Eur. Phys. J. Plus (2024) 139:403
https://doi.org/10.1140/epjp/s13360-024-05142-x

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

Unveiling realistic mobility patterns with home–origin–destination data aggregation

Yunhan Du[1,a], **Takaaki Aoki**[2,b], **Naoya Fujiwara**[1,3,4,5,c]

[1] Graduate School of Information Sciences, Tohoku University, 6-3-09 Aoba, Aramaki, Aoba, Sendai, Miyagi 980-8579, Japan
[2] Graduate School of Data Science, Shiga University, 1-1-1, Banba, Hikone, Shiga 522-8522, Japan
[3] PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan
[4] Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan
[5] Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

Received: 28 December 2023 / Accepted: 29 March 2024

Abstract The availability of increasingly abundant mobility data in recent years has opened up new avenues for researchers to unravel human mobility patterns. Data aggregation methods have been introduced to gain a quantitative understanding of collective individual movements using these data. Nevertheless, the widely adopted origin–destination (OD) aggregation method for human mobility data lacks an essential piece of information: home location, which plays a vital role in characterizing individual movement patterns. In this study, we propose a novel data aggregation approach called home–origin–destination (HOD) with the aim of improving the accuracy of human mobility estimation. We compare the performance of various data aggregation methods for estimating population distribution. Our experimental results reveal more realistic mobility patterns when incorporating estimated home information, where individuals move out in the morning and return home before midnight. To further evaluate the effectiveness of the HOD approach, we conduct an entropy analysis to measure the unpredictability of human mobility. The HOD results exhibit lower entropy values than those in the other two cases, OD and home–destination (HD). These findings underscore the importance of incorporating home information in understanding and modeling human mobility. By leveraging the HOD data aggregation method, we can achieve more accurate population distribution estimates and capture the inherent dynamics of human movement

1 Introduction

Human mobility is a fundamental aspect of the society. It serves as a proxy for human urban activities, shaping the spatial structure of cities [1–3] and providing insights into the estimation of transportation and the development of novel infrastructure [4, 5]. Furthermore, it is closely related to various social issues, including economic questions and living conditions. This encompasses a range of domains including, but not limited to, epidemic spread estimation [6–8], traffic engineering [9–11], urban planning [12], and emergency management [13–15]. Consequently, a qualitative understanding of human mobility patterns holds profound significance in empowering urban planners to design habitable and sustainable cities.

The increasing availability of human mobility data, including mobile phone records [16–18], global positioning system (GPS) tracks [19–21], and social media data [22–25], has facilitated extensive research efforts aimed at understanding human mobility patterns. The prediction of human mobility patterns using individual-level data is often challenging due to large data size and privacy considerations.

Consequently, the aggregation of raw data into an origin–destination (OD) format has become prevalent, and OD data are widely used [26–28]. Extensive efforts have been made to propose methodologies for estimating reliable OD matrices from available data [29–31], as well as general mobility models for predicting mobility flow between areas [32–35]. The use of OD data has several advantages, including important spatial information regarding the origin and destination of a trip; reduced data size compared with raw data, particularly individual-level data; and a simple structure that allows for a wide range of applications, including network analysis.

However, traditional OD data are limited because they do not include home location, which is important in showing an individual's access to opportunities, services, and amenities, as well as further shaping their mobility patterns. Previous studies have overlooked the importance of considering home location. Some studies have focused on estimating human mobility based on Markov processes [36, 37], which determine individuals' next place solely based on their current location, independent of previous places. In those

[a] e-mail: du.yunhan.q8@dc.tohoku.ac.jp (corresponding author)
[b] e-mail: takaaki.aoki.work@gmail.com
[c] e-mail: naoya.fujiwara@tohoku.ac.jp

Published online: 10 May 2024

studies, return-home trips were not considered. In non-Markov models, it is sometimes assumed that an individual might return to a previously visited place based on the frequency [38, 39]. However, despite the recognized importance of accurately identifying home locations for comprehensive data analysis, current OD methodologies lack a specific approach to aggregating data that includes home information.

In this study, we propose a novel method for data aggregation, called home–origin–destination (HOD) data aggregation, in addition to the conventional OD format and the home–destination (HD) format, which is occasionally used as an alternative to the OD format. We compare the estimated population distributions obtained from OD, HD, and HOD data. Specifically, we analyze human mobility based on a home-specific Markov process for HOD and OD cases, and a random process for HD case where the origin information is missing. Our findings demonstrate that the HOD and HD aggregation outperformed OD aggregation where home information is excluded. The absence of home information in the OD data may lead to an overestimation of the impact of human mobility. To measure the unpredictability of human mobility, we employed Shannon's entropy and demonstrate the HOD aggregation yield the best results. These findings underscore the crucial role of home information in obtaining an accurate understanding of human mobility patterns and their impact. Additionally, by comparing entropy values across different years, we observe that people tend to visit fewer places after the COVID-19 pandemic.

