Development of People Flow Data with Individual Demographics based on Mobile Phone GPS Big Data

Yoshiki Ogawa
Institute of Industrial Science
The University of Tokyo
Tokyo, Japan
ogawa@csis.u-tokyo.ac.jp

Yuki Akiyama
Center for Spatial Information Science
The University of Tokyo
Tokyo, Japan
aki@csis.u-tokyo.ac.jp

Sekimoto Yoshihide
Institute of Industrial Science
The University of Tokyo
Tokyo, Japan
sekimoto@iis.u-tokyo.ac.jp

Ryosuke Shibasaki
Center for Spatial Information Science
The University of Tokyo
Tokyo, Japan
shiba@csis.u-tokyo.ac.jp

Abstract

We develop a method to instantly visualize human flow data and demographics during tsunami evacuation using large-scale mobile phone global positioning system (GPS) data. The evacuation speed varies, and the spatial distribution of people during a disaster is greatly affected by the age and sex of the people in a crowd. To address these issues, we analyze mobile phone GPS data to estimate home/work locations, apply spatiotemporal interpolation using road/building global information systems' data, and estimate the age and sex from the National Person–Trip survey using machine learning in the Kochi city area.

1 Introduction

It is necessary to use actual data to model the initial spatial distribution of people in a city when estimating transportation demand and predicting human casualties during an earthquake or tsunami. For example, current earthquake damage assessments in Japan usually rely on static population distribution data, such as that from the nighttime population census, organized in 250 m grid units, derived from a recent national census (Central Disaster Prevention Council, 2012). However, this static data only describes the night population and aggregate data. Considering disaster prevention policies, it is important to develop a population distribution that can be used to evaluate disaster risk from a variety of viewpoints, including macro-scale (e.g., prefectures) and the micro-scale (e.g., city blocks). Additionally, there is the problem that census data cannot define trajectories of traffic modes, including cars and pedestrians. Thus, essential population distribution data, based on times and modes, have not yet been developed. Furthermore, government evacuation plans do not consider human trajectories (Zheng et al., 2010). To collect dynamic human trajectory data, there is a Person–Trip (PT) survey of people traveling in a single day obtained from a questionnaire. This survey was accomplished by collecting data in small sample sizes and low frequencies, such as with a 1 % household travel survey conducted once every decade (Yamamoto, 2009). There are problems with this, however. Survey costs are high, data update frequency is low, and data do not consider tourists (Witayangkurn et al., 2013). However, owing to the spread of GPS-equipped mobile phones in recent years, extensive geospatial data are accumulated daily on the movement of people, and represented as "big data." Such data could be used to determine the static flow of people moving around in Japan.

Because mobile-phone GPS can only acquire observation times and latitudinal and longitudinal data, we still need to estimate demographic attributes (age and gender) for evacuation simulations. However, it is difficult to obtain positional information (e.g., call detail records (CDR), GPS) with demographics (Nishimura et al., 2013).

In this study, we analyze big data that transforms mobile phone GPS raw data into people distribution data, which can reveal people flow for a number of populations with demographics (i.e., age and sex), using Kochi city as an example. By developing a GPS data-mining method, we can discover the urban flow of people at any time and can quantify the spatial distribution of people flow, reflecting gender and age.

In Section 2, we consider current literature for determining the flow of people. Additionally, we present survey areas and data, including mobile-phone GPS and PT data in Section 3. Section 4 introduces data mining and modelling methods used to extract statistically reliable flow estimates of individuals and demographic attributes from GPS and the geo information system (GIS) data. Finally, we visualize and present the spatial distribution of the developed people flow data. The results are presented in Section 5.

The objective of this study is to construct a demographic classification model using people flow data from large-scale GPS data obtained from mobile phones, estimating spatio-temporal interpolation and anytime population.

2 Literature Review

To understand the congestion and flow of people from the PT survey data, many researchers have attempted to determine human mobility by referencing mobile phone data (Horanont et al., 2013; Song et al., 2013; Akiyama et al., 2016). However, these data are not as structured as the PT survey data, which contains location and time information of higher-frequency, meaningfulness, and accuracy (Song et al., 2016).

