An Attempt to Reproduce One-day Human Mobility Using Time-Varying Markov Chain Model

Weiying WANG *, Toshihiro OSARAGI **

Abstract: Human mobility has been studied by researchers from many fields such as urban planning, transportation research, geography, computational epidemiology, etc. It has also been modeled using various data sources from different attributes. However, human mobility shows great spatial and temporal complexity, and many aspects remain unknown. In this study, we propose a method of reproducing one-day human mobility based on the time-varying Markov Chain model. A step of choosing individual's future status before the Markov process is conducted, which uses the information of previous mobility. Results show that the spatial and the temporal attributes are well-reproduced, and individual future status can be inferred from the type of previously mobility, regardless of where exactly the person has traveled and when trips happened. Our method can be a tool to infer spatial and temporal attributes of individual mobility from knowledge about aggregated mobility.

Keywords: Human mobility, One-day mobility, Person Trip survey data, Reproduction, Markov Chain model.

1. Introduction

Understanding the nature of human mobility is important in many fields such as urban planning, transportation management, geography, etc. It has raised broad attention and a large body of research has been done in the last decade. It was discovered that human mobility follows several well-defined statistical laws (González et al., 2008; Calabrese et al., 2013; Pappalardo et al., 2015) and exhibit spatial and temporal regularity (Axhausen et al., 2002; Hasan et al., 2013).

Grounded on the above findings, many models have been developed to reproduce human mobility in spatial and/or temporal dimensions. However, most of the research focus on statistical properties such as the distribution of visited locations by time, trip distance, the radius of gyration, etc. Few of them compare the generated data with the real dataset about when and where trips happen, and about the mobility of people with various socio-demographic attributes. In this research, one-day human mobility is studied. We aim at reproducing mobility for people with various individual attributes, where the spatial and temporal properties are well-reproduced. We propose a reproduction model where the future status of a person is conditioned on the types of trips the person has made previously.

2. Dataset, mobility sequences, and groups of people

2.1 Person trip survey data

Person Trip survey has been conducted every ten years by the Ministry of Land, Infrastructure and Tourism of Japan in major urban areas. It is conducted to households and collects information about their travels on a given day. "Trips" defined in Person Trip survey data (hereafter PT data) are illustrated in Fig. 1. It is a one-day dataset, and place and time of the departure and arrival, travel purpose (14 purposes in total, including commuting to work, shopping, entertainment, etc.), and means of trip (15 means: on foot, bicycle, car, bus, train, ship, airplane, etc.), as well as personal attributes (age, gender, occupation, car ownership, etc.), are included in the dataset. In our research, the survey was conducted in 2008 and it covers an area of a circle with a radius of 70 km centered on the Tokyo Railway station (Osaragi and Kudo, 2020). The day starts from 3:00 AM

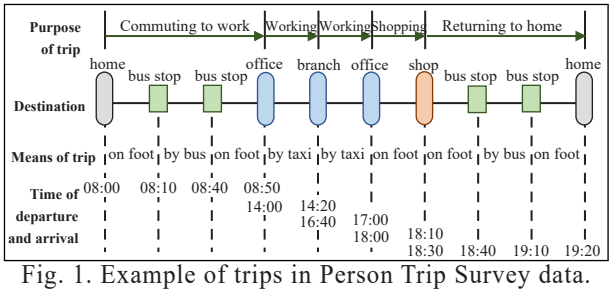

Fig. 1. Example of trips in Person Trip Survey data.

* 学生会員 東京工業大学環境・社会理工学院（Tokyo Institute of Technology）
〒 152-8550 東京都目黒区大岡山 2-12-1 E-mail: wang.w.al@m.titech.ac.jp
** 正会員 東京工業大学環境・社会理工学院（Tokyo Institute of Technology）

at night and ends at 3:00 AM the next day. We further filter out samples whose trip information, such as trip purpose, travel time, location, is missing. After data preprocessing, 581,105 samples were obtained.