2 Data

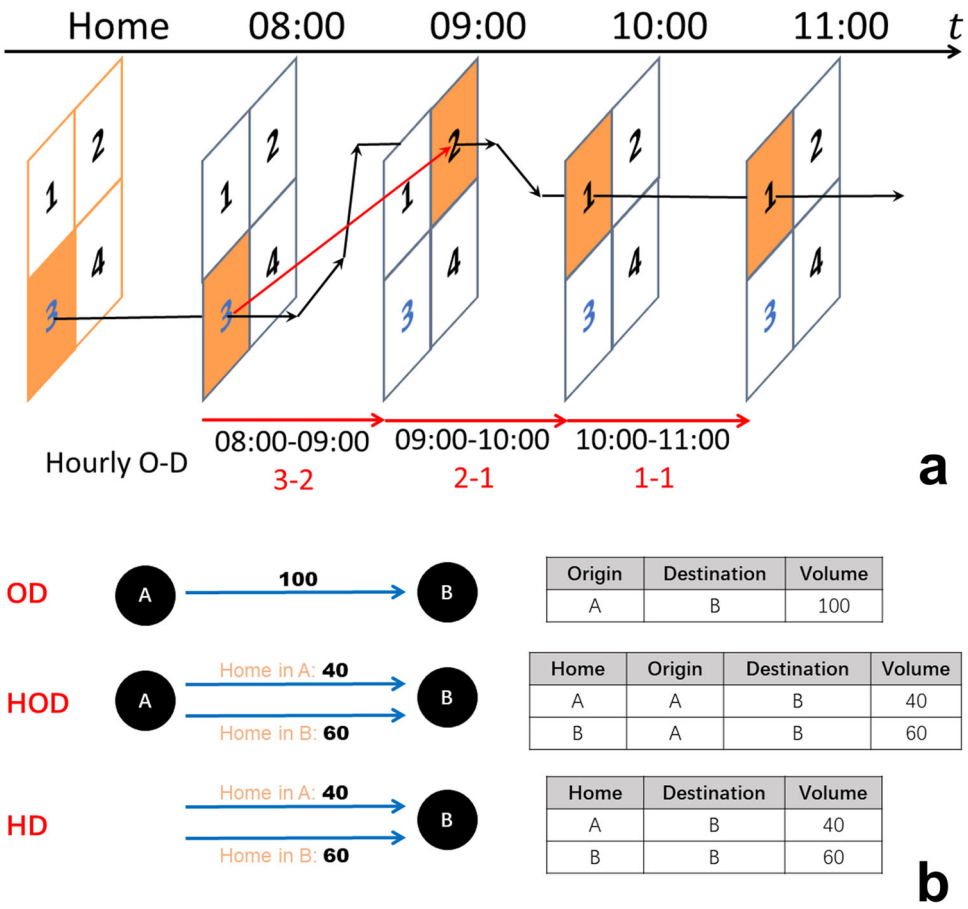
Our study is based on a set of GPS data, "LocationMind xPop", provided by the Japanese company LocationMind Inc. [40] The "LocationMind xPop" data refer to human mobility data collected through individual location information shared by mobile phone users who have provided their consent. This dataset is made available by NTT DOCOMO, Inc., Japan's largest cellular service provider. To safeguard privacy, the data are processed collectively and statistically by NTT DOCOMO, Inc. The original location data consist of GPS coordinates (latitude and longitude) sent at intervals as short as 5 min. Importantly, this dataset does not contain any information that can be used to identify specific individuals. The areas of interest are Tokyo and surrounding areas, which are divided into 2 km by 2 km meshed grids based on the grid square partition introduced by the Statistics Bureau of Japan [41]. Each grid is denoted by a 9-digit code. In this study, approximately 12 hundred populated grids are investigated.

The dataset consists of approximately 4.4 million records in the year 2019 and 4.7 million records in 2021, each detailing the hourly volume of mobility flows for each home–origin–destination set, indicating the number of individuals visiting from an origin to a destination, given their home grid. These volumes are derived from the averaged flows of specific hours across full days of May and June for either the years 2019 or 2021. Additionally, each data record is annotated with a flag indicating whether the observed average pertains to weekdays or weekends. For this study, we specifically selected the weekday subset to focus on analyzing the characteristics of daily commuting trips. This subset encompasses approximately 2.7 million records in 2019 and 3.0 million records in 2021. The data capture origin and destination information on an hourly basis, which differs from real-case OD data that accurately represents the specific origins and destinations of individual trips. However, it is worth noting that hourly OD data are still capable of illustrating the continuous movement of people and are sufficient for characterizing daily trips.

The home information in the data is estimated by the data provider using the following procedures. Initially, the GPS logs for each user, which are records of geographical positions captured by mobile phones collected over approximately one month, are mapped onto the 2 km by 2 km meshed grids mentioned previously. Following this mapping, the frequency distribution on these spatial grids for each user is generated with a bin width of 1 h. These logs are activated by smartphone movement, leading to fewer GPS records during inactive periods, such as night hours, when individuals are likely to be asleep. Frequency increases during the day and decreases at night for typical daytime workers, with the reverse pattern for night-shift workers. Subsequently, the frequency distributions for all users are clustered using the K-means algorithm, where the number of clusters k is set to 12. Notably, this method results in the formation of similar clusters that are generally narrowed down to five distinct patterns: the daytime pattern, involving individuals commuting or traveling in the morning, being active during the day, and returning home in the evening; the nighttime pattern, characterized by inactivity during the day and activity at night; the morning pattern, similar to the daytime pattern but with more active mornings; the evening pattern, akin to the daytime pattern but with more evening activities; and the stay-at-home pattern, with few trips, typical of homemakers or those working from home. Based on this clustering, users are further categorized, and then the inactive period for each user is identified. Finally, the home of the user is estimated as the location with the highest GPS log frequency during the inactive period.

The process of population magnification is undertaken because the number of mobile phone users in the data does not encompass the entire population. The magnification coefficient for each grid is derived by comparing the number of samples whose estimated home is in this grid with the actual population as per the census data. Owing to privacy concerns, any grid with a residence number less than a certain threshold is excluded from the data. This exclusion results in discrepancies between the populations in the data and the census data. The magnification of data for both 2019 and 2021 utilized the same benchmark, the Japan Population Census 2015 [42]. However, it is validated that the results in this work are not driven by the grid-wise population changes (see Supplementary Information).

The underground map layout in Figs. 2, 4 and 5 is based on map tiles by Stamen Design, under CC BY 4.0. Data by OpenStreetMap, under ODbL. The results in Figs. 2, 3, 4 and 5 are obtained by using data "LocationMind xPop (c) LocationMind Inc".



Origin	Destination	Volume
A	B	100

Home	Origin	Destination	Volume
A	A	B	40
B	A	B	60

Home	Destination	Volume
A	B	40
B	B	60

Fig. 1 **a** Individual's hourly movements. Each of the four polygons represents distinct meshed grids. The trajectory of an individual, whose estimated home is located in grid 3, is delineated by black arrows. While multiple mobile phone records can exist within each hourly segment, data are aggregated into hourly movements. For instance, within the 8AM to 9AM interval, the individual move from grids 3 to 4 and subsequently to 2. However, the OD information is documented as 3-2, as indicated by the red arrow. **b** A simple example of the OD, HD, and HOD data format in a 1-h interval

3 Methods

3.1 Time discretization

In this study, we investigate human movements based on discrete time, represented by a specific hour denoted by a non-negative integer t. As illustrated in Fig. 1a, an individual's movements are recorded through interpolation within a 1-h time window, which captures only the starting and ending grids during this hour.

3.2 Data aggregation methods

The traditional OD aggregation method primarily focuses on identifying the two important spatial factors of trips: origins and destinations. This involves recording the volume of mobility flow for each OD pair. For instance, in Fig. 1b, suppose there are 100 individuals traveling from location A at time t, to location B at $t + 1$, a single-row table can summarize these trips, incorporating information on the origin, destination, and volume.

