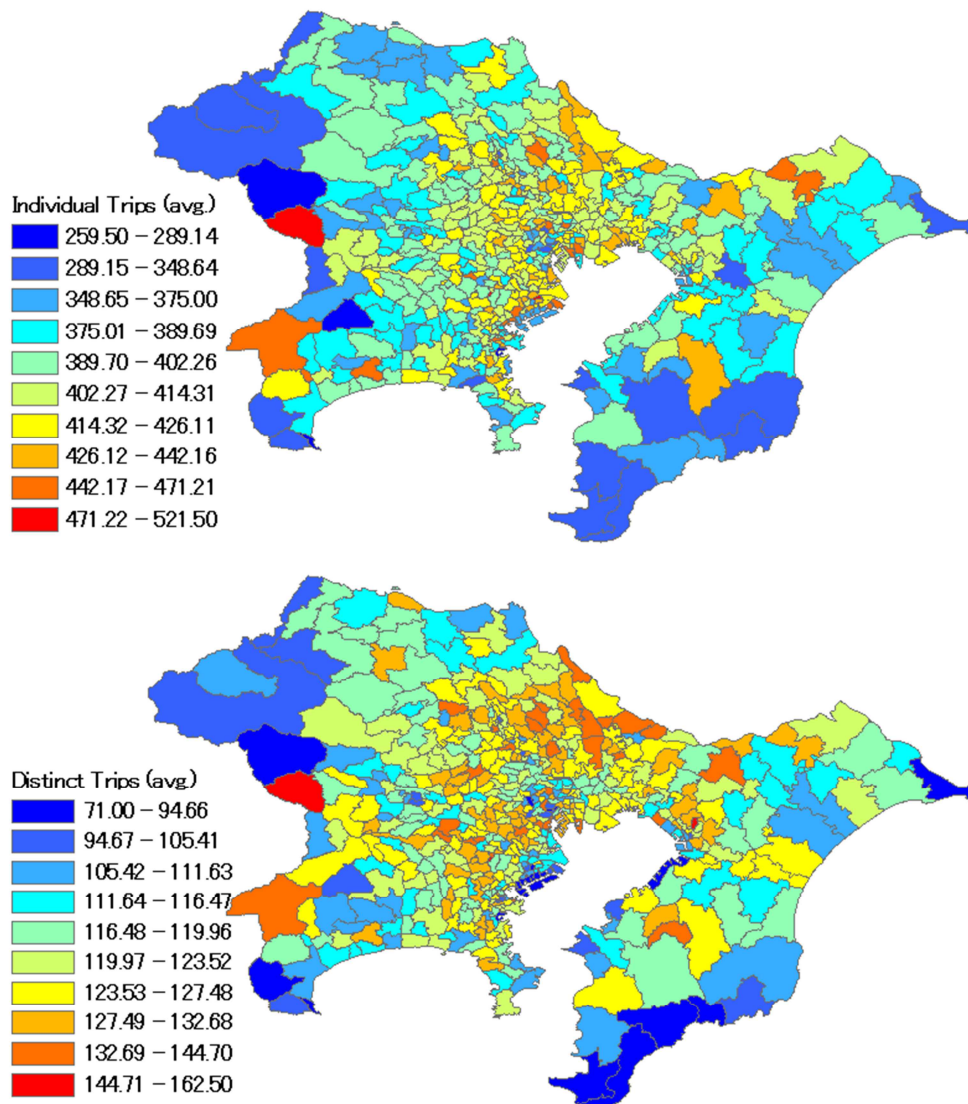


Figure 6-2 Average number of individual trips and distinct trips

6.2.3 Distance of trips

We also add together the distance of all trips, for each individual, and calculate the average value for each polygon zone, as shown in Figure 6-3. Note that areas which in the previous section had a high number of individual and distinct trips, around the border between Saitama and Chiba, are low in value, and areas in northeastern Kanagawa are high. Therefore when we divide each distance value by the average number of individual trips, calculating the average distance per trip, we see this area in northeastern Kanagawa seem the highest, along with northern Chiba.

