Identification of the Homes, Offices, and Schools from Long-Interval Mobile Phone Big Data Using Mobility Pattern Clustering

Ryosuke Kobayashi
University of Tokyo
5-1-5 Kashiwanoha,
Kashiwa-shi, Japan
koba@csis.u-tokyo.ac.jp

Satoshi Miyazawa
University of Tokyo
5-1-5 Kashiwanoha,
Kashiwa-shi, Japan
koitaroh@csis.u-tokyo.ac.jp

Yuki Akiyama
University of Tokyo
5-1-5 Kashiwanoha,
Kashiwa-shi, Japan
aki@csis.u-tokyo.ac.jp

Ryosuke Shibasaki
University of Tokyo
5-1-5 Kashiwanoha,
Kashiwa-shi, Japan
shiba@csis.u-tokyo.ac.jp

Abstract

Currently, many companies collect data from their customers, and some are trying to sell the anonymized data on the so-called "Data Transaction Market." Because this trend is likely to continue, the demand for the extraction of valuable information from anonymized data will increase. Meanwhile, Japan is considering improvements in commuting by promoting work-style reform. To this end, it needs to find regional or individual issues among commuters and develop an analysis model for detecting commuters, including nighttime workers, from geospatial big data. In consideration of this, we constructed mobility pattern clusters using long-interval mobile phone big data. Then, we identified the location of homes, offices, and schools in the Kanto region, Japan and achieved a strong correlation with the factual population.
Keywords: Mobile Phone Data, Trajectory, Home Detection, Commute, Clustering, Mobility Pattern

1 Introduction

Currently, many companies are collecting data from their customers; some companies are trying to take their anonymized data to the so-called "Data Transaction Market." For example, well-known telecommunication companies sell their mobile customers' location data (Washington Post, 2019). Of course there are issues such as privacy which have deterred some companies from selling their customers' location data. However, the anonymization of data is improving, and this marketing trend is likely to continue, so the demand for the extraction of valuable information from anonymized data will increase (Wei et al., 2018).

In Japan, work-style reform is promoting the improvement of commuting, and a part of the legislation includes allowing more flexible work-styles for workers. Over 20% of all Japanese workers are employed between 10 p.m. and 5 a.m. as nighttime workers (Kubo et al., 2014) as a result of many 24-hour stores providing nighttime jobs. To reform its work style, Japan needs to find regional or individual issues among commuters and develop an analysis model after detecting commuters, including nighttime workers, from geospatial big data. To accomplish this, the identification of the locations of homes, offices, and schools using anonymous data is required.

Many previous studies have identified the location of homes and offices. Bhattacharya et al. (2012) focused on the user's bearing changes and extracted locations, such as homes, using GPS data from the mobile phones of two people. Focusing on the most visited places, Alexander et al. (2015) extracted home locations using call detail record (CDR) data. Kung et al. (2014), also using CDR data, focused on dwell-times to extract the locations of homes and workplaces. However, few studies have considered nighttime workers.

In the current study, we constructed mobility pattern clusters using long-interval mobile phone big data and identified the locations of the homes, offices, and schools, including those of nighttime workers, in the Kanto region of Japan.

2 Data and Methods

2.1 Data

We used the following three types of data.

1) People-flow point data (Agoop Corp., 2016). Agoop Corp. collects information, such as latitude, longitude, timestamps, and ID from all users who install its applications. When collecting geospatial information, positioning systems depend on GPS, Wi-Fi, and base stations. We used the data from June 2016. The data contained 3,070,439 users in the Kanto region, which partially included the same users as those on other days.

2) Population census (Statistics Bureau of Japan, 2015). This data is collected as part of the national census of Japan and includes the daytime population, considering commuters, and the nighttime population of cities and wards in the Kanto region.

3) Person trip (PT) data and origin-destination flow (Ministry of Land, Infrastructure, Transport and Tourism, 2010). The PT survey is a well-known statistical survey in Japan. The PT questionnaire-based survey aims to investigate actual travel behavior. Its data includes the number of people in offices, in which we could observe distribution tendencies for workers by city and ward in the Kanto region.

2.2 Data Features

Our dataset, "Point data of people-flow," had three features. First, the user ID was reset each day at midnight; therefore, we could not analyze each user's trajectory across days. Second, it intentionally lacked data from 1 a.m. to 5 a.m. to prevent the easy detection of users' significant places, in terms of privacy. Generally, people tend to be at home during these hours, and, thus, it makes it difficult to identify their significant places. Third, when collecting the data, a long temporal sampling interval was used. This depended on the smartphone's operating system. For example, Android mobile phones send user data, such as latitude and longitude, to a

server every thirty minutes, and previous research has described this as a long temporal sampling interval (Alessandretti et al., 2017).

