Synthetic People Flow: Privacy-Preserving Mobility Modeling from Large-Scale Location Data in Urban Areas*

Naoki Tamura[1,2][0000−0002−3362−697X], Kenta Urano[1][0000−0003−2906−537X], Shunsuke Aoki[1,3][0000−0002−3331−2778], Takuro Yonezawa[1][0000−0001−9781−0402], and Nobuo Kawaguchi[1][0000−0002−0444−2290]

[1] Graduate School of Engineering NagoyaUniversity, Aichi, Japan
[2] tam@ucl.nuee.nagoya-u.ac.jp
[3] National Institute of Informatics

Abstract. Recently, there has been an increasing demand for traffic simulation and congestion prediction for urban planning, especially for infection simulation due to the Covid-19 epidemic. On the other hand, the widespread use of wearable devices has made it possible to collect a large amount of user location history with high accuracy, and it is expected that this data will be used for simulation. However, it is difficult to collect location histories for the entire population of a city, and detailed data that can reproduce trajectories is expensive. In addition, such personal location histories contain private information such as addresses and workplaces, which restricts the use of raw data. This paper proposes Agent2Vec, a mobility modeling model based on unsupervised learning. Using this method, we generate synthetic human flow data without personal information.

Keywords: Spatio-temporal Data Analysis · Privacy Preserving Data Mining · Unsupervised Learning.

1 Introduction

Recently, there has been an increasing demand for traffic simulation and congestion prediction for urban planning, especially for infection simulation due to the Covid-19 epidemic. On the other hand, the widespread use of smartphones and wearable devices equipped with GPS (Global Positioning System) has made it possible to collect the location history of many users with high accuracy. Therefore, it is expected that the city-level human flow data, which shows how people move and stay in the urban environment, can be used for simulation. For example, large-scale data on location history has been used to analyze the effects of policies during a pandemic[1]. However, it is difficult to collect location histories for the entire population of a city, and detailed data that can reproduce trajectories is expensive. In addition, such personal location histories contain

* Supported by AMED, JST-CREST, NICT.

private information such as addresses and workplaces, which restricts the use of raw data. In this study, we generate synthetic people flow data which can reproduce the city-level people flow based on the real location history data. This data does not contain personal information because it does not correspond to the real user's location history and can be freely processed and visualized.

In order to generate a flow of agents, we need an activity model that defines how each agent moves and stays over time. Although many user activity modeling methods have been studied in the past, they mainly require a large amount of labeled data and detailed location histories. On the other hand, a method for modeling the usage of a region and the activity tendency of users using GPS data and unsupervised learning has been studied. This allows us to model user activities based on less frequent and unlabeled data than before. As the model is based on GPS data collected on a daily basis, it also has the advantage of being able to take into account changes in the environment over time, such as changes in social policy, pandemic outbreaks, and seasonal changes. This paper proposes Agent2Vec, a mobility modelling model based on unsupervised learning. This model abstracts the tendency of users to move and stay as a distributed representation. By clustering these distributed representations, we extract groups of users with similar tendencies. We can use each group of users as an activity model of synthetic agents, and generate synthetic people flow data. The main contribution of this work are given in the following:

- Agent2Vec: Unsupervised learning model that abstracts the tendency of users to move and stay as a distributed representation;
- Generating synthetic human flow data without personal information and with higher granularity(50m mesh);
- Evaluation of synthetic data in terms of density of stay and amount of movement;

We have generated a synthetic dataset using a real GPS location dataset, and we have confirmed that the synthetic dataset reproduces real people flow by visualization and evaluation. However, we have also identified some challenges in terms of population distribution and distance traveled.

2 Related Work

In order to generate synthetic human flows, we need an activity model of how each agent moves and stays. There is a long history of attempts to model how people move and stay in their daily lives, based on person-trip surveys[2][3]. However, because of the high cost and low frequency of collection, person-trip surveys can only model typical patterns of activity and cannot model changes in people's activities as the environment changes.

On the other hand, with the recent proliferation of mobile devices, a large amount of CDR (Call Detail Record) and GPS location history has been obtained, and activity modeling using these data has become popular. For example, Song et al. [4] implemented an LSTM multi-task learning system for

learning human mobility and traffic patterns and a city-level simulation system, mainly using GPS data. Yin et al. [5] uses the movement history from CDR to model the activity with hidden Markov model. Ouyang et al. [6]uses GPS trajectory data to model human mobility and synthetically generate movement trajectories. Borysov et al. [7] use deep learning-based methods to generate a large number of and more diverse user models and aim to generate unsampled models by combining their elements. Another example is the work on modeling the daily activity schedule of users under various and complex factors [8][9]. Other research exists that uses reinforcement learning-based activity modeling to generate more natural movement trajectories[10][11]. The challenge of these methods is that they require a large amount of detailed, labeled data and the setting of an appropriate reward function to train the model. However, labeling of movement trajectories is costly since the movement of an individual is generally high dimensional information based on various factors.