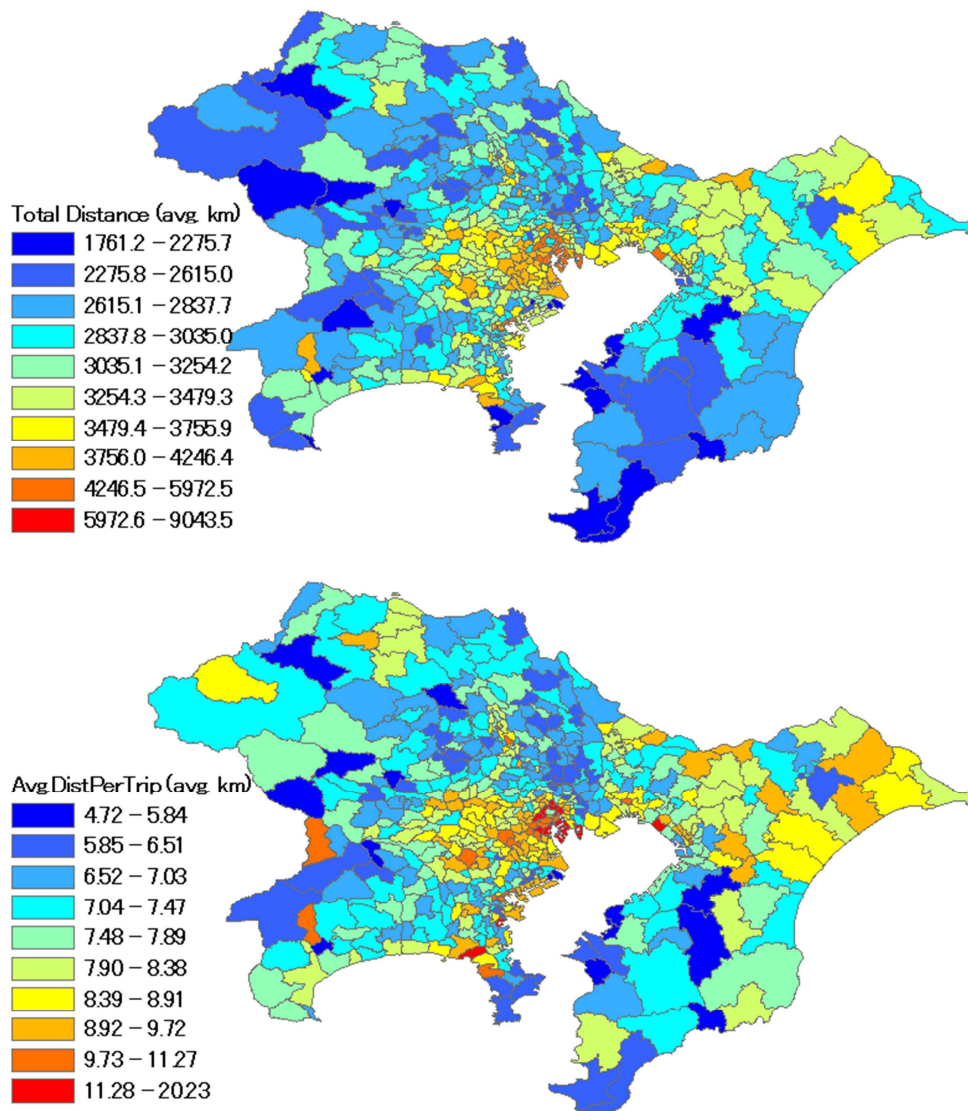


Figure 6-3 Average total distance and average distance per trip

6.3 Transportation mode results

6.3.1 Modal share

In Figure 6-5, we map the average ratio of trips for each transportation mode. As we had expected, values for rail are higher in central Tokyo, whereas values for car are lower. Zones with high rail percentages tend to have low car values, which is understandable as these two transportation modes make up the majority of our dataset.

To confirm our assumption that shares for rail and car are influenced by railway network density, we overlay such data on the map representing rail percentages (Figure 6-4). It is clear that zones with railway networks, especially networks that extend outwards from central Tokyo, are areas where users depend more heavily on rail as a transportation mode.

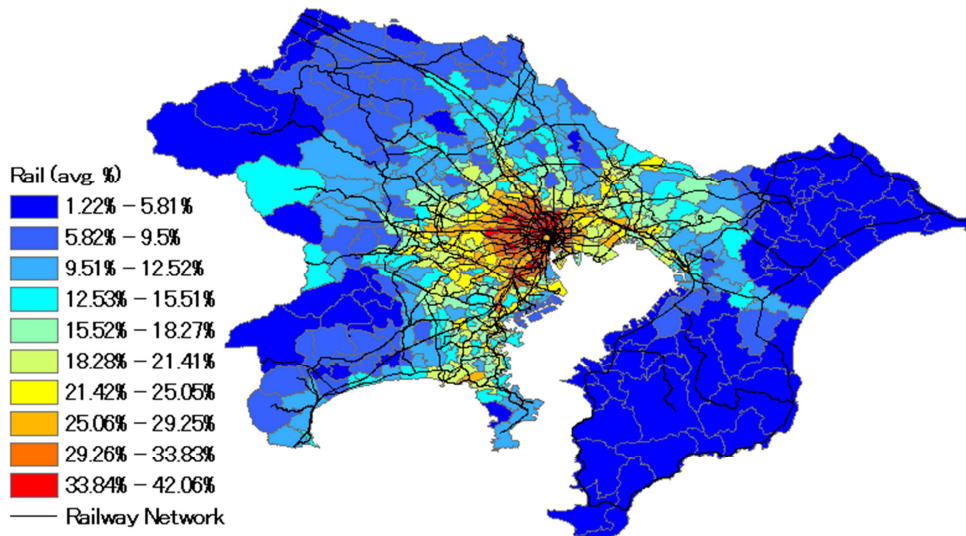


Figure 6-4 Average percentage of trips for rail, overlaid with railway network data