In the following sections, we provide a detailed review of previous studies that use the PT survey, GPS, and CDR data to determine human mobility in a wide area.

The PT survey examined the static travel patterns of humans in each survey area, and its objective was to figure

out who travelled where, when, why, and how (Sekimoto et al., 2012). Because the PT survey was a questionnaire, representative of only a part of the actual population, a sample expansion was allotted to each dataset according to the actual population. Additionally, data were aggregated for each survey area. Sekimoto et al. (2011) developed a method to understand people flow from PT data. One common weakness of this study is that it only uses weekday behavior for a single day, omitting weekend and holiday behavior (Osaragi, 2016).

Using mobile phone data, Ye et al. (2009) determined the behavioral patterns of users. The Tourism Agency studied tourist dynamics from their behavior by using location information from GPS data. However, previous studies have not reflected the use of large-scale GPS data in disaster management, because the data are not easily obtained. Moreover, when applying the data to disaster management, it is necessary to add a coefficient to match the actual population of the sample. Apichon et al. (2012) studied human behavior using large-scale mobile phone GPS data after the 2011 Great Japan Earthquake. They clarified how useful information revealed humans' reactions during disaster scenarios and how the evacuation process could be monitored in near-real time.

3 Study area and data

We use Kochi city in Japan as a case study (Fig. 1). Kochi is 20 km by 35 km. It had a population of 330,000 in 2018, of which 92,000 were over 65, and 32,000 were under 20.

Figure 1. Study area.

3.1 Mobile phone GPS data

We use data from the 2011 mobile phone GPS "Congestion Analysis®" logs provided by ZENRIN DataCom Co., LTD. Konzatsu–Tokei® refers to people flow data collected by individual locations sent from mobile phones with enabled auto-GPS function, by consent, through the "docomo map navi" service provided by NTT DOCOMO, INC. These data are processed collectively and statistically to conceal private information. Original location data is GPS data (e.g., latitude, longitude) sent in a minimum period of 5 min without information specifying gender or age. This is a large database constructed of text data from approximately nine billion records (2012) belonging to approximately 1.5 million users throughout Japan. The target sample includes data for about 30,000 people. The data processing method devised for this research is applied to GPS data by NTT DOCOMO, INC. The data are available for ages 15 to 89 with sufficient samples. The accuracy of GPS data used in this study is a maximum of 250 meters.

3.2 PT data

In this study, PT survey data is used as training data for demographics estimation. For personal trip data, we use traffic-flow planning data from national and local governments. The method of creating person flow data using PT data is comprised of geocoding the spatio-temporal position of the starting point of the sub-trip base and performing a route-search on the shortest path basis. It is obtained by interpolating detailed data from each network, including age and gender. We use personal trip data (2000) in Kochi city (about 47,000 samples).

4 Methods

This section describes GPS analysis methods for maintenance of people flow data.

4.1 Geocoding to Point of Interest (POI)

First, flags representing staying or moving were attached to each point of the target area's GPS data, principally observed at intervals of 5 min with an extraction of the staying points of each user from position coordinates and observation times. The staying points define the representative points that indicate where a user remains stationary for a certain time and within a certain range. We estimated stagnation of the observation point by dividing the trip and by judging the retention point as the threshold value of the group within a 300-m radius for 15 min or longer. Because we used GPS data for multiple days, there are several stay points (e.g., near the house). Therefore, in order to put together different spatially close dwell points, clustering was used. DBSCAN was used as an algorithm for clustering. DBSCAN uses two parameters, distance threshold (Eps) and target number threshold (Min Pts), defines the connection relationship of objects within the radius Eps, and when there are more than Min Pts connected targets, points are classified into the same cluster. In this research, Eps was set to 150 m and Min Pts was set to two. The calculation was performed and the central point of the maximum cluster from the column of points of the staying period was called the staying point (Horanont and Shibasaki, 2011; Hadano et al., 2013). We estimated the work and home locations to each staying point from their observation times. Thus, using building data (Z-map Town II (Zenrin Co., LTD.)) for POI, each point was established by taking the center of gravity of the staying point from midnight and 4 PM during 2010 and estimating the latitude and longitude of each building in the neighborhood. Subsequently, by focusing on observation

points before and after each observation, we estimated velocity, and traffic modes of each point were added. Traffic modes are defined as car, work, bike, bus, and train in this study.