Research that has actually generated city-level synthetic-population flow data includes the use of GPS, CDR, as well as comprehensive datasets such as population distribution and traffic volume [12]. However, such data are generally expensive, and although the spatial granularity is around 250-500m mesh, more granular human flow data are needed for congestion prediction and infection simulation. In this paper, we generate more granular human flows with a granularity of 50m mesh from unlabeled GPS data.

3 Proposed Method

This method aim to generate synthetic human flow data that can reproduce real human flow from the original location history data. Input data is obtained as latitude, longitude, and timestamp of a real user with a terminal. The synthetic data is also generated as latitude, longitude, and timestamp of each agent. To generate synthetic data, we need an activity model of how each agent moves and stays respectively over time. This activity model is mainly modeled using unsupervised learning in this method.

Fig. 1 shows an overview of the method. First, we generate a distributed representation of LU (Location Usage) for each mesh according to the tendency to stay and represent user's movement in the form of LU transitions. This LU is generated for each mesh by learning the tendency to stay at that mesh by Word2Vec. This allows the user's location information to be represented as information, including POIs (Points of Interest). Next, we extract travel and stay information from the location history of real users and abstract it in the form of a distributed representation for each user. Then, by clustering the distributed representation of each user, we classify users by their tendency to move and stay. The users in each cluster have a similar tendency to move and stay. For example, users in the housewife cluster tend to spend most of their time at home, while users in the salaried worker cluster tend to go to work and leave work in the morning and evening. Finally, we model the activities of agents based on the movement and stay tendencies of the users in each cluster and generate syn-

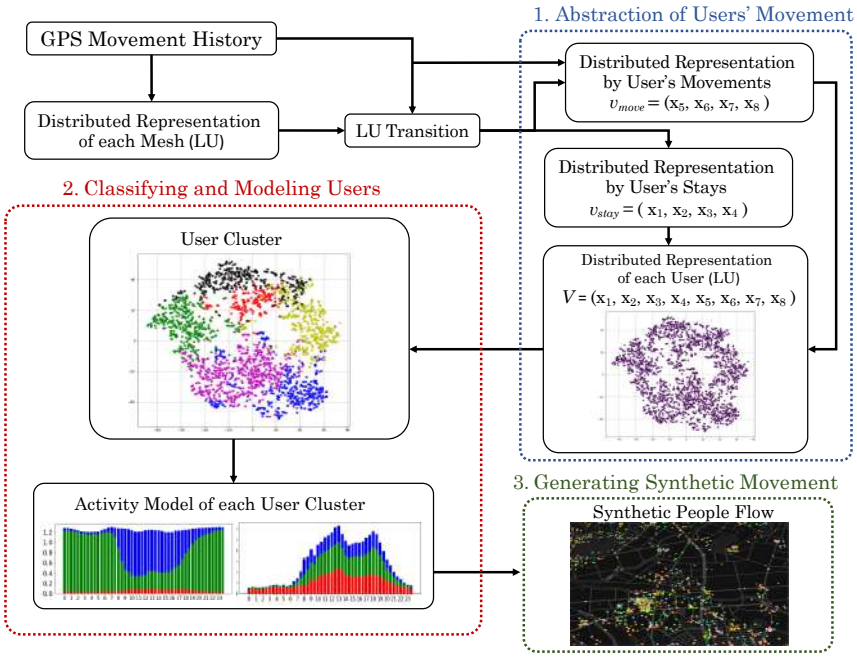

Fig. 1. Overview of the Proposed Method

thetic human flow data. These methods are explained in detail in the following sections.

3.1 Abstraction of Real Users' Movement and Stay Tendency

In this section, to classify real users, we abstract the tendency of each user to move and stay as a distributed representation V. This section describes the procedure for calculating this V. First, we divide each user's location history into 30-minute time slots and assign a mesh to stay in each slot. We then compute the LU, a distributed representation of the trend in usage over time for each mesh. It is possible to classify them according to their tendency to stay in the mesh by clustering them,. In this paper, we refer to this cluster of meshes by LU as LU cluster. These LU clusters are, for example, residential clusters for long stays in the morning and evening, and office clusters for long stay in the daytime.