2.2 Mobility sequences

Using PT data, individual locations at any time of the day can be inferred. We divide a day into 30-min slots (48 slots for a day). For every person, there is a sequence of 48 slots where each slot is assigned with his/her location during that 30-min interval (Fig. 2, **mobility sequence** hereafter). If a person is taking a trip during the 30-min interval, the location after the trip will be assigned to the time slot. The spatial resolution is the administrative region. Fig. 3 shows all the 279 regions (hereafter *Region 1*, ..., *Region 279*) in Tokyo metropolitan area in the survey. Accordingly, we obtain 581,105 mobility sequences. Our purpose is to generate mobility sequences that are similar to the original ones based on the time-varying Markov Chain model. We attempt to employ rules learned from the existing mobility sequences at the same time.

2.3 Four groups of people

We classify people into four groups based on the similarity of mobility in our previous research (Table 1). Mobility sequences are reproduced for each group in this study.

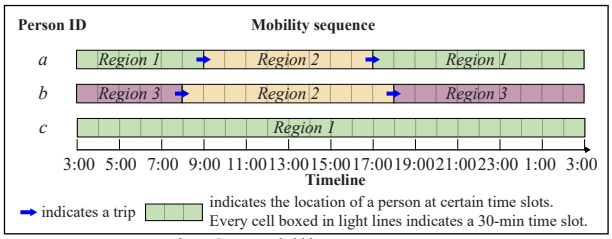

Fig. 2. Mobility sequence.

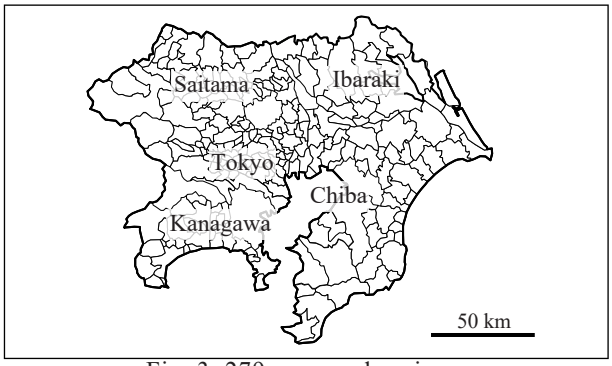

Fig. 3. 279 surveyed regions.

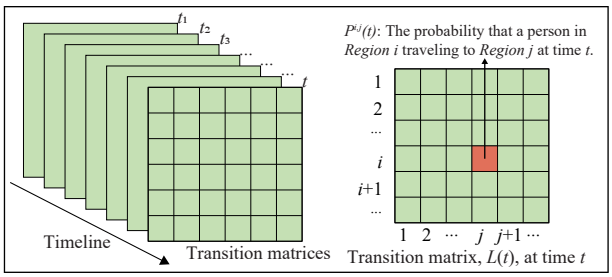

Fig. 4. Time-varying location transition matrix.

Table 1. Four groups of people.

Group ID	Attributes	Counts
Group 1	Household wives/husbands, the unemployed, and farmers.	199,017
Group 2	Workers and college students with ages greater than 14.	307,673
Group 3	High school students or people between 15-19 years old.	19,208
Group 4	People between 5-14 years old.	55,207

3. Methodology

To reproduce mobility sequences that are spatially and temporally similar to the original mobility sequences, we develop a method based on the time-varying Markov Chain model. The method uses the time-varying location transition matrix (3.1) and uses the information of types of current mobility sequences (3.2). Our assumptions and methodology are detailed in section 3.3 and 3.4.

3.1 The time-varying location transition matrix

The time-varying location transition probability matrix, $L(t)$, is a matrix indicating the probability for an individual to have a trip between two regions at time t. The entry in the i-th row and the j-th column, $P^{i,j}(t)$, is the probability of a person in *Region i* beginning a trip to *Region j* at time t (Fig. 4). For each group of people, transition probabilities can be aggregated from the mobility sequences. Moreover, the probability of being at each region at the time t=1 can be calculated. $\pi(i)$ stands for the probability being at *Region i* when t=1.

In simple scenarios, human mobility can be reproduced using such matrices in a Markov process, where the next location is probabilistically chosen from $L(t)$ at time t. The next location only depends on the current location. This is the so-called time-varying

Markov Chain model. However, the spatial and temporal characteristics of human mobility are complicated. The mobility sequences reproduced using a simple Markov process are remarkably different from real human mobility.