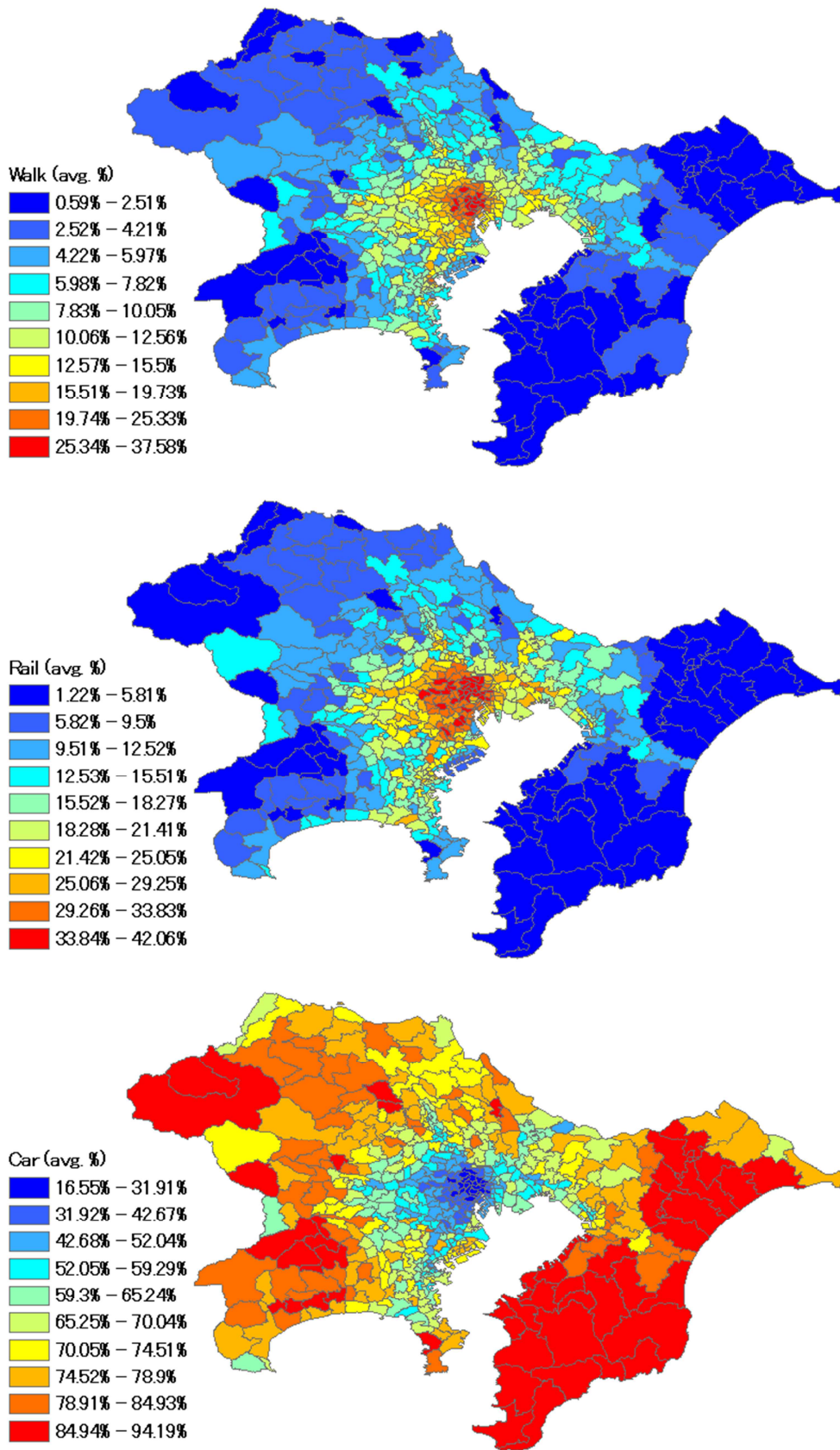


Figure 6-5 Average percentage of trips for each transportation mode

6.3.2 Relation to railway station proximity

Furthermore, we assume that station proximity is a large factor in determining mode choices. We estimate the distance between home locations and nearest stations for each user, then calculate the average value for each zone and map results in Figure 6-6. As we expected, those living in zones with dense railway networks had shorter distances.

We then compared these distances with modal shares, by creating scatter plots for each transportation mode with distance values as the x value and percentages of modal share as the y value. In the graphs on the left side of Figure 6-7, each point represents one user; on the right side, each point represents one polygon zone. User-based comparisons did not show a significant correlation, as coefficients of determination values were very low (all below 0.1) due to those who live near stations yet have a relatively low percentage of trips that use rail. However, we can observe from the scatter plots that users who live some distance away from stations will not use rail or walk as often. Zone-based comparisons, which used average values of residents for each polygon, showed more significant correlation. For (b) walk and (d) rail there is a slight negative correlation, showing that individuals are more likely to depend on these two modes of transportation if they live closer to railway stations. In addition, there is a slight positive correlation for (f) car, showing that users whose homes are located further from railway stations are more likely to rely on car for their trips.