2.3 Methods

Shiotsu et al. (1998) stated that people, even if working, spend more time in their homes each day. Therefore, we focused on the stay time at each point by using a first significant place, where people stayed for the longest time during a day, and a second significant place, where people stayed for the second longest time during a day. In other words, the 1st significant place is likely to represent home and the 2nd significant place is likely to represent an office, school, or favorite place. In addition, we use the term "commuter" to denote workers and students, as opposed to the term "worker." Figure 1 shows a flowchart of the study's method.

The method comprised preprocessing, modeling, and evaluating. During the preprocessing, we eliminated users who moved more than 150 km during one trip and whose data contained errors to accurately extract commuters. In addition, we eliminated users that lacked data. Periods of 24 hours were divided into three timeframes, 0 to 8 a.m., 8 to 4 p.m., and 4 to 0 a.m. If there was no data in each timeframe, we removed the users' data. Then, we interpolated the information from 1 to 5 a.m., with reference to the information around these hours, regarding the 2nd data feature. Next, we transformed one user's scattered points within 365 m into a single point using density-based spatial clustering of applications with noise, DBSCAN (Ester et al., 1996).

After completing the preprocessing, we conducted the modeling by focusing on the stay time at each point and the distance from the identified home. Using stay time, we extracted the 1st and 2nd significant places.

Subsequently, we classified all the users into groups by applying a K-means clustering algorithm (MacQueen., 1967), which is a widely used method for partitioning n points into k clusters in which each point belongs to the cluster with the nearest mean. We applied the Elbow method to determine that 20 was the optimal number of clusters at which the sum of the squared errors abruptly decreased (Ketchen & Shook., 1996). For the initialization, we used K-means++ (Arthur & Vassilvitskii., 2007), which improved the seeding in the K-means clustering.

We focused on people's mobility patterns rather than on how far away people moved; therefore, we selected a distance ratio (Equation (1)) for each user as a feature. In addition, considering the 3rd data feature, we linearly interpolated the information to fix for the lack of data.

$$Distance\ Ratio\ (\sim 0\text{–}1) = \frac{Distance\ from\ one's\ home}{Maximam\ distance\ from\ one's\ home} \quad (1)$$

After the clustering, we modified some of the clusters (described in section 2.4) and selected for commuter clusters; the shape of a time-series graph indicated whether it was a commuter or not. During the night, daytime commuters, such as a daytime worker or a student, stay in their homes to rest. During the day, they travel to their offices or schools. If the graph was mountain-shaped, like a trapezoid, it was likely that it was a daytime commuter's mobility pattern. In contrast, if the graph was valley-shaped, like an inverted trapezoid, it was likely to be a nighttime worker's mobility.

Finally, we used the population census data to evaluate the estimated population.

2.4 The Switching Process after Clustering

A problem occurred when using the stay time to decide which point was home; many of the clusters comprised valley-shaped graphs (Figure 2). This was because some daytime workers spent more time in their offices than their homes. Figure 2 shows only the valley-shaped 24h time-series graphs of all the clusters, and its y-axis shows the distance ratio from the 1st significant place. Even though there are many nighttime workers in Japan, these graphs indicated that it is co -rrect to consider people who spend more time in their offices.

Consequently, some of the clusters needed to be modified to solve this issue. Therefore, we used the PT survey dataset, including information about how many people work in each district, because it suggested which cities contain more workers. In other words, the dataset represented whether it was an office-like area, where more people work, or not. In general, people tend to live in less urban districts, which do not correspond with office-like areas, than urban districts, which do correspond to office-like areas. The PT-survey-based analysis indicated that the 1st significant place, the home, holds 1.5 to 2.8 times more workers than the 2nd significant place for each cluster, 1 to 5. Each 1st significant place for clusters 1 to 5 was in much more office-like areas than the 2nd significant place. In regard to cluster 20, the 1st and 2nd significant places held approximately the same amounts. Therefore, we switched the 1st and 2nd significant places for clusters 1 to 5, and classified as a daytime worker cluster. In addition, cluster 20 was classified as a nighttime worker cluster.

Figure 1: Flowchart of the study's method used in the study.