However, as argued previously, we contend that home location plays an important role in characterizing human mobility patterns. Consequently, we propose including the home location, in addition to the OD pair, when aggregating the data. Specifically, we aim to ascertain how many of the 100 individuals reside in location A and how many reside in location B. In this example, HOD data can be summarized in a table in which each specific home location is associated with the corresponding OD pair. In detail, 40 individuals reside in home A and travel from A to B, whereas 60 reside in home B and also travel from A to B.

Another aggregation method, HD, considers only the home and destination information and does not include origin information. It can be viewed as a proxy for commuting trips because the origin is assumed to be the individual's home. Similarly, the HD table can be summarized as shown in Fig. 1b, where the origin information is missing.

Remarkably, the HOD format can generate both OD and HD data by aggregating the home or origin information because it contains a complete set of home, origin, and destination information. As shown below, even when only one additional feature, the home location, is included in the data, which does not significantly increase the data size compared with individual-level data, using HOD data for estimating the human movements proves to be more reliable than using OD data alone.

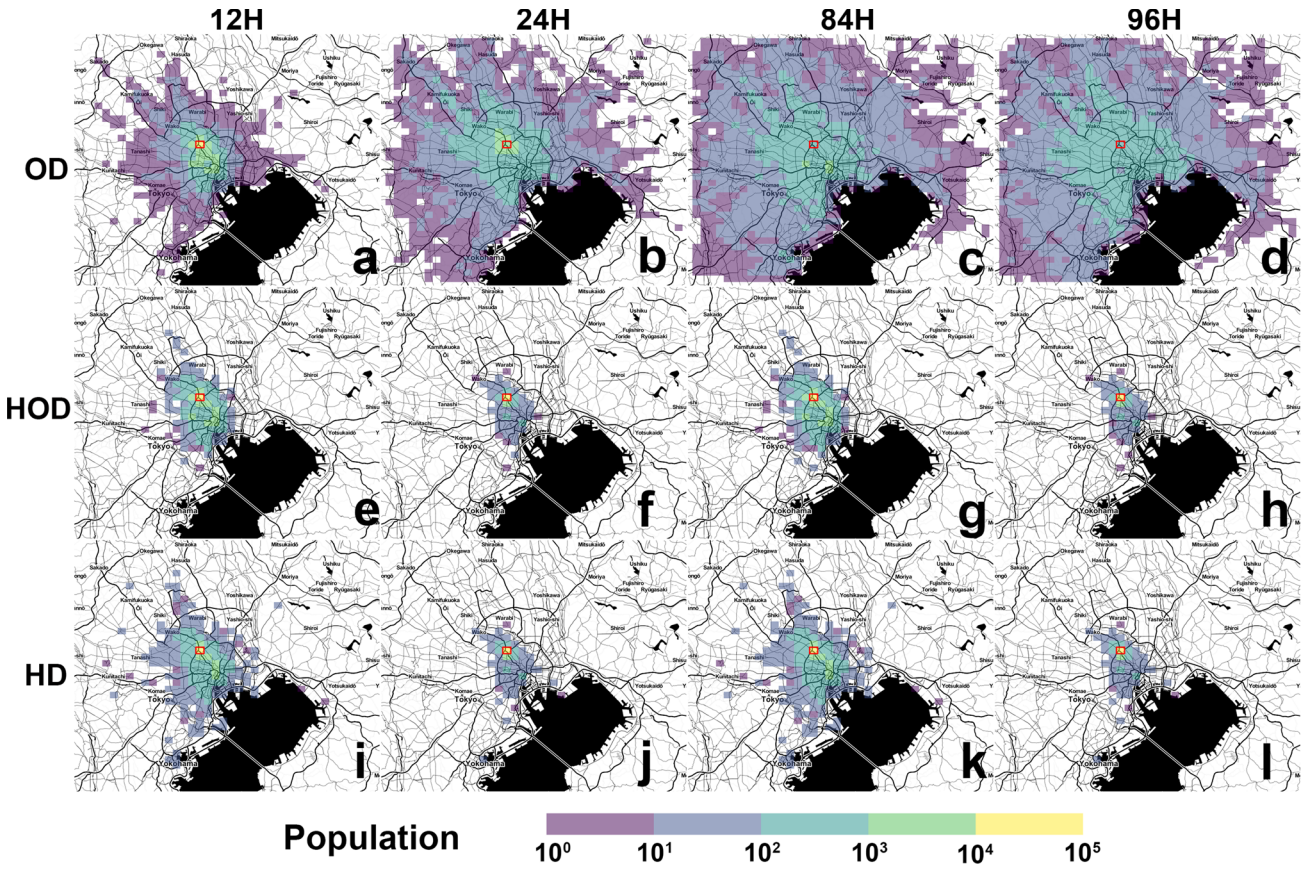


Fig. 2 Estimated population distribution. The population distributions in 12, 24, 84, and 96 h later in the OD (**a**–**d**), HOD (**e**–**h**), and HD (**i**–**l**) cases are shown. The black lines correspond to big streets, railways, and subways. The black areas are rivers, inland lakes, and the sea. Initially, only the home location (hinted by the red square) is populated. The colors represent the population volume in each 2 km by 2 km grid square at a certain hour. The underground map layout is based on map tiles by Stamen Design, under CC BY 4.0. Data by OpenStreetMap, under ODbL

3.3 Transition matrix

To estimate the population distribution using OD, HD, and HOD data, we adopt the following procedure to extract the transition matrix from the data at a specific hour t, represented by $\mathbf{T}(t)$.