4.2 Road network interpolation

Considering the estimated traffic modes, we used road and railroad network data to interpolate the spatio-temporal data. We also processed route placements from the points observed from the starting points. If the connection of the observation point to the road network was large, the error grew. Thus, we considered the link connection by searching for the nearest neighbor on the link. Thus, it was possible to shrink the error via the spatio-temporal interpolation method of Sekimoto et al. (2011) (Fig. 2.).

4.3 Scaling factor

To match the GPS data to the population, we estimated sample expansion from estimated work and home locations. To estimate the sample expansion, we used the statistical data of the population and number of employees. We used the 250 m-mesh data of the 2010 residential population census and the 250 m-mesh data of the 2010 economic census for the number of employees. Figure 3 shows the method of estimating the magnification number, estimated by dividing the residential population by the number of home terminals by the number of employees and by the number of terminals at the work location. From the average of the magnification coefficients (Mh and Mw) we obtained the final magnification coefficient. By imparting the magnification factor to each datum, it was possible to estimate the distribution of the population rather than the distribution of a sample.

4.4 Estimation of attribute

The flow of the estimation method is shown in Fig. 4. First, we constructed a model using PT data (47,000 sample) as training data. Then we applied the model to GPS data. Because PT is questionnaire and GPS is observational data, the properties of both are different. Therefore, when applying a model to GPS data, some areas do not match. As a countermeasure, we weighed the attributes among groups using census data when applying the construction model. Because there were variations after their application, we again revised those with low probability to match the census. The learner model adopted the LASSO (Least Absolute Shrinkage and Selection Operator) support vector machine (SVM), a machine learning method, showing an excellent classification performance by internally performing nonlinear mapping (i.e., kernel trick) (Boser, Guyon, and Vapnik, 1992). Using LASSO (i.e., L1 regularization), variable selection and model construction were done simultaneously. The number of features used for model construction was calculated using the total time to leave, the home departure time, the home return time, total staying time, total travel distance, and distance from home to workplace. We used the same feature quantities for both gender and age, creating a learning model using LASSO SVM. Parameters of the SVM are tuning by 10 cross validations; set radial for the SVM kernel; SVM type: eps regression; gamma: 0.125; and epsilon: 0.1. We also attempted to use other major classification models, such as random forests and the lasso logit model, but the highest cross validation accuracy was achieved by the LASSO SVM (Table 1).

4.5 Validation

The reliability of the model, based on personal trip data, was made by a 10-fold cross validation with an error matrix. The results of the 10-fold cross validation are shown in Table 1. The accuracies were 64.4% in the case of sex and 67.7% in the case of age using the LASSO SVM. The results of the error matrix of the age presumption were found to have variations of accuracy. The accuracy of 20 years or less was 40%, and the accuracy of 65 years or more was 54%. However, the estimation of 21-to-65 year-olds was 80% (see Table 2). This is probably because many college and high school students behave just like workers (21-to-64 years), even when they are under 20. Looking at the results of the error matrix of the gender estimation model, the overall accuracy was 63%. However, the male estimation accuracy was lower than that of females (Table 3), because men and women have followed similar lifestyles in recent years, indicating that it is difficult to distinguish only by feature quantity. With the estimated people flow and demographics for each phone user in the study, we can then estimate population with a high spatial resolution that is not available in the census statistics.