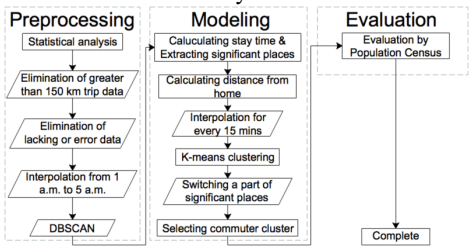
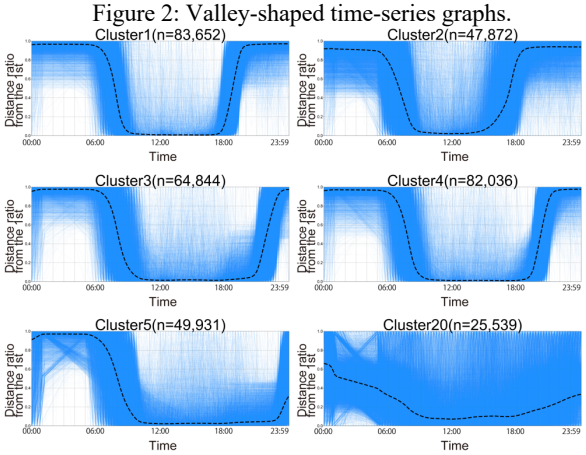


Figure 2: Valley-shaped time-series graphs.



3 Results and Discussion

Table 1 shows the results of a day-of-the-week comparison of the average number of users on weekdays and weekends. As seen in the right-hand column, over 60% of clusters 1 to 10 consisted of weekday users. In contrast, about 60% of cluster 11 to 15 consisted of weekend users. Clusters 16 to 20 included equal numbers of weekday and weekend users. According to Japanese statistics (Japan Broadcasting Corporation, 2016), ~ 88% of adults commute on weekdays, and ~ 45% of adults commute on weekends. Therefore, clusters 1 to 10 were likely to be clusters of commuters in terms of the day-of-the-week comparison.

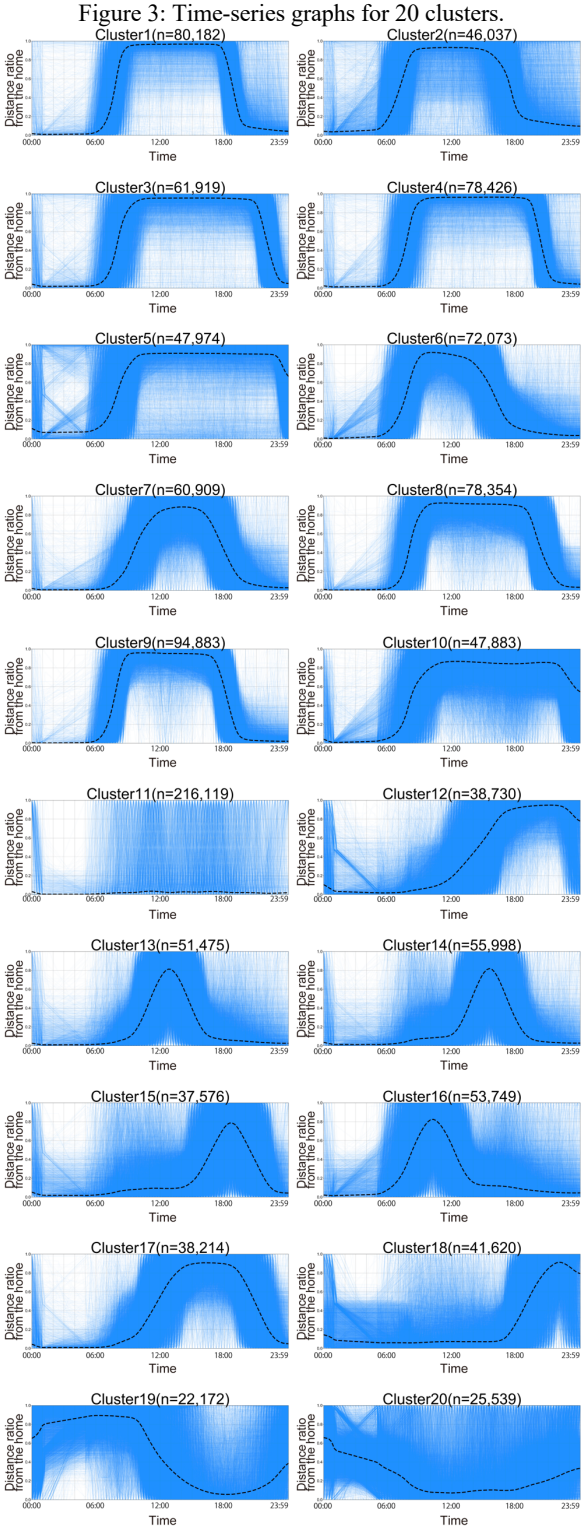
Figure 3 shows each cluster's time-series graph from 0 to 23 p.m. In each graph, a maximum of 20,000 blue lines was randomly selected as samples. In addition, the black dotted line indicates the mean of each cluster. At first glance, there appeared to be a difference between clusters 3 and 4; however, when looking closer, we observed a difference on an hourly scale. As noted in section 2.4, we switched the 1st and 2nd significant places for clusters 1 to 5. Therefore, the shapes of the graphs became mountain-shaped opposite to valley-shaped. Focusing on the shapes of the graphs, clusters 1 to 10 were mountain-shaped and cluster 20 was valley-shaped. However, clusters 13 to 18 were of people moving temporarily, who did not seem to be commuters.