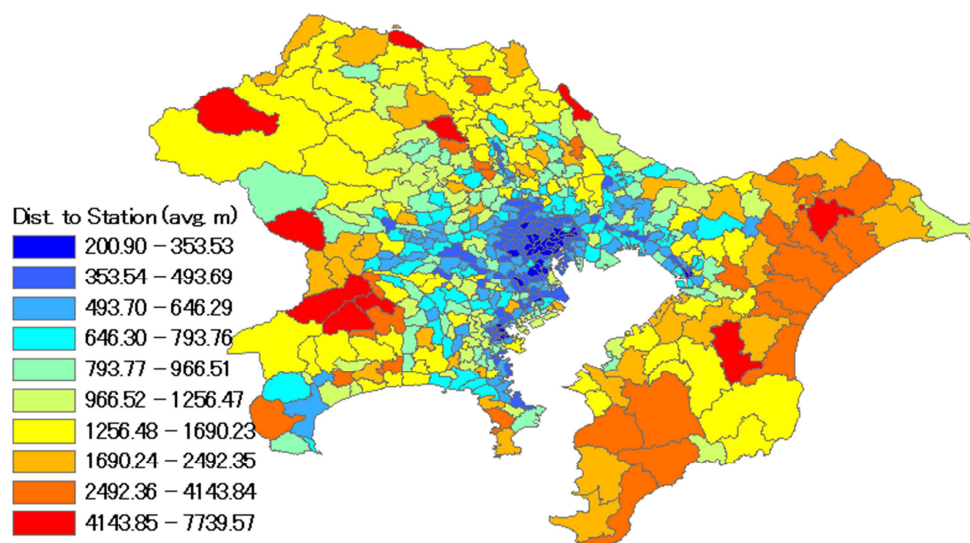


Figure 6-6 Average distance between home and nearest station

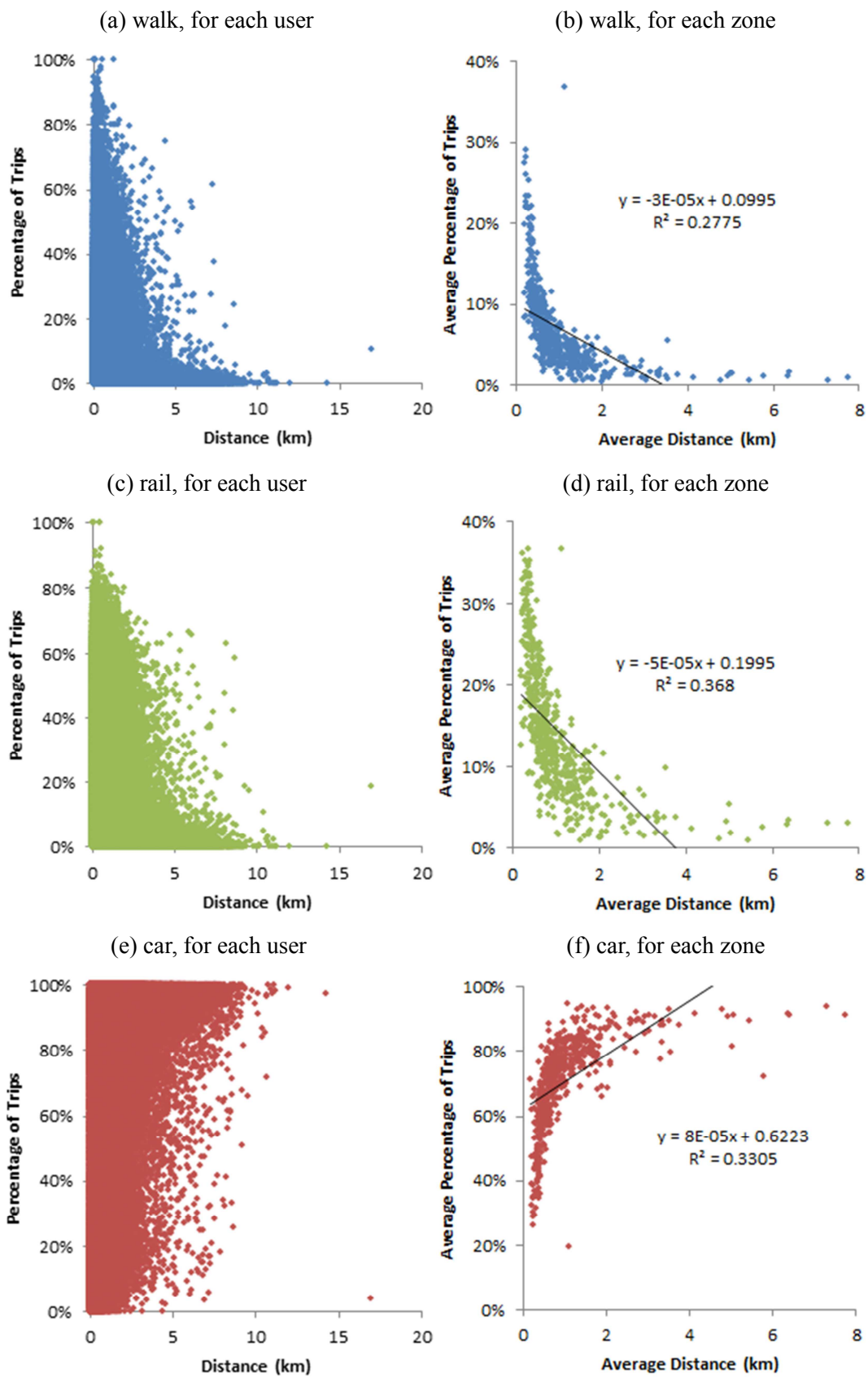


Figure 6-7 Scatter plots showing relation between modal share and proximity to stations

6.3.3 *Relation to trip distances*

Finally, we noted some similarities in the distribution of areas where rail dependency was high and those where total trip distances were relatively high (6.2.3). Therefore we conducted a correlation analysis similar to the one comparing station