3.2 Types of mobility sequences and types of current mobility sequences

There are various **types of mobility sequences** (M_a for person a), such as '*A*' type, '*A-B-A*' type, '*A-B-C-A*' type, etc. (Fig. 5a). '*A*' stands for the region that shows up first in the mobility sequence; '*B*' stands for the second appearing region; and '*C*' indicates the third one, etc. For any person, the **type of current mobility sequence** ($M_a(t)$, t=1,2,…,48) is the type of mobility sequence the person has until the current time (Fig. 5b). M_a and $M_a(t)$ are both sequences of letters, but M_a is the type of mobility for a whole day, while $M_a(t)$ is that for a given time. M_a and $M_a(t)$ only suggest the type of mobility sequence of a person, but indicate nothing about where those regions are, and at exactly what time each trip happened. For different individuals, even if they have the same type of current mobility sequence at the same time, the regions traveled and the time of trips could vary greatly.

3.3 Next-status probability

We assume that people who have the same types of current mobility sequence are likely to have similar patterns of mobility in the next time slot. Grounded on this idea, we predict the mobility of these people in the same way. For people with any type of current mobility sequence at time t, $M(t)$, there can be three types of statuses at time t+1:

▸S_1. Staying in the current region and making no trip with probability denoted by $P_t(S_1|M)$.

▸S_2. Having a trip to a region visited. The probability is denoted by $P_t(S_2|M)$. It can be a collection of probabilities if the person has multiple visits until time t. $Sum(P_t(S_2|M))$ is the sum of probabilities in the collection.

▸S_3. Traveling to an area that the person has never been to. The probability of such a case is denoted by $P_t(S_3|M)$ (=1 - $P_t(S_1|M)$ - $Sum(P_t(S_2|M))$).

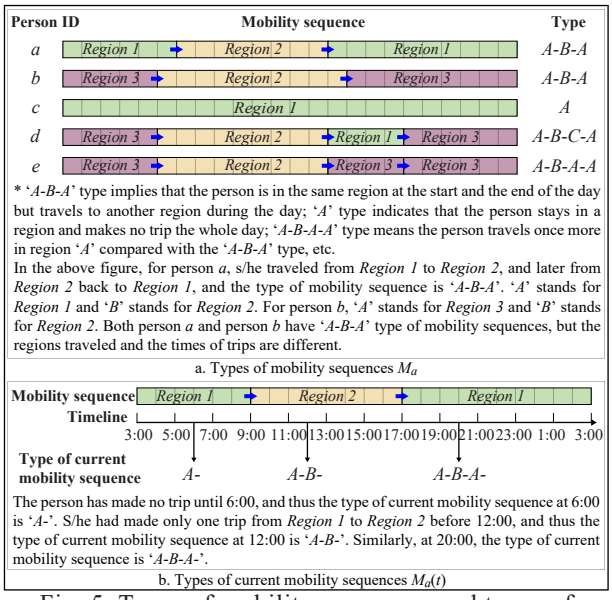

Fig. 5. Types of mobility sequences and types of current mobility sequences.

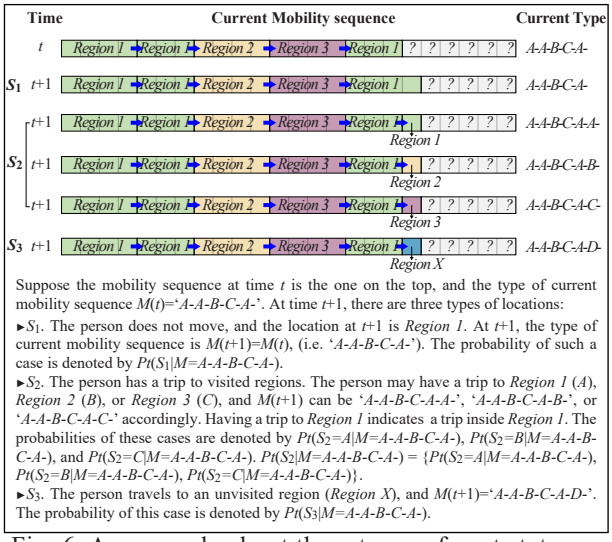

Fig. 6. An example about three types of next statuses.