Considering the day-of-the-week comparison and the graph shapes described above, clusters 1 to 10 were classified as daytime commuters and cluster 20 was classified as nighttime commuters. Although cluster 20 included many weekend users, it is not surprising as there are demands for nighttime workers on weekends such as 24-hour store's clerks. In addition, Cluster 20 appeared to hold ~ 3.5% of all the workers; however, this number differed largely from the factual data. According to Kubo et al. (2014), nighttime workers account for 21.8% of all workers in Japan. This indicates that focusing only on the graph's shape will not extr-act the nighttime workers with high accuracy. Clusters 13 to 18 appeared to be clusters of people spending time on leisure activities, for example, going shopping and to restaurants, because there were many weekend and temporary outgoings. In contrast, cluster 11 seemed to be a cluster of people staying in their homes.

Table 1: Day-of-the-week comparison.

Cluster	Weekday (user)	Weekend (user)	Percentage of weekday users (%)
1	15,543.4	2,967.5	84.0
2	7,968.2	4,015.5	66.5
3	11,957.4	2,528.5	82.5
4	15,334.0	2,683.0	85.1
5	8,412.2	3,935.0	68.1
6	11,601.6	7,032.5	62.3
7	9,584.8	6,492.5	59.6
8	14,326.8	3,360.0	81.0
9	17,097.2	4,698.5	78.4
10	8,082.6	3,735.0	68.4
11	27,207.6	40,040.5	40.5
12	5,208.4	6,344.0	45.1
13	6,528.2	9,417.0	40.9
14	6,734.4	11,163.0	37.6
15	4,501.0	7,535.5	37.4
16	7,779.0	7,427.0	51.2
17	5,485.4	5,393.5	50.4
18	5,979.4	5,861.5	50.5
19	3,211.4	3,057.5	51.2
20	3,670.0	3,594.5	50.5

Figure 3: Time-series graphs for 20 clusters.



4 Evaluation

After identifying the commuters, we evaluated the model's accuracy by focusing on the correlations between the detected population and the population census data. Figure 4 shows the results of this evaluation.

The left-hand graph in Figure 4 shows the correlation between the detected population and the nighttime population; the Pearson correlation coefficient was 0.98 with high significance, which indicated that the identification of homes was very accurate. The Pearson correlation coefficient would have been 0.81 if we had not switched the 1st and 2nd significant places. This means that the reversal improved the accuracy of the model.

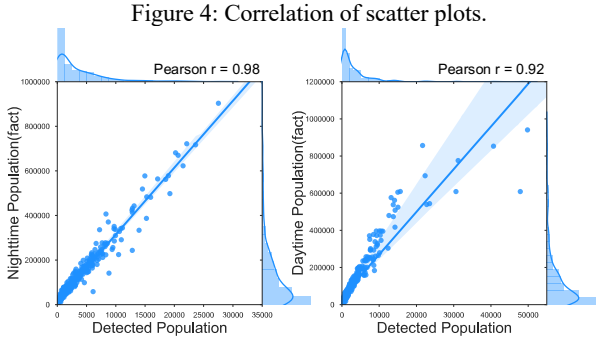
The right-hand graph in Figure 4 shows the correlation between the detected population and the daytime population, considering commuters. The daytime population from the population census data conveniently included nighttime workers. In the census data, the daytime population is defined by Equation (2) using an example of city A.

