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

学籍番号　　47-116721
氏　　名　　大野 夏海（Ono, Natsumi）
指導教員　　柴崎 亮介 教授

1. INTRODUCTION

In recent years, information regarding the flow of people is becoming increasingly important. Furthermore, the spread of mobile phones has made it possible to collect Global Positioning System (GPS) data in large scales (hundreds of thousands of people) for long durations (from several months to several years). While such data is already being used for purposes such as estimating population distribution, records of individuals are simply a sequence of points indicating when and where they traveled, but not how they traveled between these locations.

Studies identifying transportation modes from GPS data exist, but most are conducted using loggers (Zenji, 2005; Bohte, 2008; Gong, 2011). While logs are frequent, the scale of study is often limited (Table 1). In contrast, mobile phones provide sparse and inconsistent data (due to battery constraints), yet on a large scale.

Table 1　Comparison of related works

	intervals	participants	duration
Zenji	10 sec	N/A	1 day
Bohte	6 sec	1104	1 week
Gong	5 sec	63	5 days

Therefore, the objective of this study is to identify mobile phone-based GPS data, and use the results to analyze the long-term mobility of individuals in the Tokyo Metropolitan Area.

2. METHODOLOGY

Large-scale, long-term GPS data is used to extract and identify trips as rail, car, or walk. To gather enough information from such sparse data, we assume that individuals traveling from one location to another multiple times will use the same mode and group GPS logs accordingly. First we identify and cluster similar trip nodes, "stay points", and then we extract and group similar trips as "distinct trips". Ground truth data is used to create a classifier for identifying each distinct trip. Finally, we process our entire dataset, compare the numbers of individual trips for each mode with validation data, and then conduct a mobility analysis (Figure 1).

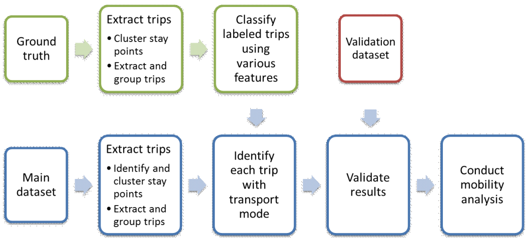
Ground truth
Extract trips
• Cluster stay points
• Extract and group trips
Classify labeled trips using various features
Validation dataset
Main dataset
Extract trips
• Identify and cluster stay points
• Extract and group trips
Identify each trip with transport mode
Validate results
Conduct mobility analysis

Figure 1　Overall process

2.1 *Main dataset*

GPS logs of mobile phone service users were collected from August 2010 to July 2011. Logs were recorded at a minimum of 5 minutes, but only when movement was detected and when reception was available. We use the data of roughly 221,100 individuals residing in Tokyo, Kanagawa, Chiba, and Saitama.

2.2 *Ground truth dataset*

In a separate study, 160 individuals used the GPS features in their mobile phones between November 28 and December 22, 2011. Trips were automatically extracted and labeled with the correct transportation mode by participants using an online application.

2.3 *Validation dataset*

We use aggregated results of single-day travel diaries, Person Trip Surveys for the Tokyo Metropolitan Area from 2008. About 2% of the residents were surveyed and weighted to represent the actual population. All trips from one area zone to another are added together by transport mode as "grouped trips" (Figure 2).

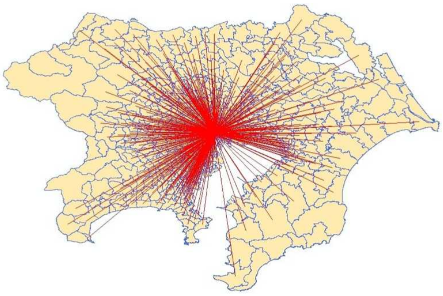

Figure 2 Aggregated trip data from PT Surveys (Source: National Land-Information Office)

3. PREPARATION OF TRIP DATA

GPS logs between clusters of stay points are extracted and grouped as trip data (Figure 3).