Then, we calculate the distributed representation reflecting the tendency of each user to stay and move, v_{stay} and v_{move}, independently in Agent2vec. Agent2Vec is an application of Word2Vec that generates a distributed representation of each user by learning the user's tendency to move and stay. Fig. 2 illustrates the architecture of Agent2Vec. The input layer is a one-hot vector

for each user, and the output layer is a one-hot vector with flagged dimensions for each movement or stays feature. The weights of the middle layer obtained by training the input vectors produce a distributed representation that reflects the tendency of each user to stay. The number of dimensions of this distributed representation for each user is equal to the one of the hidden layer N, In this case, the number of dimensions of both $v_{stay}$and $v_{move}$was set to $N = 4$. The stay features are learned by Agent2Vec using three pieces of information for each stay: a weekday or a holiday, the period of the stay, and the stay LU cluster. This makes it possible to compute a distributed representation v_{stay} based on the tendency of each user to stay in the mesh, i.e., when and what attributes they stayed in. The movement features are learned by Agent2Vec using four pieces of information for each travel: a weekday or a holiday, the period of the travel, the distance of the travel, and whether the travel is to the main mesh. This main meshes are the meshes in which each user has stayed for many days, such as home and office. In our method, we define the main mesh as the mesh where the user is observed to stay for more than half of the days in the data obtained for each user. This allows us to compute the distributed representation v_{move} based on the movement tendency, i.e., when, at what distance ,and to what mesh the user frequently visits. Finally, we combine v_{stay} and v_{move} to create V, a distributed representation of each user's tendency to move and stay.

3.2 Real User Classification

In this section, we classify users according to their tendency to move and stay by clustering their distributed representation V. We used Kmeans++ for clustering. Fig. 3 - 5 shows some examples from the actual clustering into 40 clusters. Fig. 3 shows the percentage of LU clusters that stayed in each time slot on weekdays and holidays, while Fig. 4 and Fig. 5 show the number of moves, the percentage of distance traveled, and the percentage of moves to the main mesh in each time slot. The vertical axis in Fig. 4 and Fig. 5 is the percentage of moves in each period (every 30 minutes), normalized for the number of people in each user cluster, weekdays, and days off. From Fig. 3, clusters 1 and 2 tend to stay in the residential cluster during the evening hours and in the office cluster during the daytime, which is similar to that of a typical office worker. Cluster 3 stays in the residential area cluster except a few times during the daytime when it stays in the restaurant cluster, indicating that it tends to stay like a homemaker. On the other hand, cluster 2 has the peaks of travel between 7:00 and 9:00 in the morning and between 17:00 and 20:00, and the travel of 3km to 5km is conspicuous. In cluster 3, the amount of travel is lower than clusters 1 and 2, and the distance traveled is higher than 0.5km. Fig. 5 shows that clusters 1 and 2 have a high proportion of travels to the main mesh in the morning and evening, going to work and returning home. Cluster 3 has more travels to non-main meshes in the afternoon than clusters 1 and 2. These clusters are just examples, but by clustering, V in this way, we can classify not only the tendency of users to stay but also the tendency of users to move.