Regarding the OD data, we construct a matrix denoted as $\mathbf{N}^{\mathrm{OD}}(t)$, where each element $n_{ij}^{\mathrm{OD}}(t)$ represents the number of trips that start from grid j at hour t and end at grid i at hour $t+1$. To obtain the transition matrix $\mathbf{T}^{\mathrm{OD}}(t)$, we normalize the columns of matrix $\mathbf{N}^{\mathrm{OD}}(t)$, such that each element of $\mathbf{T}^{\mathrm{OD}}(t)$ is defined as follows:

$$t_{ij}^{\mathrm{OD}}(t) = \frac{n_{ij}^{\mathrm{OD}}(t)}{\sum_k n_{kj}^{\mathrm{OD}}(t)}. \tag{1}$$

Similarly, for HD data, we construct a matrix $\mathbf{N}^{\mathrm{HD}}(t)$, where each element $n_{ij}^{\mathrm{HD}}(t)$ represents the number of individuals whose homes are located in grid j and who visit grid i at hour t. Then, an element of the transition matrix $\mathbf{T}^{\mathrm{HD}}(t)$ is defined as

$$t_{ij}^{\mathrm{HD}}(t) = \frac{n_{ij}^{\mathrm{HD}}(t)}{\sum_k n_{kj}^{\mathrm{HD}}(t)}. \tag{2}$$

Regarding the HOD data, as it includes information on the estimated home grid in each OD pair, we define a flow matrix specific to each estimated home grid h as $\mathbf{N}^{\mathrm{HOD}}(h,\, t)$, whose element $n_{ij}^{\mathrm{HOD}}(h,\, t)$ represents the mobility flow whose estimated home grid is h, visiting from grid j to i at time t. In this case, the transition matrix of the estimated home h is defined as

$$t_{ij}^{\mathrm{HOD}}(h,t) = \frac{n_{ij}^{\mathrm{HOD}}(h,t)}{\sum_k n_{kj}^{\mathrm{HOD}}(h,t)}. \tag{3}$$

In the following analysis, we assume that transition matrix $\mathbf{T}(t)$ is periodic over a 24-h time interval, as expressed by the relationship $\mathbf{T}(t) = \mathbf{T}(t+24)$ for any non-negative hour t. This assumption is consistent with our intent to examine average mobility patterns. It is worth noting that the periodicity assumption is usually not necessary when the focus shifts to studying anomalous mobility, such as evacuation from a major earthquake.

3.4 Random walk

By extracting the transition matrices from the data as described above, it is possible to estimate the population distribution in future hours. In this study, we adopt a random walk methodology to replicate mobility patterns, where the population distribution is estimated using a Markov process for HOD and OD cases, and a random process in which the next movement does not depends on the previous one for HD case. By employing this approach, we gain insights into the spatial dynamics of a population, particularly its movement within urban areas over time. In this framework, the probable movements of individuals at each step are based on their current states, with no dependence on the previous positions. Although it is acknowledged that this memoryless model cannot fully encompass the complexities of real-world population dynamics, it serves as a valuable simplification. This allows us to concentrate on analyzing distinct population dynamics across various scenarios, including OD, HD, and HOD.

As our focus is particularly on whether home information is available, we examine the time-varying distribution of the population from a certain estimated home grid h. Specifically, we start from an initial condition in which only grid h is populated such that the population distribution is defined as follows:

$$\hat{\mathbf{q}}^{\mathrm{HOD}}(h,0) = \hat{\mathbf{q}}^{\mathrm{OD}}(h,0) = \hat{\mathbf{q}}^{\mathrm{HD}}(h,0) = (0,\ldots,0,q_{h,0},0,\ldots,0)^{\mathrm{T}}, \tag{4}$$

where $q_{h,0}$ is taken as the total number of people whose estimated home grid is h from the data, and the superscript T denotes the transpose of a vector.

In the OD and HOD cases, given the estimated population distribution at any hour t, the population distribution at the next hour can be determined by multiplying the transition matrix by the estimated population distribution at the current hour. The equations for the OD and HOD cases are as follows:

$$\hat{\mathbf{q}}^{\mathrm{OD}}(h,t+1) = \mathbf{T}^{\mathrm{OD}}(t)\hat{\mathbf{q}}(h,t), \tag{5}$$

$$\hat{\mathbf{q}}^{\mathrm{HOD}}(h,t+1) = \mathbf{T}^{\mathrm{HOD}}(h,t)\hat{\mathbf{q}}(h,t). \tag{6}$$

For the HD case, because the origin information is not provided, the starting place is always considered as home. Given the initial population distribution $\mathbf{p}(h,\,0)$, the population distribution at hour t is estimated as:

$$\hat{\mathbf{q}}^{\mathrm{HD}}(h,t+1) = \mathbf{T}^{\mathrm{HD}}(t)\hat{\mathbf{q}}(h,0). \tag{7}$$

4 Results

4.1 Estimation of population distribution

Figure 2 presents the results of random walk experiments discussed above as an example to illustrate the estimated population distribution. The initially populated area, $h = 533945865$ (indicated by the red square), is located near Ikebukuro, one of Tokyo's city centers. The population value is taken as the total number of people whose estimated home is located in grid h based on the HOD data, which is 61,666 in this case.

The OD results reveal a discernible trend of population spread within the area of interest over a 96-h period, as shown in Fig. 2a–d. The absence of home information in the OD data results in the limited predictive capabilities in determining return-home trips. Even if all individuals share the same estimated home grid, forecasting their journeys back home remains difficult. In contrast, the HD and HOD scenarios exhibit distinct patterns of expansion at hour 12 and 84, corresponding to the noon, followed by contractions at hour 24 and 96, representing the midnight. These observations indicate a recurring movement pattern in which individuals leave during the day and return home in the evening. These results provide evidence that relying solely on the OD information may lead to an overestimation of the impact of human mobility. Incorporating estimated home information in the HOD and HD scenarios enables a more comprehensive understanding of population dynamics, highlighting the importance of considering human mobility patterns and home-based movements in urban areas.

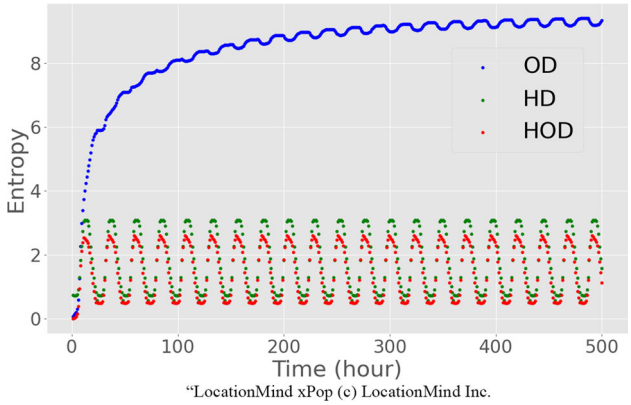


Fig. 3 Hourly entropy change of population distributions in a 500-h period for the OD, HD, and HOD cases. The initial condition of population distribution is the same as that in Fig. 2