Table 1: 10-fold cross validation result

Demographics	**Accuracy (%)**
	LASSO SVM
Gender	64.4
Age	66.6

Table 2: Error matrix of age estimation by LASSO SVM

	Prediction data				
Classified data	***Age***	***Under 20***	***21-65***	***Over 65***	***User's accuracy***
	Under 20	589	835	66	40.0%
	21-65	183	3798	767	80.0%
	65 over	77	929	1173	53.8%
	Producer accuracy	70.0%	68.3%	58.5%	Overall: 66.0%

Table 3: Error matrix of gender estimation by LASSO SVM

	Predict data			
Classified data	***Gender***	***Men***	***Women***	***User's accuracy***
	Men	1803	2056	45.7%
	Women	1072	3486	76.5%
	Producer accuracy	62.7%	62.9%	Overall: 62.8%

5 Results

Figure 5 shows an example of people flow data with individual demographics on October 17, 2012 in Kochi City for approximately 150 thousand people. Figure 6 presents the estimated population (for people older than 5–20 years old, older than 21–64 years old, older than 65 years old) in Kochi at a 500 m-mesh by using the GPS data. Over time, it became possible to determine the time when a person with an attribute moved to a location. For example, various scenarios, such as weekdays, holidays, and events could be generated. This is a pseudo-distribution and does not indicate the actual distribution (i.e., individual identification is impossible). By using mobile phone GPS data with age and sex as described above, it is now possible to follow the flow of people every day, which cannot be obtained by conventional census and PT data. By conducting evacuation simulations using people flow data, based on actual observed data, it should become possible to more accurately determine the vulnerability of each area.

Figure 2. Method of spatio-temporal interpolation (Sekimo et al., 2012).

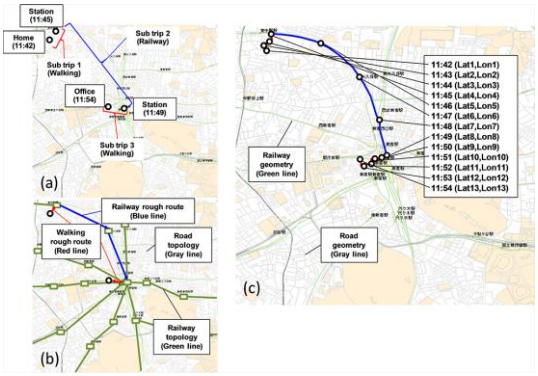


Figure 3. Method of estimating scaling factor (people/ a phone).

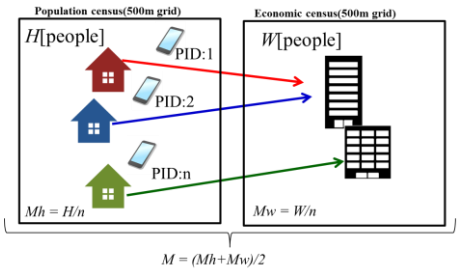
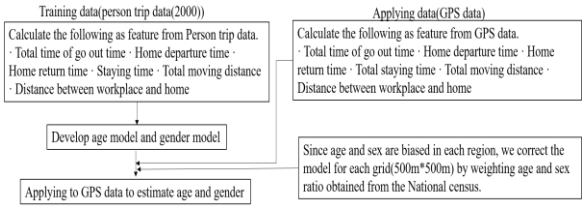


Figure 4. Method of estimating individual demographics (age and gender).



6 Conclusion

In this study, we developed people flow data applicable to time-based micro-simulation from large-scale mobile phone GPS data around Kochi City. By interpolating spatial and temporal data that described the distribution of people and flow on a person basis, the data for the year 2012 was prepared for Kochi. The maintenance of wide-area people flow data from GPS data for disaster simulation will be important essential data for future disaster simulation.

We have two issues for future work. Firstly, our estimates for GPS data only used night population data and work place population data obtained from the population census and economic census to estimate a scaling factor of the population distribution. In addition, there is data available for the number of passengers on the trains in Japan; it is possible to make a more reliable estimation of population data by utilizing those data. In the future, to improve accuracy, we want to improve the estimation accuracy by considering traffic modes, such as whether cars are used, and by considering POI.