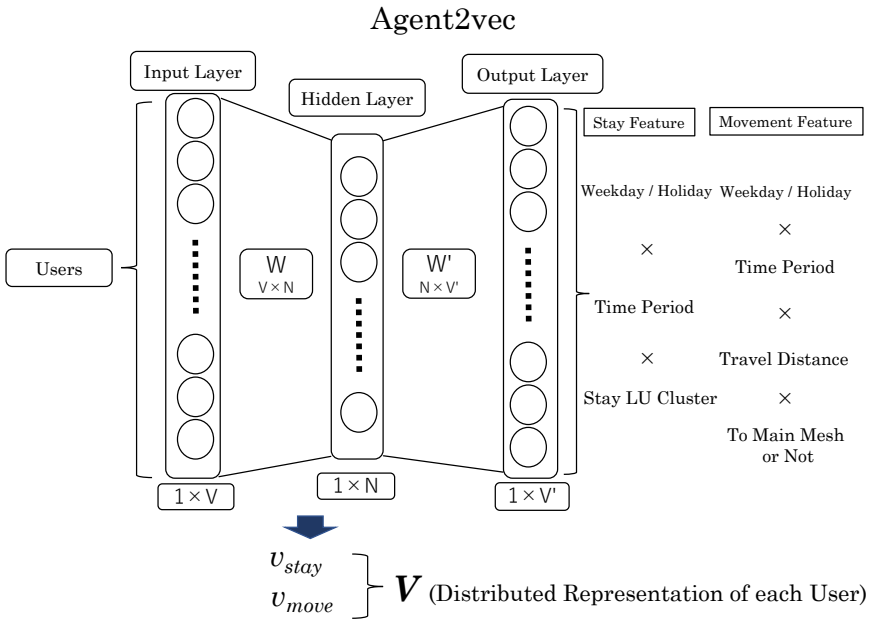

Fig. 2. architecture of Agent2Vec

3.3 Generating Synthetic Flow Data

In this section, we explain how to model the activity of agents from each classified user cluster and how to generate the location history of each agent. First, we decide which user cluster each agent will use as a model. The probability $P(u)$ that an agent uses the user cluster u as a model is defined as follows, using N_u that is the number of users belonging to the user cluster u(U is the number of user clusters).

$$P(u) = \frac{N_u}{\sum_{k=0}^{U} N_k} \tag{1}$$

That is, a typical cluster with many users is likely to be chosen as a model for agents.

Next, the location history of the agent is generated according to the selected user model. Specifically, for the mesh in which the agent is currently staying, we probabilistically select "1. whether to move or stay" and "2. to which mesh to move if to move" during each 30-minute time slot. The decision to move from the current stay mesh is made according to a Poisson distribution based on the average number of moves of real users λ. The probability P_{stay} that a agent does

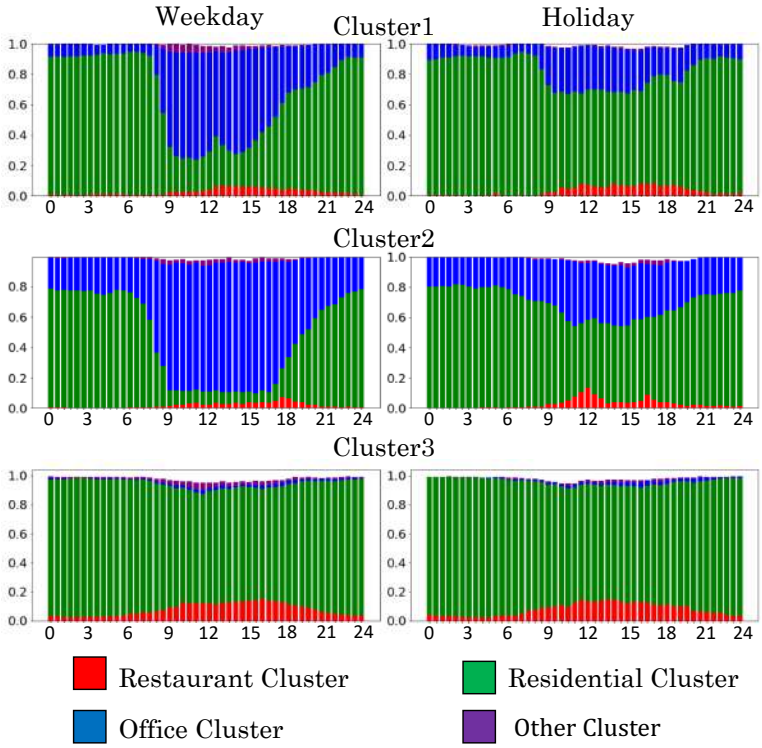

Fig. 3. Stayed LU Cluster

not move in a given time slot $t$is

$$P_{stay} = \frac{\lambda_t^0 exp(-\lambda_t)}{0!} = exp(-\lambda_t) \tag{2}$$

At the same time, the probability that a agent moves is

$$P_{move} = 1 - exp(-\lambda_t) \tag{3}$$

This average number of travels λ is averaged for each time slot and for each LU cluster before the movement. This is based on the idea that the occurrence of travel is correlated not only with the time of day but also with the LU cluster before travel. For example, the probability that a user in the salaryman cluster moves to the residential cluster at 8:00 in the morning is considered to be different from that in the office cluster.

Finally, we explain the decision of which mesh to move to. This is mainly based on three factors: the density of stay in each time zone, the LU similarity, and the distance traveled.