4.2 Entropy of population distribution

We incorporate Shannon entropy as a measure to quantify the information loss and degree of unpredictability of human mobility [43–45]. Given a population distribution estimated from the initial condition that only grid h is populated, $\hat{\mathbf{q}}(h, t) = (\hat{q}_{h,1}, \hat{q}_{h,2}, \ldots, \hat{q}_{h,M})$, where M denotes the total number of grids, we define the Shannon entropy as follows:

$$H(h,t) = -\sum_{j=1}^{M} \frac{q_j(h,t)}{\sum_{k=1}^{M} q_k(h,t)} \log_2 \frac{q_j(h,t)}{\sum_{k=1}^{M} q_k(h,t)}. \tag{8}$$

In this measure, a higher entropy value indicates greater unpredictability of human mobility, implying a larger number of expected locations that individuals might visit. Figure 3 presents the average entropy values of the estimated population distributions across all grids, $\sum_{h=1}^{M} H(h, t)/M$, for the OD, HD, and HOD scenarios. The HOD and HD scenarios exhibit consistent periodic variations from the initial stages. Specifically, during nighttime hours, the entropy values are lower, reflecting the predictable trend of individuals returning home. Moreover, the entropy of the HOD scenario is lower than that of HD scenario. In the OD scenario, the entropy values increase at least over the 200-h observation period, eventually reaching a steady-state with higher entropy than that of the HOD and HD scenarios.

As shown in Fig. 3, the absence of estimated home information in the OD data results in the failure to capture returning trips, leading to high entropy values. During the evening hours, when individuals are expected to start returning home, the entropy values are lower in the HOD and HD cases than in the morning hours. However, in the OD case, the entropy values remain high, indicating a counter-intuitive scenario in which individuals are not considered to return home but are assumed to explore further areas instead. These findings further highlight the critical importance of incorporating estimated home information to achieve more accurate predictions of human mobility.

4.3 Spatial distribution of the maximum entropy

Upon observing the dynamics of entropy values within a singular grid in Fig. 3, several questions emerge regarding the comparative analysis of entropy values across all regions, focusing on identifying which locations exhibit relatively higher or lower entropy values in comparison to others and exploring the formation of specific spatial patterns among these values. Because of the temporal variability of entropy values for each location, we adopt the methodology of selecting the maximum entropy value for each grid as a standardized measure for comparison. Specifically, since a 24-h oscillation is observed in the case of HOD and HD for the first simulation day in Fig. 3, the maximum entropy value during the initial 24-h period was chosen for analysis. The spatial distribution of the maximum entropy values with respect to the home locations is shown in Fig. 4, where colors represent the maximum of the time-dependent entropy of the population distribution estimated from the initial condition that only this grid is populated, within a 24-h period, represented by $\max_{t\in(0,24]} H(h, t)$.

Figure 4 demonstrates that the maximum entropy in the OD case is higher than that in the other two cases, encompassing all grids within the designated area of interest in both 2019 and 2021, which is consistent with the results in Fig. 3. It is also validated that OD aggregation results in a higher entropy than the other two cases, while comparing the same hour in the simulation (see Supplementary Information). The maximum entropy quantifies the potential complexity that different aggregation methods can introduce. The high entropy values in the OD aggregation suggest an exaggerated number of potential places that individuals might visit. This indicates that predictions based solely on the OD data can overestimate the actual spatial extent of human movement. In contrast, the entropy values corresponding to the population distribution estimated using HOD data exhibit the lowest values. This implies a higher level of predictability for the HOD case, highlighting the global significance of home locations in determining the predictability of human mobility patterns.

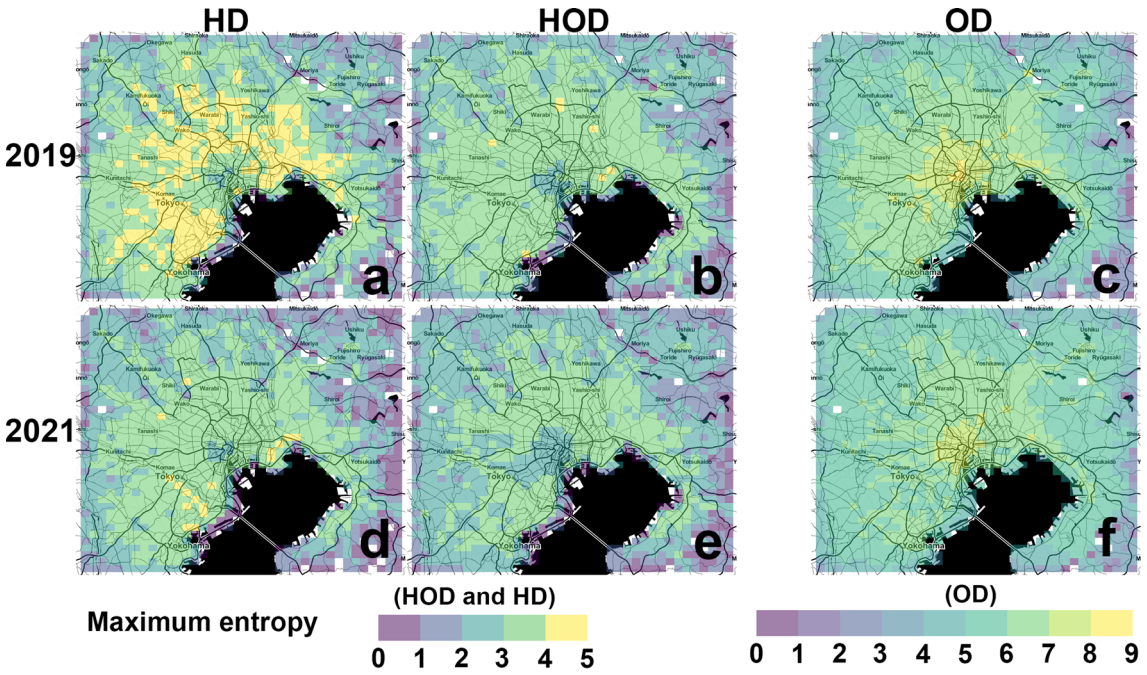


Fig. 4 Spatial distribution of maximum entropy values in a 24-h period for the HD (**a**), HOD (**b**), and OD (**c**) in 2019, and HD (**d**), HOD (**e**), and OD (**f**) cases in 2021. The underground map layout is based on map tiles by Stamen Design, under CC BY 4.0. Data by OpenStreetMap, under ODbL

In all cases examined, the peripheral areas exhibit relatively lower entropy values. This can be attributed to the lower population density and the presence of established trip routines among the individuals residing in these areas. In the OD case (Fig. 4c, f), the entropy values generally increase from the periphery toward the center. However, in the HOD and HD cases (Fig. 4a–d), the central areas display relatively lower entropy values than their immediate surroundings. This discrepancy arises because the maximum entropy in OD case is in the midnight, due to the continuous increase in the entropy value for the first several hours, while HOD and HD predict the recurrent movements of people and the maximum entropy appears in the daytime.