proximity (6.3.2), using scatter plots with percentages of modal share as the y value. For the x value Figure 6-8 uses the total distances traveled, whereas Figure 6-9 uses the average distance per trip, as calculated earlier. In both figures, scatter plots on the left are user-based comparisons and those on the right are zone-based.

Similar to the analysis regarding station proximity, user-based comparisons indicated very low coefficients of determination, while zone-based results showed weak correlation. Albeit only slightly, walk and rail percentages indicated positive correlation, meaning that those who traveled longer distances were more likely to use these transportation modes. On the other hand, users who were more dependent on car tended to travel shorter distances. These analyses are another example of how long-term GPS data can be used to compare the mobility of residents in different areas.

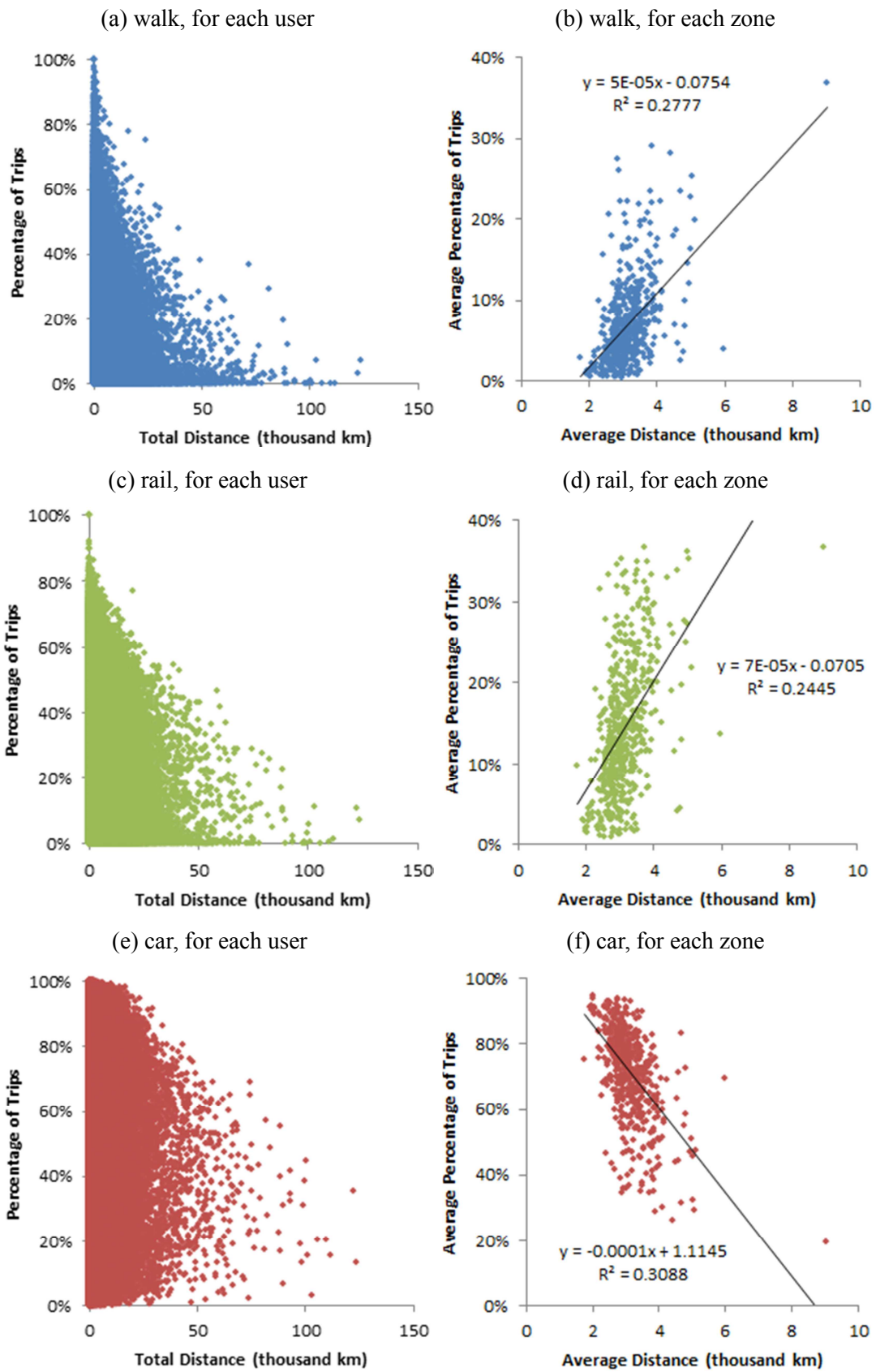
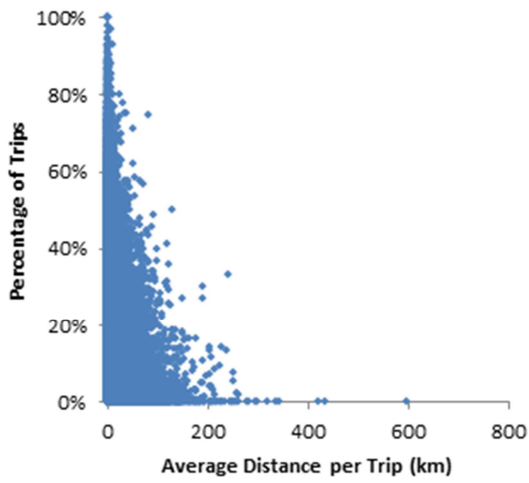
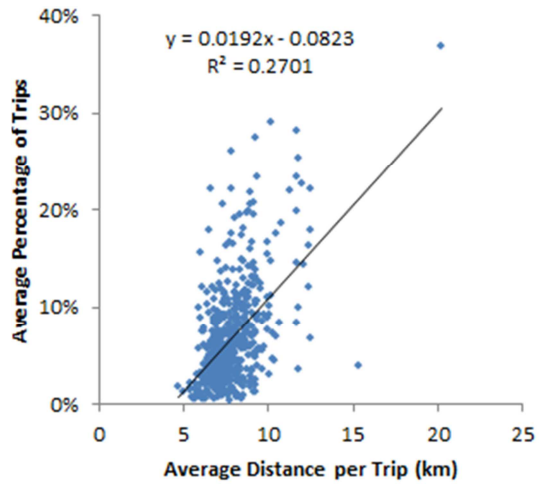