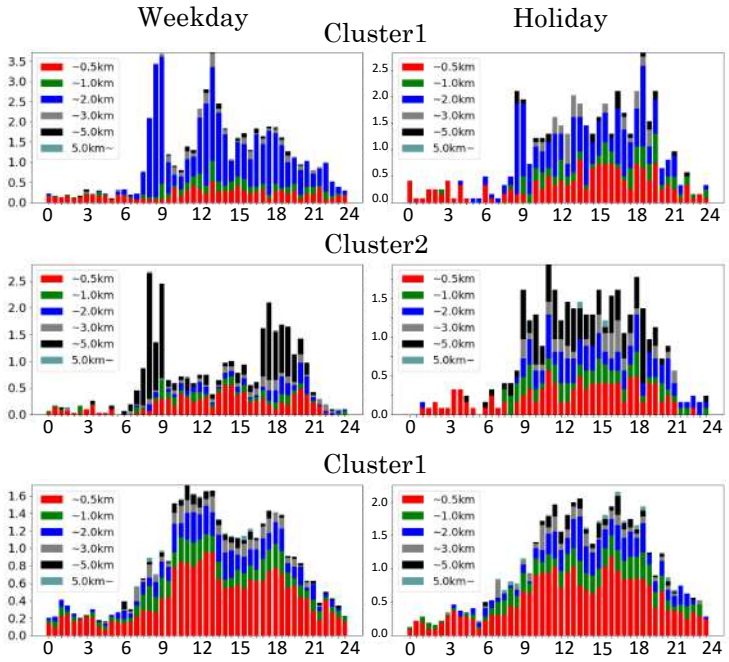

Fig. 4. Travel Distance Tendency

For the density of stay, the thickness of stay of actual users in each time zone is calculated in advance. The density of stay D_t in a given time slot t is calculated using the number of users $C(m)$ in the mesh m as follows(M is the number of meshes).

$$D_t(m) = \frac{C(m)}{\sum_{k=0}^{M} C(k)} \tag{4}$$

By considering this as the probability of migration, we can set a higher probability of migration to a mesh with more stays.

For LU similarity, we set the probability of moving to a mesh with LUs that have high resemblance to the destination LU cluster to generate a destination along with the user's POI. The destination LU cluster is determined according to the proportion of LU clusters (3) that stay in each time zone of the user cluster to which it belongs. To reduce the amount of computation, we take the average LU as a representative of each LU cluster and calculate the probability of moving based on the similarity of the average LU of each LU cluster.

As for the travel distance, we set the travel probability to the one with the appropriate travel distance to be high based on the trend of travel distance of real users (mean and variance of distance). Specifically, the probability of movement

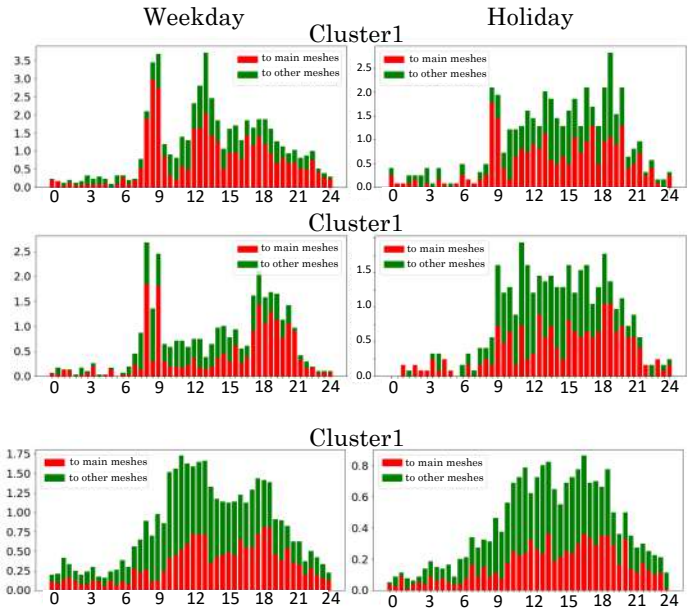

Fig. 5. Travel to Main Mesh Tendency

is assigned to each mesh according to a normal distribution based on the mean distance and variance.

In this way, the probability of moving to each mesh is calculated for each density of stay, LU similarity, and travel distance, the product of these probabilities generates a move to the mesh. Therefore, agents decide the destination mesh for each period based on the density of stay, POI, and distance from the source mesh. By generationg agent's travel probabilistically from each user model, we can generate synthetic people flow dataset for an arbitrary number of users.

4 Evaluation

To evaluate the proposed method, we generate synthetic human flow data using a GPS location history dataset provided by Blogwatcher. The target area was Nisshin City, Aichi Prefecture, and the period of the data was March 2020. We used the data of 2155 users whose location information was sufficiently available in the target area. The number of generated agents was 30000, the number of LU clusters was 4, and the number of user clusters was 40. Fig. 6 shows a visualization of the generated data. The plots show the position of each agent, which moves over time according to the generative model. We can see that during the daytime hours there are more stays in the office area near the centre, while during the night there is a decrease in stays.