Furthermore, as shown in Fig. 4d, relatively higher entropy values are observed alongside railway or subway lines in the HD case in 2019. This observation suggests that areas near railway or subway lines exhibit a higher probability of individuals moving to various destinations, which aligns with an intuitive understanding of the influence of transportation infrastructure on human mobility patterns.

Moreover, across all three cases, the entropy values in 2021 are lower than those in 2019. This suggests that individuals tend to visit fewer locations during the post-COVID-19 period than during the pre-pandemic era. This trend can be attributed to several factors, including changes in people's attitudes toward travel after the COVID-19 pandemic and the increased prevalence of remote work facilitated by the enhanced remote working environment, including the introduction of various online meeting applications.

4.4 Population and maximum entropy: fitting and residuals

Next, we seek to understand what brings the mobility patterns indicated by the maximum entropy analysis depicted in Fig. 4. The size of the population stands out as a factor contributing to the variability in movement patterns observed. It is intuitive to posit that larger populations are associated with more complex behavior patterns, which, in turn, are likely to lead to higher entropy values. To segregate the impact of population size on maximum entropy, a regression analysis was conducted, correlating the maximum entropy, $\max_{t\in(0,24]} H(h, t)$ for each home grid h, and its respective population size, as shown in Fig. 5. Throughout the evaluated HOD, HD, and OD cases, maximum entropy values correlate with the population size through logarithmic functions. The correlations establish a foundational understanding of how population size influences maximum entropy values.

The objective extends to identifying regions where maximum entropy deviates from population-based predictions. By employing residual plots in Fig. 5, we spotlight regions characterized by significant discrepancies, revealed either as overestimations or underestimations when compared to population forecasts. Interestingly, an observed negative bias in the fitting for HOD and HD cases within central regions suggests a diminished level of daytime mobility than what would be anticipated based on population size alone. This observed reduction in movement could be attributed to a phenomenon where individuals both live and work in the same central grid, thereby reducing the need for daytime travel to other areas.

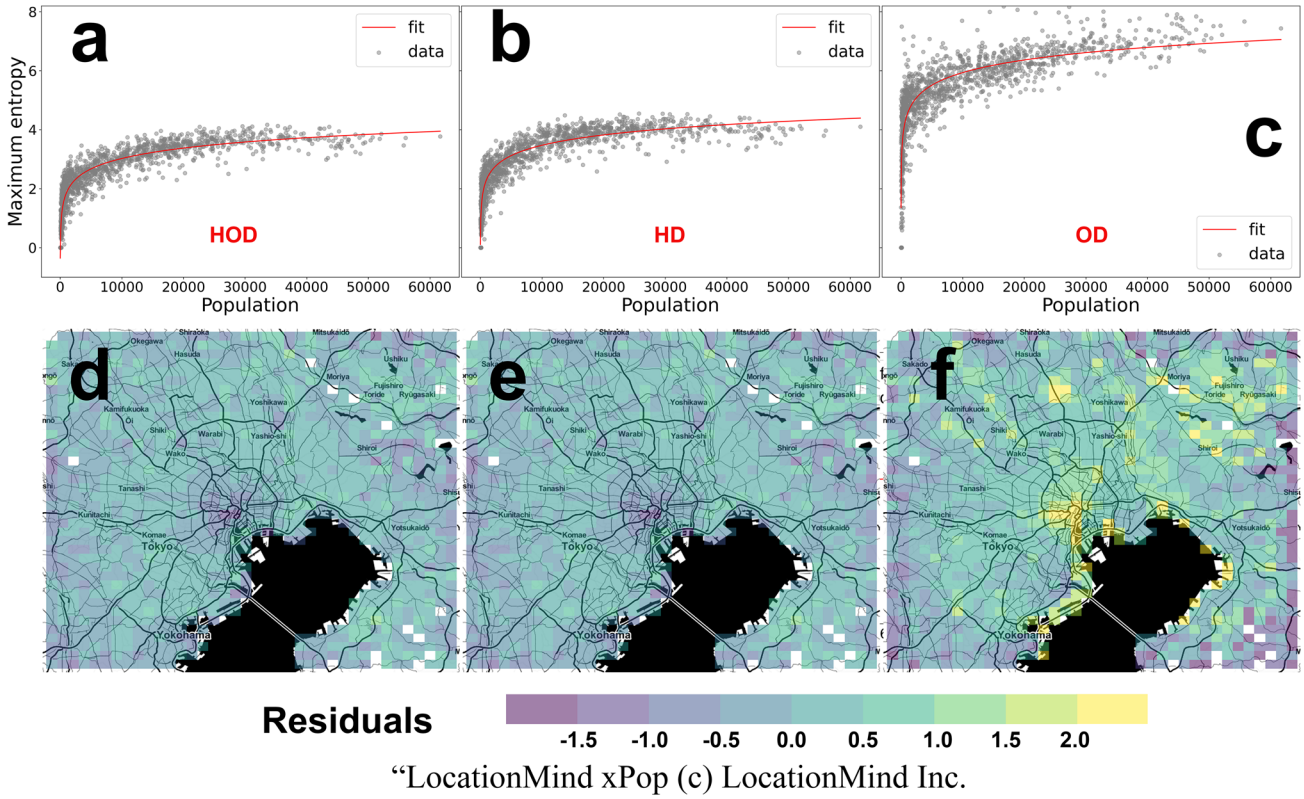


Fig. 5 Correlation between the maximum entropy and population. The logarithm functions represented by red curves are fitted into the relationship between maximum entropy and population in each grid in the HOD (**a**), HD (**b**), and OD (**c**) cases. The correlation coefficients are 0.706, 0.720, and 0.743 for the OD, HD, and HOD cases, respectively. The spatial distribution of residuals is plotted for HOD (**d**), HD (**e**), and OD (**f**) cases. Colors in each grid represent the residual values

5 Discussion

This study aimed to address the limitations of the widely adopted OD aggregation method for human mobility data by introducing a novel approach called the HOD aggregation method. The HOD method incorporates the estimated home location information, which is crucial for understanding individual movement patterns and improving the accuracy of human mobility estimation.