Acknowledgment

The present study was made possible by data provided by NTT Docomo, Inc., ZENRIN DataCom Co., Ltd. Furthermore, this study was carried out as part of joint research No. 122 with the Center for Spatial Information Science (CSIS), The University of Tokyo. This study received funding support under the "Post-K computer priority issue application development" program (Priority issue 3: Construction of an integrated forecasting system for complex disasters based on earthquake and tsunami) sponsored by the Ministry of Education, Culture Sports, Science, and Technology, and the Strategic Basic Research Program and AIP (Artificial Intelligence, Big Data, IoT, Cyber Security Integration Project) Network Public/Private R&D Investment Strategic Expansion Program (PRISM) of the Japan Science and Technology Agency (JST). The authors wish to express their gratitude to the individuals, companies, and sponsor organizations noted above.

References

Akiyama, Y., Ueyama, S., Shibasaki, R., and Adachi, R. (2016) Event Detection Using Mobile Phone Mass GPS Data and Their Reliability Verification by MDSP/OLS Night light Image.

Apichon, W., Horanont, T., and Shibasaki, R. (2012) Performance comparisons of spatial data processing techniques for a large scale mobile phone dataset. Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. ACM.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* - COLT '92, 144.

Central Disaster Prevention Council (2012) About the Nankai Trough earthquake measures of building damage and human damage (in Japanese).

Hadano, A., Wiayangkurn, M., Akiyama, A., Horanont, Y., and Sibasaki, R. (2013) A Study on Extracting Characteristic of Visitors at Commercial Area from GPS Data, *IGU 2013 Kyoto regional conference*, 01435.

Horanont, T. and Shibasaki, R. (2011) Nowcast of Urban Population Distribution using Mobile, Phone Call Detail Records and Person Trip Data, Proc. Int. Conf. Computers in Urban Planning and Urban Management 2011, CD-ROM.

Horanont, T., Witayangkurn, A., Sekimoto, Y., Shibasaki, R. (2013) Large-Scale Auto-GPS Analysis for Discerning Behavior Change during Crisis, *Intelligent Systems, IEEE*, 28, 4, 26–34.

Nishimura, T., Akiyama, Y., Kanasugi, H., Horanont, T., Shibasaki, R., and Sekimoto Y. (2013) Analysis and Evaluation of Human Lifestyle Pattern Using Mobile Phone GPS Log Data, *Proc. Int. Conf. The 34th Asian Conference on Remote Sensing* 2013, CD-ROM.

Osaragi, Toshihiro (2016) Estimation of Transient Occupants on Weekdays and Weekends for Risk Exposure Analysis. *Proc. Int. Conf. ISCRAM*.

Sekimoto, Y., Shibasaki, R., Kanasugi, H., Usui, T., and Shimazaki, Y. (2011) PFLOW: Reconstruction of people flow by recycling large-scale fragmentary social survey data, *IEEE Pervasive Computing*, 10, 4, 27–35.

Sekimoto, Y., Watanabe, A., Nakamura, T., Kanasugi, H., and Usui, T. (2012), Combination of spatio-temporal correction methods using traffic survey data for reconstruction of people flow, *Pervasive and Mobile Computing Journal*, Elsevier.

Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., and Shibasaki, R. (2013) Intelligent System for Human Behavior Analysis and Reasoning Following Large-scale Disasters, *Intelligent Systems, IEEE*, 28, 4, 35–42.

Song, X., Zhang, Q., Sekimoto, Y., Shibasaki, R., Yuan, N., and Xie, X. (2016) Prediction and Simulation of Human Mobility Following Natural Disasters, *ACM Transactions on Intelligent Systems and Technology* (ACM-TIST), 8–37.

Witayangkurn, A., Horanont, T., Sekimoto, Y., and Shibasaki, R. (2013) Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud platform. In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, 1219–1228.