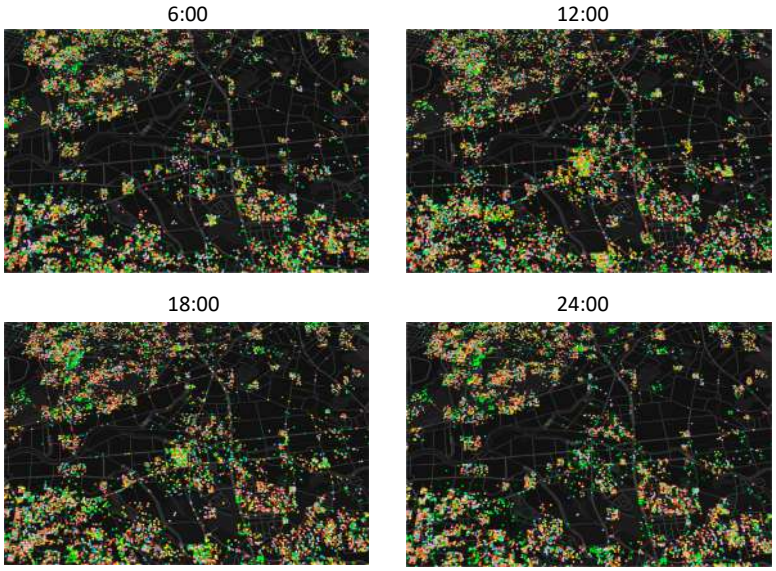

Fig. 6. Visualization of Syntheic Data

In the following, we evaluate the generated data regarding density of stay, amount of travel, and distance traveled. The evaluation method is based on the method used in [12]. As there is no correct data for urban flows, we basically compare the data with the original data, and for the density of stay, we compare the data with the population distribution dataset of mobile spatial statistics at each period.

4.1 Stay Density

We evaluated whether the generated data could reproduce the population distribution at each time. Fig. 7 compares the density of stay of the generated data at 6:00, 12:00, and 18:00 with the respective data. From the top, the density of stay for Mobile Spatial Statistics (500m mesh), Blogwatcher (500m mesh), Blogwatcher (50m mesh), and the density of stay for the generated data are plotted for each mesh. Since the population distribution data of Mobile Spatial Statistics is available only at the granularity of 500m mesh, the comparison is made according to this mesh granularity. If the positive correlation between the density of stay in both data is high, the generated data will likely reproduce the actual population distribution. As for the plot with the Blogwatcher data, there is a positive correlation between 50m and 500m meshes. As for the plot with

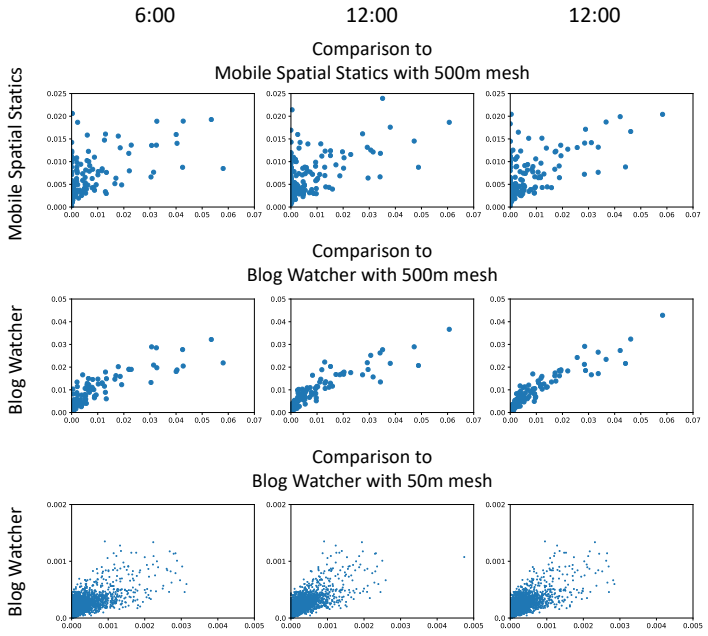

Fig. 7. Comparison of Stay Density

mobile spatial statistics, there are some meshes where the density of stay is high in the mobile spatial statistics data but low in the generated data.