A comparison of various data aggregation methods for estimating the population distribution revealed that our proposed HOD aggregation yielded a more realistic moving pattern than the OD aggregation. Furthermore, an entropy analysis was conducted to measure the unpredictability of human mobility and evaluate the effectiveness of different approaches. In comparison to the OD aggregation, the lower entropy values in the HOD aggregation suggest that incorporating estimated home information is crucial for the prediction of return-home trips. Compare with the HD aggregation, the lower entropy values in the HOD aggregation underscore the importance of considering origin information, even when analyzing individuals from the same residential area. This indicates that an individual’s current location also plays a pivotal role in shaping their movement patterns. In summary, by incorporating home, origin, and destination information, the HOD aggregation captures an additional layer of context compared with the OD or HD aggregation, which enhances our understanding of how individuals make trips, further reduces the uncertainty and improves the predictability of human mobility patterns.

It is worth noting that each aggregation method has distinct advantages tailored for different scenarios. OD aggregation is suitable when focusing on overall mobility flow, such as in assessing citywide traffic demand. HOD aggregation, although more computationally demanding, provides relatively precise predictions. Meanwhile, HD aggregation might be especially valuable in simulating epidemic spread. This strikes a balance between computational efficiency and the integrating of the origin data for a more accurate depiction of human mobility patterns.

From the comparison of mobility patterns between 2019 and 2021, there was a tendency for individuals to visit fewer places in 2021, which is attributable to the fact that the COVID-19 pandemic had a considerable influence on human mobility patterns. This implies that the variations reflected in the mobility data may not only enhance the current understanding of the spread and size of epidemics but can also provide insights into the social and economic impacts of the pandemic. Some studies have indicated that the pandemic could exert a long-lasting impact on our society, including reshaping working styles [46] and reducing income diversity [47]. Incorporating home information may improve our knowledge of these aspects and reveal the influence of the pandemic. To

encapsulate the changes in mobility patterns induced by the pandemic, future studies could examine more recent HOD data to determine whether mobility patterns have reverted to the pre-pandemic state or remained altered.

A potential extension to this study lies in adopting a train-test split methodology to predict population distribution several hours into the future using different aggregation methods. While the current dataset, limited to the average of aggregated 24-h mobility flow, constrains our ability to conduct such predictive analysis, the examination of Fig. 2 suggests that the HOD aggregation may offer promising accuracy in population distribution predictions. Should the predictive potential of HOD aggregation be validated through the analysis of more granular mobility data collected over extended periods, it would significantly refine our aggregation approach, offering a more comprehensive metric for assessment beyond traditional methods.

A primary limitation of this study is the dependence of population estimation on a random walk process. Human mobility is inherently complex, and individual movements are multifaceted and challenging to predict. In this case, using a Markov process as the sole representation of these intricate behaviors is not a perfect approach to fully represent the nature of human mobility in the real world. However, previous studies have demonstrated the effectiveness of Markov models in predicting human movements, particularly in terms of the accuracy of next-location prediction [36, 37]. Therefore, they can serve as suitable simplifications when the primary focus is on comparing the outcomes of different aggregation methods. Future extensions of this research should incorporate home information into these Markov models or explore the integration of more realistic models for mobility pattern prediction.

In addition, there are limitations related to the data. First, the GPS sample did not cover the entire population in the area of interest, primarily because the data were sourced from mobile phone users. Furthermore, because the data were not randomly selected, some certain groups, such as older adults and low-income individuals, were less likely to use a mobile phone, potentially leading to biased results. Moreover, to protect the privacy of individuals, the data from grids with fewer than a certain threshold were truncated, resulting in the exclusion of nearly half of the users from the data. A thorough robustness check using less biased datasets represents an avenue for future research exploration.

Despite the limitations, this study provides valuable insights into human mobility research, with implications that extend beyond the academic realm. Policymakers, urban planners, and transportation authorities can benefit from the improved accuracy of the HOD aggregation method. The ability to accurately estimate the population distribution and understand human mobility patterns enables better resource allocation, infrastructure planning, and targeted interventions to enhance public services and optimize transportation systems. Moreover, the application of the HOD aggregation method is particularly valuable for the estimation of epidemic spread. As demonstrated in this study, traditional OD methods tend to overestimate the influence of human mobility on epidemic spread. Using the HOD method, researchers can effectively reduce the risk of misguided interventions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1140/epjp/s13360-024-05142-x.

Acknowledgements The authors would like to thank prof. Albert Díaz-Guilera and prof. Jürgen Kurths for the discussions and their helpful comments on this work. Y.D. is supported by JST SPRING, Grant Number JPMJSP2114. N.F. is supported by JSPS KAKENHI Grant Number JP21H03507 and JST PRESTO Grant Number JPMJPR21RA, Japan. This work additionally receives support from the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University, and involves the joint research with the Center for Spatial Information Science, the University of Tokyo (No. 1237).

Data availability statement The mobility flow data used in this study can be purchased from the Japanese company, LocationMind Inc. [40] (Contact form [48]). The product name is "LocationMind xPop".

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

1. M. Batty, *The New Science of Cities* (MIT Press, Cambridge, 2013)
2. M. Barthelemy, *The Structure and Dynamics of Cities* (Cambridge University Press, Cambridge, 2016)
3. L.M.A. Bettencourt, The origins of scaling in cities. Science **340**, 1438–1441 (2013)
4. J. Dios Ortúzar, L.G. Willumsen, *Modelling Transport* (Wiley, New York, 2011)
5. M. Barthelemy, Spatial networks. Phys. Rep. **499**, 1–101 (2011)
6. J.S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, Population flow drives spatio-temporal distribution of covid-19 inchina. Nature **582**, 389–394 (2020)
7. S. Hazarie, D. Sorianos, A. Arenas, J. Gómez-Gardeñes, G. Ghoshal, Interplay between population density and mobility in determining thespreadof epidemics in cities. Commun. Phys. **4**, 191 (2021)