In section 3, we identified clusters 1 to 10 and 20 as commuters. Applying the numbers of these clusters' commuters to Equation (2), we calculated the number of nighttime commuters for each city or ward, taking the commuter clusters into account.

$$\begin{aligned}&(Daytime\ population\ in\ city\ A)\\&= (\text{Nighttime population } in\ city\ A)\\&- (Commuters\ from\ city\ A\ to\ another\ city)\\&+ (Commuters\ from\ another\ city\ \ to\ city\ A)\end{aligned} \tag{2}$$

The Pearson correlation coefficient was 0.92 with high significance (Figure 4; right-hand graph); therefore, the identification of the commuters was well done. However, when we looked closer at the right-hand graph, we observed a gap between the detected and the factual data for a few points. These points represented areas of the highest population densities in Japan, such as Minato-ku and Chou-ku in Tokyo. These gaps could be related to a lack of the number of nighttime workers, as mentioned in section 3. Therefore, the Pearson correlation coefficient was 0.92 because of these gaps.

Figure 4: Correlation of scatter plots.



5 Conclusion

The current study revealed that the clustering of mobility patterns can identify significant places, such as homes, offices, and schools, even if the dataset used long-interval geospatial big data. Additionally, it revealed that it is difficult to extract nighttime workers using the shape of a graph and a day-of-the-week comparison.

Three issues arose during this study. First, the PT survey does not target all of Japan; therefore, when applying this model to other areas of Japan, a PT survey dataset might not be available. Second, the detected population suggests fewer nighttime workers than the factual population. Although the evaluation results were good, the census data do not distinguish a daytime commuters from a nighttime commuters. Additional features such as land use data could be beneficial for extracting data regarding nighttime workers more accurately. Furthermore, we revised a few of the clusters' significant places on the assumption that people tend to live in less office-like and less urban areas, and their offices tend to be located in more office-like and more urban areas. However, recently, to save tax, help the areas prosper, etc., some companies, especially tech companies, have moved to less urbanized districts. This trend is likely to continue; therefore, it is possible that this assumption will become unsuitable.

In future work, we aim to accommodate those trends and demands as the work-style reform scenarios to produce an improved traffic demand model.

Acknowledgements

This research was supported by a grant from the H–UTokyo Lab, which is the joint research center of Hitachi Ltd. and University of Tokyo. In addition, we would like to thank the Center for Spatial Information Science (CSIS) (Joint Research Program No. 794), University of Tokyo, for the contributions.

References

Alessandretti, L., Sapiezynski, P., Lehmann, S., & Baronchelli, A. (2017) Multi-scale spatio-temporal analysis of human mobility. PLoS One, 12(2), e0171686.

Alexander, L., Jiang, S., Murga, M., & González, C. M. (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data, Transportation Research Part C: Emerging Technologies, 58, 240-250.

Arthur, D. & Vassilvitskii, S. (2007) k-means++: the advantages of careful seeding, SODA '07 Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 1027-1035.

Bhattacharya, T., Kulik, L., & Bailey, J. (2012) Extracting significant places from mobile user GPS trajectories: a bearing change based approach, SIGSPATIAL '12 Proceedings of the 20th International Conference, 398-401.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996) A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, KDD'96 Proceedings of the 2nd International Conference, 226-231.

Japan Broadcasting Corporation. (2016), Comprehensive Survey of Time Use,

https://www.nhk.or.jp/bunken/research/yoron/pdf/20160217_1.pdf [Accessed February 22th 2019].

Ketchen, J. D. & Shook, L. C. (1996) The application of cluster analysis in strategic management research: an analysis and critique, Strategic Management Journal, 17(6), 441-458.

Kung, S. K., Greco, K., Sobolevsky, S., & Ratti, C. (2014) Exploring universal patterns in human home-work commuting from mobile phone data, PLoS ONE, 9(6), e96180.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations, Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

Shiotsu, M., Yoshizawa, S., Ikeda, K., & Nozaki, A. (1998) Survey on human activity patterns according to time and place, The Architectural Institute of Japan's Journal of Architecture and Planning, 63(511), 45-52.

Washington Post. (2019) Congress should make it harder for cellphone carriers to sell your location data. https://www.washingtonpost.com/opinions/congress-should-make-it-harder-for-cellphone-carriers-to-sell-your-location-data/2019/01/09/93602366-1451-11e9-803c-4ef28312c8b9_story.html [Accessed February 22th 2019].

Wei, R., Tian, H., & Shen, H. (2018) Improving k-anonymity based privacy preservation for collaborative filtering, Computers & Electrical Engineering, 67, 509-519.