The Fig. 8 shows the correlation coefficients calculated for each period for each dataset. There is a strong correlation with the Blogwatcher data, especially at the 50m mesh granularity (0.75-0.8), thus reproducing a highly granular population distribution. However, the correlation with mobile spatial statistics is low in all periods. This is because the sample size of the original Blogwatcher data is smaller than that of the mobile spatial statistics, and the stay history is biased. Fig. 9 shows the difference in density of stay between the mobile spatial statistical data and the geographically drawn data, with the meshes with large discrepancies in the density of stay colored darker. The plots in the figure show Akaike and Nisshin stations, which are the main stations in the target area, and the meshes with a large difference in density of stay are located around the stations and at the periphery of the target area. The dataset from which the synthetic data is generated focuses on users in the area and whose positions are available for many hours during the period and excludes users who stay outside the area for many hours a day. This may be the reason for the dissociation between the density of stay in the synthetic data and the density of stay in the mobile spatial statistics around the station and the periphery of the area, and

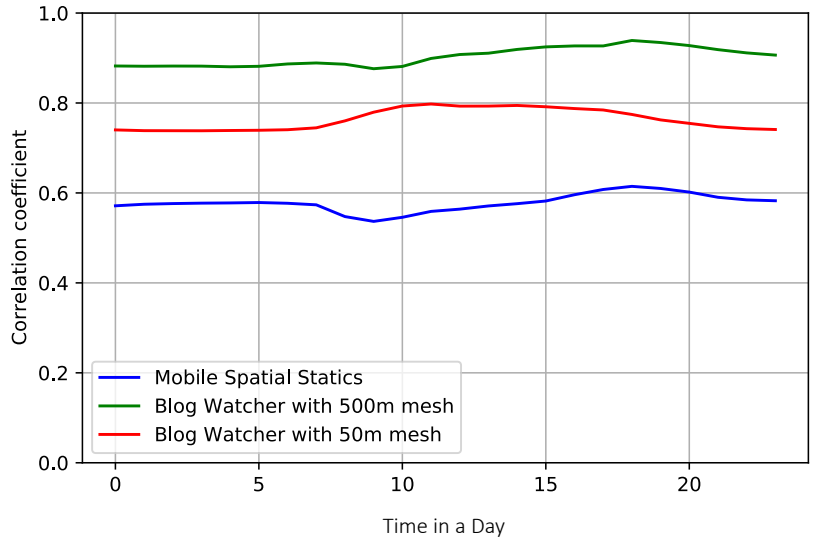

Fig. 8. Comparison of Correlation

the red mesh in the figure is considered the area where users frequently enter and leave the area.

4.2 Amount of Travel

Fig. 10 shows the comparison between the synthetic data and the original data, where the horizontal axis shows the hourly time, and the vertical axis shows the percentage of the travel. Compared with the original data, the generated data reproduces the peak hours, but the amount of travel during the midnight hours is lower than that of the original data.

Fig. 11 shows the comparison of the travel distance between the synthetic data and the original data. We can see that the synthetic data tends to move longer distances than the original data, and especially the distance below 0.5 km is smaller. This may be because the density and LU similarity are too important when selecting the destination mesh. In addition, the distance traveled depends on the user's geographic location, such as the location of the user's home and workplace, this may be because these are not sufficiently modeled in the activity modeling of agents.

Fig. 9. Distribution of Meshes with High RMSE

5 Conclusion

This paper proposes a method to generate synthetic human flow data in an urban environment by utilizing large-scale GPS location history data. In particular, we realize unsupervised user activity modeling using Agent2Vec for distributed representation and Kmeans++ for clustering. By using this method, we were able to generate synthetic people flow data using only unlabeled data. This data reproduces the density of stay in the real world with finer granularity than the conventional data and models the agent's travel by LU transitions. Therefore, the synthetic traffic flow data can reproduce a more realistic flow by creating the travel along with the POI of real users. As for the prospects, we would like to generate more accurate human flow data for the distance traveled, which was not accurate in this evaluation. We are considering an approach that adds a geographical element to the user activity modeling. We also plan to evaluate our dataset by comparing it with various other datasets. In addition, we would like to improve our method to reproduce each agent's travel route, travel speed, and means of travel, as these are considered essential elements for reproducing urban human flows.