A detailed example is shown in Fig. 6. For any type of current mobility sequence, the probabilities of being at the three types of statuses at the next time slot can be aggregated from the mobility sequences for each group of people. We call such probabilities the **next-status probabilities** hereafter.

3.4 Method of reproducing human mobility

For each group of people, after calculating the time-varying transition matrix, $L(t)$, the probability of being at each region at the start time, $\pi(i)$, and the next-status probabilities, $P_t(S_1|M)$, $P_t(S_2|M)$, and $P_t(S_3|M)$, we do

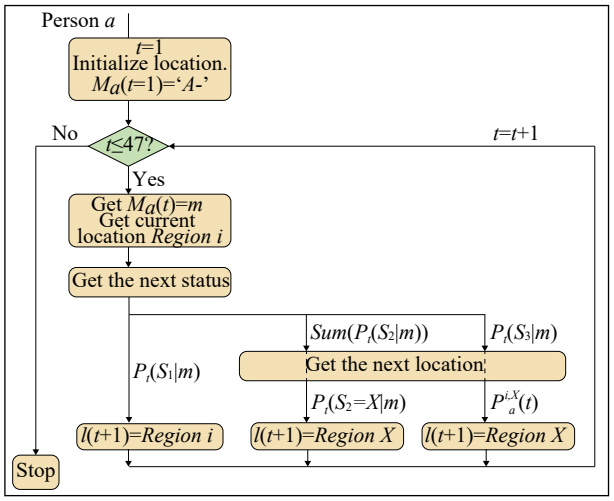

Fig. 7. Flowchart of reproducing human mobility.

the following steps (Fig. 7):

Step 1. At the time t=1, individual location is randomly chosen where the probability for a region to be chosen is $\pi(i)$. The type of current mobility sequence of any person is set to be '*A*-', and '*A*' stands for the first location of the person.

Step 2. When $1 \le t \le 47$, there can be three types of statuses at time t+1: staying in the current region and making no trip (S_1), having a trip to a previously visited region (S_2), and traveling to an unvisited region (S_3). The probabilities of these statuses are $P_t(S_1|M)$, $P_t(S_2|M)$, and $P_t(S_3|M)$ accordingly.

Step 3. When the next status is S_1 or S_2, the location of the person in the next time slot is the current location or one of the previously visited locations. When the next status is S_3, the next location (denoted by $l(t+1)$) is chosen from unvisited regions. The probability of choosing an unvisited region is given by:

$$P_a^{i,j}(t) = \frac{P^{i,j}(t)}{\sum_j P^{i,j}(t)} \quad (1)$$

where $P_a{}^{i,j}(t)$ is the probability for person a in *Region i* to choose *Region j* at time t; *Region i* is the region the person in at time t; *Region j* is an unvisited region; $P^{i,j}(t)$ is the entry in the i-th row and the j-th column of transition matrix $L(t)$.

Step 4. By looping *Step 2* and *Step 3* until t=*47*, mobility sequences can be generated.

4. Evaluation

Human mobility is complicated and there are a large number of types of current mobility sequences (i.e. $M(t)$). For each group, only the types of current mobility sequences that cover more than 0.1% of the population are focused. The next-status probabilities of these types are calculated. For other types of current mobility sequences, the next location is chosen from the time-varying location transition matrix $L(t)$ in a Markov process. After reproducing mobility sequences for each group, we evaluate the spatial and temporal similarity.

4.1 Spatial similarity

In our research, spatial characteristics of mobility sequence are what regions are traveled by an individual in what order. Here, we define a **region sequence** as a sequence of regions that a person traveled in the day. Any mobility sequence obtained from the original or the generated dataset can be converted to a region sequence. For example, if a person has the first trip from *Region i* to *Region j* at 9:00, and the second trip from *Region j* to *Region i* at 17:00, then we have the region sequence *Region i*→*Region j*→*Region i*. Different from the type of mobility sequence, region sequence claims what regions are traveled. One region sequence belongs to one type of mobility sequence, while one type of mobility sequence includes multiple region sequences.