Figure 6-8 Scatter plots showing relation between modal share and total distance traveled

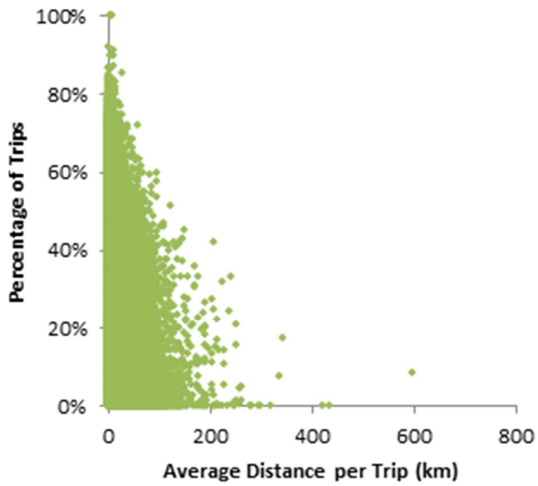
(a) walk, for each user



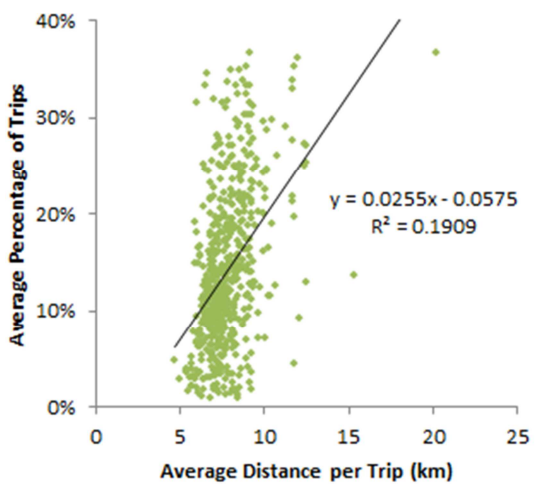
(b) walk, for each zone



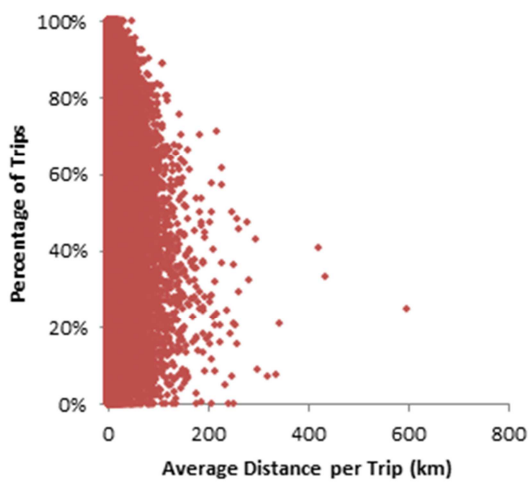
(c) rail, for each user



(d) rail, for each zone



(e) car, for each user



(f) car, for each zone

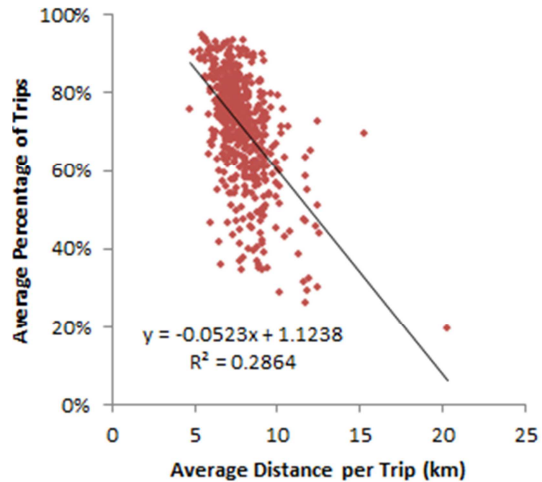


Figure 6-9 Scatter plots showing relation between modal share and average distance per trip

7. CONCLUSION

7.1 In summary

In this study we used a large dataset of GPS logs collected from mobile phones, and used it to extract trips and identify the main transportation modes as walk, rail, or car. As such data was sparse, we assumed that individuals traveling from one location to another multiple times used the same route each time and grouped GPS logs accordingly. By increasing data for each trip, we were able to use proximity to railway networks as a primary parameter for identifying transportation modes.

We validated the results of our trip extraction and identification process by using single-day data, Person Trip Survey data, to compare the numbers of trips from one area zone to another. Finally, we used the identified trips to conduct a mobility analysis of the Tokyo Metropolitan Area, where we compared how home locations affect the frequency of using specific transportation modes.

During the validation process, we observed some difficulty in identifying certain trips, especially those more likely to be affected by limitations of GPS data, low accuracy and frequency, such as trips between short distances and that take place in densely urbanized areas. At the same time, we noticed limitations with single-day travel survey data. In addition to the risk of human error, such information, while suitable for understanding commuting travel patterns, does not include travel behavior for weekends, when people move more freely.

Furthermore, our analysis of results succeeded in highlighting how geography, in particular accessibility to public transport networks, affects dependency on different transportation modes. This sort of information is helpful for urban planning purposes, especially for the maintenance and improvement of transport-related infrastructure. This aim, combined with the aforementioned limitations of survey data, help to emphasize the importance of collecting and analyzing long-term data collected from a reasonably large sample set, which is easy to accomplish with the use of GPS features in mobile phones.

7.2 Future works

In the future, our methodology may be applied to other regions besides the Tokyo Metropolitan Area. Parameters and thresholds may need to be reconsidered depending on the characteristics of each area, but we believe that Tokyo presents the most difficulty in that densely urbanized areas can cause inaccurate and missing GPS logs, a fact which seemed to be confirmed during our study.

Furthermore, our methodology can be expanded upon and improved. First of all, preparation of the GPS dataset can be refined, either by removing inaccurate logs or interpolating logs to fill gaps. Inaccurate logs may be detected by searching for unrealistic speeds or directions, and also by identifying base stations where such logs may be concentrated. Next, the trip extraction process can include segmentation of individual trips for identifying more specific transportation modes, such as the walk trips before and after a rail trip. Furthermore, we can increase the variety of modes to be identified. As our results indicate a large ratio of car trips, identification of more specific transportation modes, such as bus and bicycle trips, may be of particular interest.

Finally, our study results and mobility analysis— including the locations of users' homes— may help to identify the personal attributes of individual users, such as gender or age, or even further characteristics of the trips themselves, such as trip purpose. This would provide similar information to survey data in a more efficient manner.