References

1. Yabe, T. Tsubouchi, K. Fujiwara, N. Wada, T. Sekimoto, S. Ukkusuri, S.: Non-compulsory measures sufficiently reduced human mobility in Tokyo during the

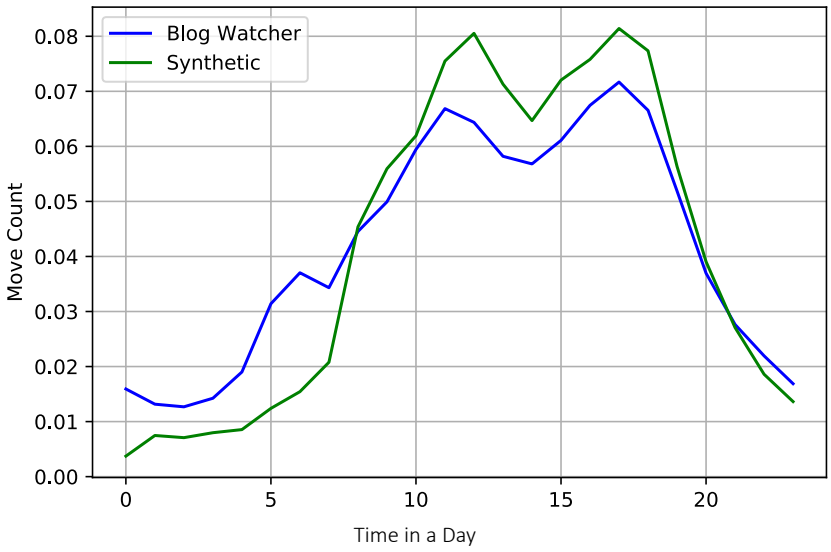

Fig. 10. Comparison of Amount of Travel

COVID-19 epidemic. Scientific Reports volume 10, Article number: 18053. (2020)
2. Bhat, C.-R. Frank, S.-K.: Activity-based modeling of travel demand. Handbook of transportation Science, International Series in Operations Research & Management Science, vol 23, Springer, Boston, MA (1999), pp. 35-61, (2003)
3. Bowman, J.-L. Moshe, E.-B.: Activity-based disaggregate travel demand model system with activity schedules. Transport. Res. Part A: Policy Pract., 35 (1) (2001), pp. 1-28, (2001)
4. Song, X. Hiroshi, K. Ryosuke, S.: DeepTransport: Prediction and simulation of human mobility and transportation mode at a citywide level. IJCAI (2016), p. 16, (2016)
5. Yin, M. Sheehan, M. Feygin, S. Paiement, J.-F. Pozdnoukhov, A.: A generative model of urban activities from cellular data. IEEE Trans. Intell. Transp. Syst., 19 (6) pp. 1682-1696, (2017)
6. Ouyang, K. Shokri, R. Rosenblum, D.-S. Yang, W.: A Non-Parametric Generative Model for Human Trajectories. IJCAI(2018), pp. 3812-3817, (2018)
7. Borysov, S.-S. Rich, J. Pereira, F.-C.: How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation research part C: emerging technologies, vol. 106, pp.73-97, (2019)
8. Drchal, J. Čertický, M. Jakob, M.: Data-driven activity scheduler for agent-based mobility models. Transportation research part C: emerging technologies, vol. 98, pp.370-390, (2019)

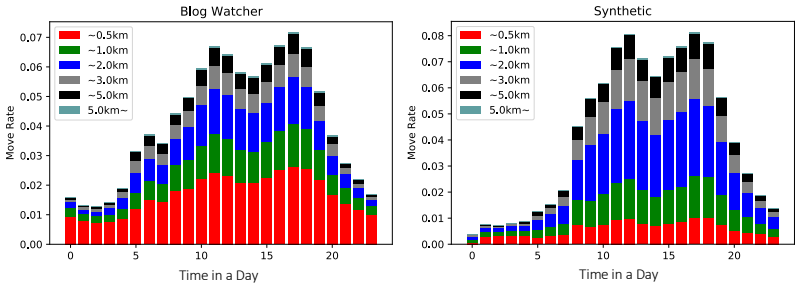

Fig. 11. Comparison of Travel Distance

9. Vecchio, P.-D. Secundo, G. Maruccia, Y. Passiante, G.: A system dynamic approach for the smart mobility of people: Implications in the age of big data. Technological Forecasting and Social Change, vol. 149, (2019)
10. Pang, Y. Tsubouchi, K. Yabe, T. Sekimoto, Y.: Development of people mass movement simulation framework based on reinforcement learning. Transportation research part C: emerging technologies, vol. 117, (2020)
11. Pang, Y. Tsubouchi, K. Yabe, T. Sekimoto, Y.: Replicating urban dynamics by generating human-like agents from smartphone GPS data. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 440-443, (2018)
12. Kashiyama, T. Pang, Y. Sekimoto, Y.: Open PFLOW: Creation and evaluation of an open dataset for typical people mass movement in urban areas. Elsevier, Vol. 85, pp.249-267, (2017)