We evaluate the spatial similarity by comparing the counts of region sequences from the original dataset and the reproduced dataset (Fig. 8). To clearly present our results, for each group of people (see Table 1), the comparison is shown in three figures. In Fig. 8, each dot represents one region sequence. The x-value is the count of mobility sequences belonging to such a region sequence in the original dataset, and the y-value is the corresponding count in the reproduced dataset. If the x-value and the y-value are close (i.e. the dot is close to the diagonal line drawn from the origin to top right), the numbers of people with such a region sequence in the two datasets are close, which means the spatial attributes are well-reproduced. Results show that, for

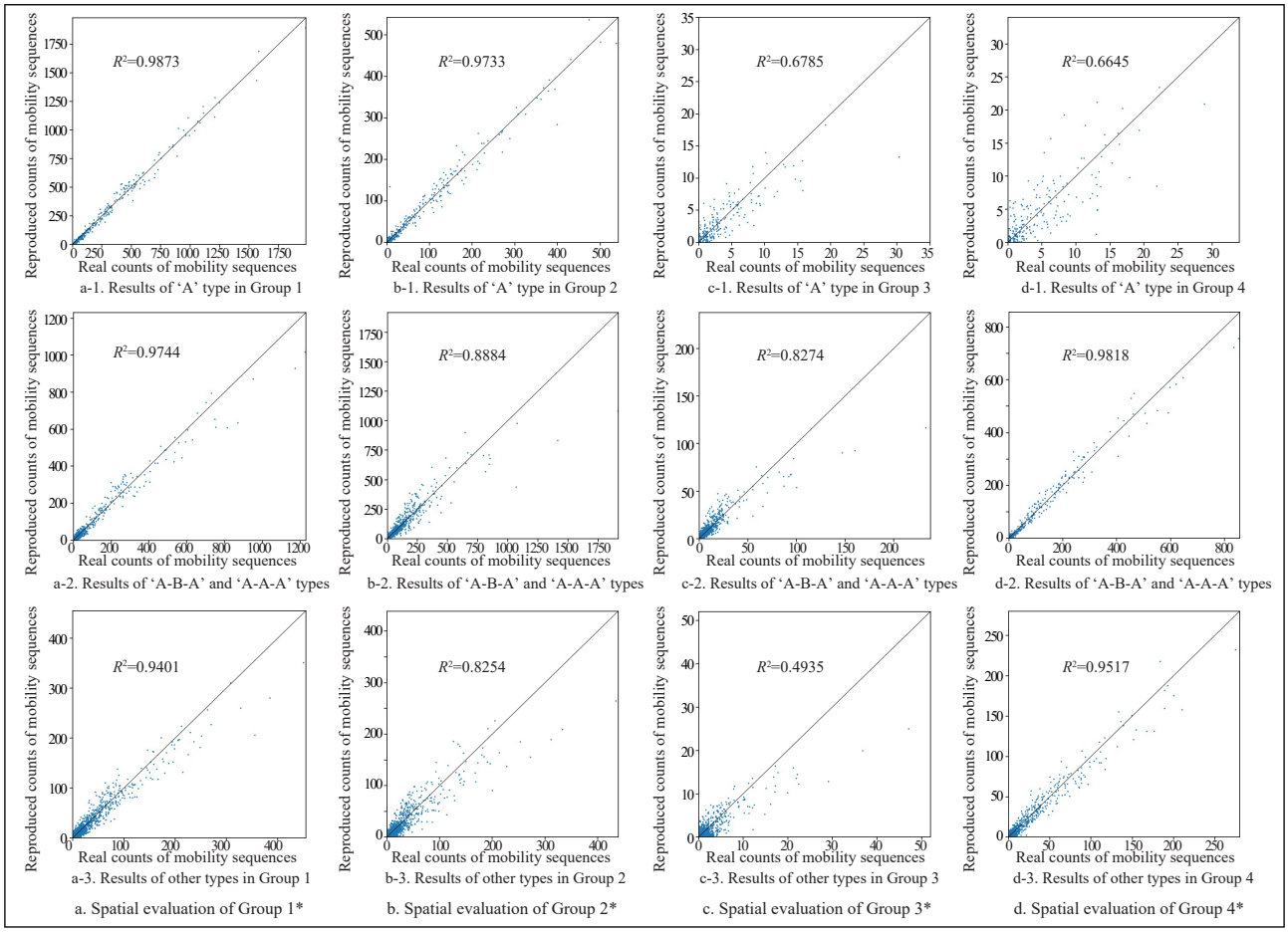

Fig. 8. Spatial evaluation of each group. * Groups details are in Table 1

Group 1 and Group 4 (see Table 1), the spatial attributes of reproduced mobility sequences are similar to those of the real mobility sequences; for Group 2 and Group 3, the '*A*' type is well-reproduced spatially, while the region sequences of other types with large counts are generated fewer than real counts. This may be owing to the differences between next-status probabilities for frequent region sequences (with large counts) and the next-status probabilities for less frequent region sequences. However, we treat all these region sequences as the same type of current mobility sequences in our model. Overall, our method can reproduce the spatial attributes well. The results suggest that people who have the same type of mobility sequences are likely to have the same future status, no matter when they had each trip and where they have visited.

4.2 Temporal similarity

Temporal characteristics of mobility sequences suggest when trips happen. Mobility sequences with more than one trips are very complicated. We compare the temporal similarity only for '*A-B-A*' and '*A-A-A*' types of mobility sequences with two trips, which cover a large proportions of people in the research (about 50%, see Fig. 9). For each group, we do the following steps for the original and reproduced datasets for these people:

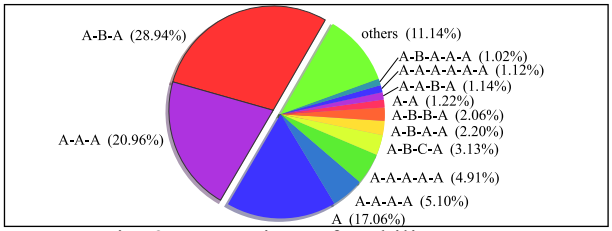

Fig. 9. Proportions of mobility types.

Step 1. For any target person *a*, we get its region sequence, and the times of the two trips, $T_a=(T_a^1,T_a^2)$.

Step 2. For people (*a*, *b*, ...) whose mobility sequences belong to the same region sequence *R*, we have a collection of times of the two trips of these people, $T^R=\{(T_a^1,T_a^2),(T_b^1,T_b^2),\ldots\}$.

Step 3. The collection of times of trips, T^R, is plotted

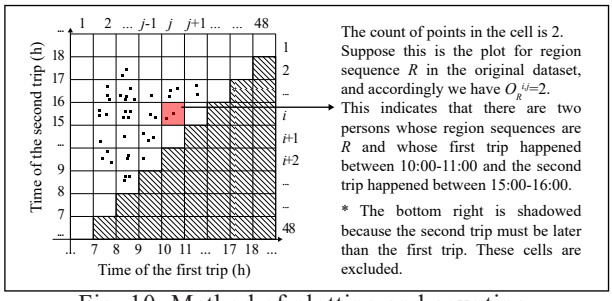

Fig. 10. Method of plotting and counting.

on a 2-D space where the *x*-axis is the time of the first trip and the *y*-axis is the time of the second trip. One point for one mobility sequence (Fig. 10).

Step 4. Cut the 2-D space into 1h-by-1h cells and count the number of points in each cell (also Fig. 10).

For region sequence R, $O_R^{i,j}$ is the number of points in cell in the i-th row and the j-th column for the original dataset and $G_R^{i,j}$ is the number of points in the same cell for the reproduced dataset. $O_R^{i,j}$ indicates the number of people in the original dataset whose region sequence is R and whose first trip happened in the i-th time interval and the second trip happened in the j-th time interval. $G_R^{i,j}$ indicates the same in the generated dataset. If $O_R^{i,j}$ and $G_R^{i,j}$ are close for any i, j, then the temporal attributes of mobility sequences of R type are well-reproduced. For each group of people, $O_R^{i,j}$ and $G_R^{i,j}$ for all Rs are shown in one plot (Fig. 11). We can notice that for Group 1 and Group 4, the temporal attributes are well-reproduced, while the numbers of points in favored time-cells (with large counts) are underestimated for the other two groups. The inaccuracy should be caused by differences between the next-status probabilities for various region sequences, the same reason as is introduced in section 4.1. Overall, it is surprising that the temporal attributes are reproduced in such high accuracy. The results suggest that the probability of having a trip at a certain time can be approximated only using the information of the type of current mobility sequences, regardless of previous locations and times of trips.