Yamamoto, T. (2009) Comparative analysis of household car, motorcycle and bicycle ownership between Osaka metropolitan area, Japan and Kuala Lumpur, Malaysia. *Transportation*, 36(3), 351–366.

Ye, Y., Zheng, Y., Chen, Y., Feng, J., and Xie, X. (2009) Mining Individual Life Pattern Based on Location History, *Proc. Int'l. Conf. on Mobile Data Management Systems, Services and Middleware*, 1–10.

Zheng, V. W., Zheng, Y., Xie, X., and Yang, Q. (2010) Collaborative location and activity recommendations with GPS history data, *Proc. 19th Int. Conf. World wide web - WWW* 10, 1029.

Figure 5. Developed people flow data with individual demographics on October 17, 2012 in Kochi City.

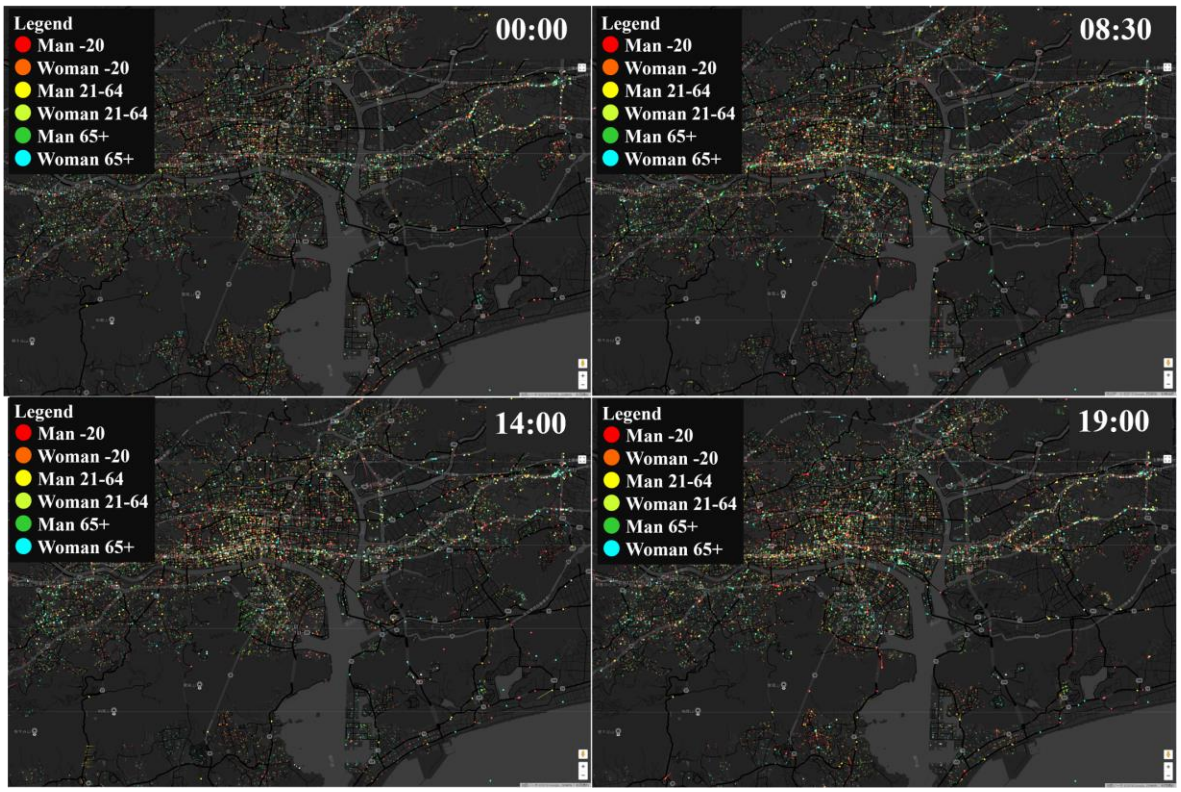


Figure 6. Developed people flow data with individual demographics on October 17, 2012 in Kochi City (upper: aggregated to 500 m mesh).