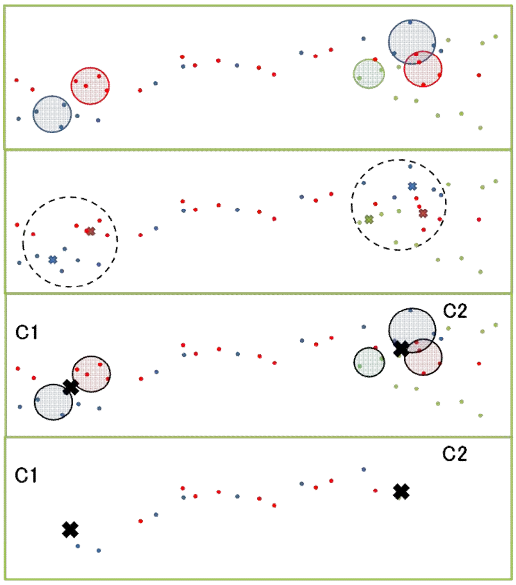

Figure 3 Image of trip extraction process

3.1 *Identifying stay points*

While most studies use speed or log density as thresholds, due to the inconsistency of our logs, we extracted stay points if located within a distance of 150 meters for over a time frame of 20 minutes (Witayangkurn et al., 2010).

3.2 *Clustering stay points*

Stay point centroids are clustered using the k-means algorithm, which groups data into a predetermined *k* number of clusters where each point belongs to the cluster with the nearest mean (MacQueen, 1967). Canopy clustering is used to form temporary subsets to determine *k*.

3.3 *Extracting and grouping trips*

Consecutive non-stay points are extracted as individual trips, and grouped into distinct trips. During this extraction process, instances where different stay points followed one another are counted as trips despite the lack of data, as the grouping process may provide such data.

4. IDENTIFICATION OF TRIPS

We use ground truth data to prepare similar groups of distinct trips, calculate various candidate features for each trip, and then use software to classify all of these labeled trips into a decision tree (Figure 4).

Most studies use speed as a primary feature to separate walk from other modes, but accurate speeds were difficult to determine from our data. Instead, railway proximity (ratio of logs located within 100m of network) works best as the primary feature, separating rail and car, as there is no need for cars to use congested streets near railways. Trip distance is used to separate walk from rail as it is unlikely for rail trips to be shorter than the distance between two stations. On the other hand, walk is separated from car by using the average speed value of all trip data, but tends to return an overestimation of walk trips, possibly due to inaccurate logs or the characteristics of slower bicycle trips (Table 2).

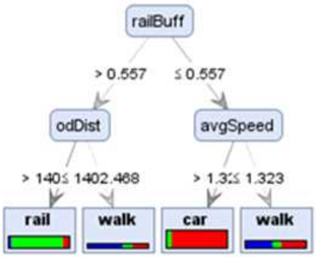

Figure 4 Decision tree results

Table 2 Accuracy results for decision tree

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

5. FINAL RESULTS

Next, we use the parameters and thresholds of our decision tree results to label the 89,077,507 individual trips extracted within our target area.

Table 3 Identified modes of all individual trips

Rail	Walk	Car	Unclassified*
13,944,005	6,448,940	62,105,171	337,439
16.8%	7.8%	75.0%	0.4%

*trips with no data

To validate our results we aggregate our trips in the same way as PT data, grouping trips that start and end in the same combination of zones, then compare the numbers of individual trips for each grouped trip between the two datasets. Scatter plots use PT trips as x values and GPS trips as y values. In general, correlation for total trips and car were strong, and for walk were weak. Coefficients of determination improved when trips to and from the same zone were removed, especially when GPS values were weighted according to users' home locations, using national census population data(Figure 5).