For region sequences with n trips, we need an n-dimensional space to plot the times of trips and as a result, cells are also n-dimensional. However, the number of cells increases exponentially as n grows and the count of points in each cell is too few for comparison. Thus, we only take into account the '*A-B-A*' and '*A-A-A*' types of mobility sequences.

5. Conclusions and discussion

5.1 Conclusions

In this research, we propose a method of reproducing human mobility based on the idea that people who have the same type of current mobility sequences are likely to behave in similar patterns in the future. In the framework of the model, individual status in the next time slot is chosen based on the type of current mobility sequences. If the next status is traveling to an unvisited region, the location will be selected from the time-varying transition matrix in a Markov process. The method can successfully reproduce the spatial and temporal attributes of human mobility. The results show that individual future status and mobility can be inferred based on the type of current mobility sequences of the person, regardless of where exactly the person traveled and when trips happen.

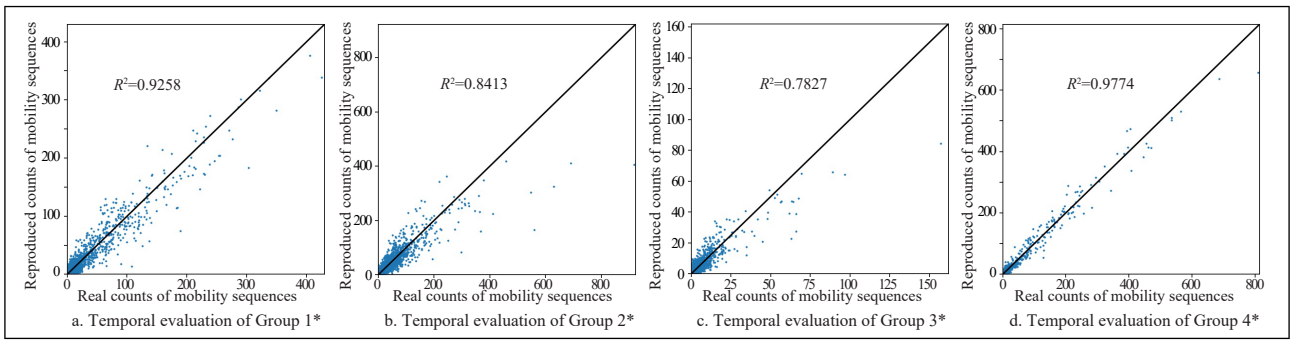

Fig. 11. Temporal evaluation of each group.

* Groups details are in Table 1

5.2 Discussion

Existing literature has demonstrated many methods of estimating transition probability matrices using various data sources. However, reproducing individual mobility using a simple Markov Chain process is far from perfect. This study introduces a method of reproducing human mobility using the information of the type of current mobility sequences. Our method is a tool to infer the spatial and temporal attributes of individual mobility from knowledge about aggregated mobility. In this research, we use data collected from Tokyo metropolitan area in 2008, and we will use data from other years and areas and see if the method still functions. We will also make attempts to simplify the current model in the future.

References

[1] González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958

[2] Calabrese F, Diao M, Di Lorenzo G, Ferreira J, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies 26:301–313. https://doi.org/10.1016/j.trc.2012.09.009

[3] Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6(1):8166. https://doi.org/10.1038/ncomms9166

[4] Axhausen KW, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T (2002) Observing the rhythms of daily life: A six-week travel diary. Transportation 29(2):95–124. https://doi.org/10.1023/A:1014247822322

[5] Hasan S, Zhan X, Ukkusuri SV (2013) Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp ’13. ACM Press, Chicago, Illinois, p 1

[6] Osaragi T, Kudo R (2020) Enhancing the Use of Population Statistics Derived from Mobile Phone Users by Considering Building-Use Dependent Purpose of Stay. In: Kyriakidis P, Hadjimitsis D, Skarlatos D, Mansourian A (eds) Geospatial Technologies for Local and Regional Development. Springer International Publishing, Cham, pp 185–203