8. REFERENCES

- Ashbrook, D., Startner, T. (2003). Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, v.7 n.5, p.275-286.
- Bohte, W., Kees, M. (2008). Deriving and validating trip destinations and modes for multiday GPS based travel surveys: An application in the Netherlands. Paper presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC
- Chung, E., Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Journal of Transportation Planning and Technology* 28 (5), 381–401.
- de Jong, R., Mensonides, W. (2003). Wearable GPS device as a data collection method for travel research. Working Paper, ITS-WP-03-02, Institute of Transport Studies, University of Sydney
- Gong, H., Chen, C., Bialostozky, E., and Lawson, C. (2011). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, doi:10.1016/j.compenvurbsys.2011.05.003
- Horanont, T., Witayangkurn, A., Sekimoto, Y., Shibasaki, R. (2013). Large-Scale Auto-GPS Analysis for Discerning Behavior Change during Crisis. *IEEE Intelligent Systems*, 11 Jan. 2013. IEEE computer Society Digital Library. IEEE Computer Society, <<http://doi.ieeecomputersociety.org/10.1109/MIS.2013.3>>
- Khetarpaul, S., Chauhan, R., Gupta, S. K., Subramaniam, L. V., Nambiar, U., Mining GPS data to determine interesting locations. (2011). Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011, p.1-6, March 28-28, 2011, Hyderabad, India
- Lerin, P.M., Yamamoto, D., Takahashi, N. (2010). Inferring and Focusing Areas of Interest from GPS Traces. In: Kim, K.-S. (ed.) W2GIS 2011. LNCS, vol. 6574, pp. 176–187. Springer, Heidelberg
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J. (eds), *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. University of California Press.
- Montoliu, R., Blom, J., and Gatica-Perez, D.. (2012). Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, 1-29.
- Schuessler, N., Axhausen, K.W. (2009). Processing GPS Raw Data without Additional

- Information. In Paper presented at the 88th annual meeting of the transportation research board, Washington, DC
- Stopher, P., Clifford, E., Zhang, J., and FitzGerald, C. (2007). Deducing Mode and Purpose from GPS Data, paper presented to the Transportation Planning Applications Conference of the Transportation Research Board, Daytona Beach, Florida, May.
- Stopher, P., Qingjian, J., FitzGerald, C. (2005). Processing GPS Data from Travel Surveys. In: 28th Australasian Transport Research Forum, Sydney, Australia.
- Witayangkurn, A., Horanont, T., Sekimoto, Y., and Shibasaki, R. (2010). Large Scale Mobility Analysis: Extracting Significant Places using Hadoop/Hive and Spatial Processing. Technical Report.
- Xu, Ch., Ji, M., Chen, W., Zhang, Z. (2010). Identifying Travel Mode from GPS Trajectories through Fuzzy Pattern Recognition. In Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery
- Zheng, Y., Zhang, L., Xie, X., Ma, W. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. In Proc. of the 18th Intl. Conf. on World Wide Web (Madrid Spain, 2009), ACM Press: 791-800.
- 前司敏昭, 堀口良太, 赤羽弘和, 小宮粹史: GPS 携帯端末による交通モード自動判定法の開発, 第4回 ITS シンポジウム 2005 論文集, 2005

RapidMiner, <http://www.rapidminer.com/>

ZENRIN DataCom CO., Ltd., <http://lab.its-mo.com/densitymap>

国土数値情報ダウンロードサービス, 国土交通省国土政策局, 2013年1月2日

<http://www.mlit.go.jp/kokudoseisaku/gis/index.html>

社団法人 電気通信事業者協会, 2013年1月24日

<http://www.tca.or.jp/database/index.html>

東京都市圏交通計画協議会, 2013年1月2日

<http://www.tokyo-pt.jp/>

ACKNOWLEDGEMENTS

I would like to thank the following people for their assistance in this study.

First, I would like to express my appreciation towards my advisors for guiding me during these past two years. Thank you to Professor Ryosuke Shibasaki and Associate Professor Yoshihide Sekimoto, for supporting me in choosing a topic that was of particular interest to me, for listening to my ideas about how to go forth with the research, and for providing me with valuable advice. I would like to give special thanks to Associate Professor Sekimoto for his close guidance. Thank you for providing me with useful references and data, as well as various opportunities for me to learn more about this field of study.

I would also like to extend my sincere appreciation to my secondary advisor, Associate Professor Shin'ichi Konomi, who helped me look at my study from an objective perspective and in particular, helped me in preparing the presentation of my results.

This study would not have been possible without the various types of data that we used. I am extremely grateful to ZENRIN DataCom, Ltd., who provided us with the massive dataset of GPS logs used in this study. I am also thankful to the Japanese mobile phone service provider who helped to provide the labeled data used as ground truth.

I especially owe much thanks to Teerayut Horanont and Apichon Witayangkurn, who conducted similar research on the same dataset. Thank you both for preparing the data, for advising me on processing methods, and for providing me with the technical assistance in handling the large dataset.

Also, I would like to thank Takehiro Kashiya, Hiroshi Kanasugi, and Toshikazu Nakamura, who helped provide me with ground truth data and validation data, and who assisted me greatly during the coding process, especially in reducing the processing speed of my program to handle such a massive dataset.

I am grateful to the other students who used the same data, Mayumi Hadano and Mariko Shibasaki, for discussing and sharing various ideas with me. Thank you also to students Yuki Okamoto and Yoshiki Ogawa for their support at Komaba lab. Finally, I would like to express my deepest gratitude towards everyone else in our laboratory, for helping me in various ways during these past two years.