8. S. Chang, E. Pierson, P.W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of covid-19 explain inequities and inform-reopening. Nature **589**, 82–87 (2021)
9. G. Krings, F. Calabrese, C. Ratti, V.D. Blondel, Urban gravity: a model for inter-city telecommunication flows. J. Stat. Mech: Theory Exp. **2009**, 07003 (2009)
10. Z. Huang, X. Ling, P. Wang, F. Zhang, Y. Mao, T. Lin, F.-Y. Wang, Modeling real-time human mobility based on mobile phone andtransportation data fusion. Transp. Res. Part C: Emerg. Technol. **96**, 251–269 (2018)
11. M.M. Vazifeh, P. Santi, G. Resta, S.H. Strogatz, C. Ratti, Addressing the minimum fleet problem in on-demand urban mobility. Nature **557**, 534–538 (2018)
12. M. Batty, The size, scale, and shape of cities. Science **319**, 769–771 (2008)
13. X. Lu, L. Bengtsson, P. Holme, Predictability of population displacement after the 2010 haitiearthquake. Proc. Natl. Acad. Sci. **109**, 11576–11581 (2012)
14. T. Yabe, K. Tsubouchi, N. Fujiwara, Y. Sekimoto, S.V. Ukkusuri, Understanding post-disaster population recovery patterns. J. R. Soc. Interface **17**, 20190532 (2020)
15. T. Yabe, K. Tsubouchi, N. Fujiwara, T. Wada, Y. Sekimoto, S.V. Ukkusuri, Non-compulsory measures sufficiently reduced human mobility in tokyoduring the covid-19 epidemic. Sci. Rep. **10**, 18053 (2020)
16. M.C. González, C.A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns. Nature **453**, 779–782 (2008)
17. S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, G. Pujolle, Estimating human trajectories and hotspots through mobile phone data. Comput. Netw. **64**, 296–307 (2014)
18. M. Schläpfer, L. Dong, K. O'Keeffe, P. Santi, M. Szell, H. Salat, S. Anklesaria, M.C. Ratti, G.B. West, The universal visitation law of human mobility. Nature **593**, 522–527 (2021)
19. D. Ashbrook, T. Starner, Using gps to learn significant locations and predict movement acrossmultiple users. Pers. Ubiquit. Comput. **7**, 275–286 (2003)
20. A. Lima, R. Stanojevic, D. Papagiannaki, P. Rodriguez, M.C. Gonázlez, Understanding individual routing behaviour. J. R. Soc. Interface **13**, 20160021 (2016)
21. A. Cuttone, S. Lehmann, M.C. Gonázlez, Understanding predictability and exploration in human mobility. EPJ Data Sci. **7**, 2 (2018)
22. P.A. Grabowicz, J.J. Ramasco, B. Gonçalves, V.M. Eguluz, Entangling mobility and interactions in social media. PLOS ONE **9**, 1–12 (2014)
23. M.G. Beiró, A. Panisson, M. Tizzoni, C. Cattuto, Predicting human mobility through the assimilation of social mediatraces into mobility models. EPJ Data Sci. **5**, 30 (2016)
24. Q. Wang, N.E. Phillips, M.L. Small, R.J. Sampson, Urban mobility and neighborhood isolation inamerica's 50 largestcities. Proc. Natl. Acad. Sci. **115**, 7735–7740 (2018)
25. M. Mazzoli, A. Molas, A. Bassolas, M. Lenormand, P. Colet, J.J. Ramasco, Field theory for recurrent mobility. Nat. Commun. **10**, 3895 (2019)
26. G. Varga, Z. Néda, Commuting patterns: the flow and jump model and supporting data. EPJ Data Sci. **7**, 37 (2018)
27. X.-Y. Yan, T. Zhou, Destination choice game: a spatial interaction theory on humanmobility. Sci. Rep. **9**, 9466 (2019)
28. T. Aoki, S. Fujishima, N. Fujiwara, Urban spatial structures from human flow by hodgekodairadecomposition. Sci. Rep. **12**, 11258 (2022)
29. M.G.H. Bell, The estimation of origin-destination matrices by constrainedgeneralised least squares. Transp. Res. Part B: Methodol. **25**, 13–22 (1991)
30. L. Alexander, S. Jiang, M. Murga, M.C. Gonózlez, Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol. **58**, 240–250 (2015)
31. L. Zhu, F.R. Yu, Y. Wang, B. Ning, T. Tang, Big data analytics in intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. **20**, 383–398 (2019)
32. H. Barbosa, M. Barthelemy, G. Ghoshal, C.R. James, M. Lenormand, T. Louail, R. Menezes, J.J. Ramasco, F. Simini, M. Tomasini, Human mobility: models and applications. Phys. Rep. **734**, 1–74 (2018)
33. G.K. Zipf, The p 1 p 2/d hypothesis: on the intercity movement of persons. Am. Sociol. Rev. **11**, 677–686 (1946)
34. S.A. Stouffer, Intervening opportunities: a theory relating mobility and distance. Am. Sociol. Rev. **5**, 845–867 (1940)
35. F. Simini, M.C. Gonózlez, A. Maritan, A.-L. Barabási, A universal model for mobility and migration patterns. Nature **484**, 96–100 (2012)
36. M. Chen, X. Yu, Y. Liu, Mining moving patterns for predicting next location. Inf. Syst. **54**, 156–168 (2015)
37. M. Yan, S. Li, C.A. Chan, Y. Shen, Y. Yu, Mobility prediction using a weighted markov model based on mobile userclassification. Sensors **21**, 1740 (2021)
38. C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility. Nat. Phys. **6**, 818–823 (2010)
39. E. Moro, D. Calacci, X. Dong, A. Pentland, Mobility patterns are associated with experienced income segregationin large us cities. Nat. Commun. **12**, 4633 (2021)
40. LocationMind Inc. Accessed 11 November 2023.https://locationmind.com/
41. Standard grid square and grid square code used for the statistics. Accessed 10 October 2023.https://www.stat.go.jp/english/data/mesh/02.html
42. Official Statistics of Japan. Accessed 1 March 2024. https://www.e-stat.go.jp/en
43. C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility. Science **327**, 1018–1021 (2010)
44. S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, M. Alava, Patterns, entropy, and predictability of human mobility and life. PLOS ONE **7**, 1–8 (2012)
45. X. Lu, E. Wetter, N. Bharti, A.J. Tatem, L. Bengtsson, Approaching the limit of predictability in human mobility. Sci. Rep. **3**, 2923 (2013)
46. Covid-19 pandemic continues to reshape work in America. Accessed 10 October 2023. https://www.pewresearch.org/social-trends/2022/02/16/covid-19-pandemic-continues-to-reshape-work-in-america/
47. T. Yabe, B.G.B. Bueno, X. Dong, A. Pentland, E. Moro, Behavioral changes during the covid-19 pandemic decreased incomediversity of urban encounters. Nat. Commun. **14**, 2310 (2023)
48. LocationMind contact form. Accessed 11 November 2023. https://locationmind.com/#contact-us