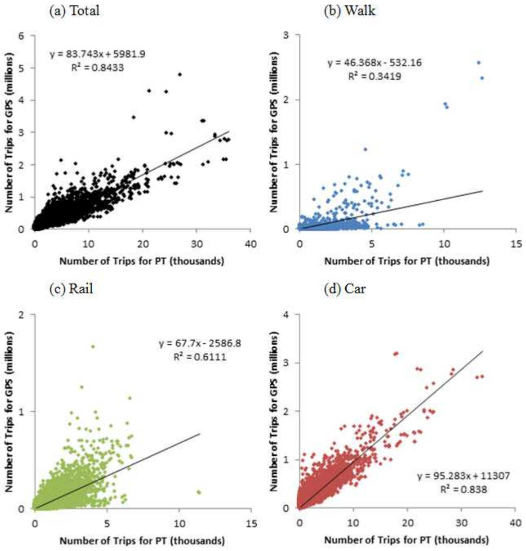

Figure 5 Scatter plots for weighted values of trips between different zones

We assume shorter trips (mostly walk, some car) posed more difficulty due to a lack of data, and trips common in dense urban areas (walk, rail) probably included missing or inaccurate logs. In addition, a slight overestimation of car may be attributed to the inclusion of weekends.

6. MOBILITY ANALYSIS

Finally, identified trips are used to analyze long-term mobility patterns. For example, we calculate the modal share for each individual:

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

where T_m indicates the number of individual trips for mode *m*. Individuals are grouped according to their home locations and average values for each area are mapped (Figure 6).

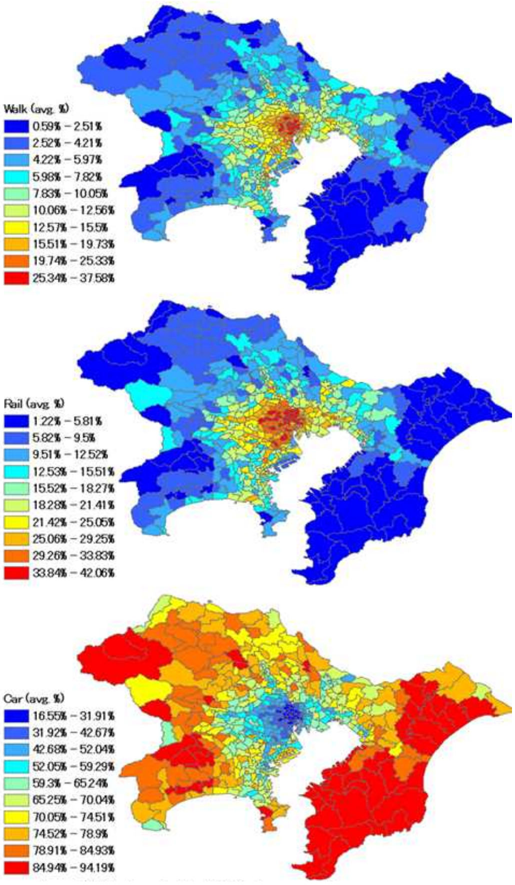

Figure 6 Percentages of trips for each mode

Results indicate that residents of central Tokyo are more likely to travel using rail or walk. In addition, these modal shares are compared with other calculated values such as total distance traveled and home proximity to stations.

7. CONCLUSION

Transportation modes were identified fairly accurately, especially for trips less likely to be affected by the limitations of GPS from mobile phones (i.e. long intervals, inconsistency) such as long-distance car trips. At the same time, validation results emphasized the limitations of single-day survey data. Finally, mobility analyses helped to understand the advantages of collecting large-scale, long-term data.

Future works may add parameters to improve identification accuracy, as well as to include other transportation modes such as bike or bus.

8. ACKNOWLEDGEMENTS

We are grateful to ZENRIN DataCom Co., Ltd., for providing the GPS data used in this study.

9. REFERENCES

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

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

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J. *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability,*

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.

前司敏昭, 堀口良太, 赤羽弘和, 小宮粋史：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

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 http://www.mlit.go.jp/kokudoseisaku/gis/