Metropolitan Scale and Longitudinal Dataset of Anonymized Human Mobility Trajectories

Takahiro Yabe[1,3,†,*], Kota Tsubouchi[2,†,*], Toru Shimizu[2], Yoshihide Sekimoto[3], Kaoru Sezaki[3], Esteban Moro[1,5], and Alex Pentland[1,4]

[1]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Yahoo Japan Corporation, Kioicho, Tokyo 102-8282, Japan
[3]Center for Spatial Information Science,the University of Tokyo, Kashiwa, Chiba 277-8568, Japan
[4]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain
[*]corresponding authors: Takahiro Yabe (tyabe@mit.edu) and Kota Tsubouchi (ktsubouc@yahoo-corp.jp)
[†]these authors contributed equally to this work

ABSTRACT

Modeling and predicting human mobility trajectories in urban areas is an essential task for various applications. The recent availability of large-scale human movement data collected from mobile devices have enabled the development of complex human mobility prediction models. However, human mobility prediction methods are often trained and tested on different datasets, due to the lack of open-source large-scale human mobility datasets amid privacy concerns, posing a challenge towards conducting fair performance comparisons between methods. To this end, we created an open-source, anonymized, metropolitan scale, and longitudinal (90 days) dataset of 100,000 individuals' human mobility trajectories, using mobile phone location data. The location pings are spatially and temporally discretized, and the metropolitan area is undisclosed to protect users' privacy. The 90-day period is composed of 75 days of business-as-usual and 15 days during an emergency. To promote the use of the dataset, we will host a human mobility prediction data challenge ('HuMob Challenge 2023') using the human mobility dataset, which will be held in conjunction with ACM SIGSPATIAL 2023.

Background & Summary

Understanding, modeling, and predicting human mobility trajectories in urban areas is an essential task for various domains and applications, including human behavior analysis[1], transportation and activity analysis[2], disaster risk management[3], epidemic modeling[4], and urban planning[5]. Traditionally, travel surveys and census data have been utilized as the main source of data to understand such macroscopic urban dynamics[6]. The recent availability of large-scale human movement and behavior data collected from (often millions of) mobile devices and social media platforms[7] have enabled the development and testing of complex human mobility models, resulting in a plethora of proposed methods for the prediction of human mobility traces[8].

Despite its academic popularity and societal impact, human mobility modeling and prediction methods are often trained and tested on different proprietary datasets, due to the lack of open-source and large-scale human mobility datasets amid privacy concerns[9]. This makes it difficult to conduct fair performance comparisons across different methods. Several efforts have created open-source datasets of human mobility. Real-world trajectory datasets include the GeoLife dataset, T-Drive trajectory dataset, and NYC Taxi and Limousine Commision dataset. The GeoLife dataset[10] provides trajectory data of 182 users across a period of over three years, containing 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000 hours. The T-Drive trajectory dataset contains trajectories of 10,357 taxis across a one week timeframe[11]. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches 9 million kilometers. Similarly, the New York City Taxi and Limousine Commission (NYC-TLC) provides pick-up and drop-off locations and timestamps data [1]. Although T-Drive and NYC-TLC datasets provide massive amounts of trajectory information, they are limited to taxi trips. There has also been several synthetic datasets produced from open-source data, including the Open PFLOW[12] and Pseudo-PFLOW datasets[13]. While such datasets are valuable in conducting large-scale experiments on human mobility prediction, the lack of metropolitan-scale, longitudinal, real-world, and open-source datasets of individuals has been one of the key barriers hindering the progress of human mobility model development.

To this end, we created an open-source and anonymized dataset of human mobility trajectories from mobile phone location

[1]https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

data provided by Yahoo Japan Corporation. The dataset contains 100,000 individuals' mobility trajectories across a 90 day period collected from an undisclosed, highly populated metropolitan area in Japan. The location pings are discretized into 500 meters x 500 meters grid cells and the timestamps into 30 minute bins. The actual date of the observations are not available either (i.e., timeslot t of day d) to protect privacy. The 90 day period is composed of 75 days of business-as-usual and 15 days during an emergency with unusual behavior.

To promote the use of the dataset, we will host a human mobility prediction data challenge ('HuMob Challenge 2023') using the human mobility dataset of 100K individuals' trajectories across 90 days. The workshop will be held in conjunction with ACM SIGSPATIAL 2023 [2]. Participants will be tasked to develop and test methods to predict human mobility trajectories using the provided open-source dataset (for details, see Section 'Human Mobility Prediction Data Challenge').

Methods

Observation of Smartphone GPS records

GPS location data were collected from smartphones which have installed the Yahoo Japan Application, and were anonymized so that individuals cannot be specified, and personal information such as gender, age and occupation are unknown. Each GPS location record contains the user's unique ID, timestamp of the observation, longitude, and latitude, and the data has a sample rate of approximately 5% of the entire population. The data acquisition frequency of GPS locations varies according to the movement speed of the user to minimize the burden on the user's smartphone battery. If it is determined that the user is staying in a certain place for a long time, data is acquired at a relatively low frequency, and if it is determined that the user is moving, the data is acquired more frequently.

Spatio-temporal Processing and Anonymization

The set of mobile phone users included in the dataset were selected by spatially and temporally cropping the raw dataset. To spatially crop the raw dataset, we created a boundary box around an undisclosed metropolitan area in Japan, and selected mobile phone users who were observed within the boundary box more than 10 times during 10 day period (dates undisclosed for privacy reasons). To make the mobile phone users unidentifiable, the location pings are discretized into 500 meters x 500 meters grid cells and the timestamps into 30 minute bins. The actual date of the observations were also masked (i.e., timeslot t of day d) to protect privacy. The movement (encoded into 500m grid cells) of the mobile phone users was tracked across a total of 90 days (again, dates are undisclosed), including a 75-day period of business-as-usual (*Dataset 1*) and another 15-day period under an emergency situation (*Dataset 2*), where we can assume human behavior and mobility patterns could be shifted. The dataset was finally cropped by selecting users with a sufficient number of 30-minute timeslot observations to ensure that the mobility patterns could be studied (see Figure 2 for distribution of pings per user). Observations outside of the target boundary box were discarded. For *Dataset 1*, 100,000 users were selected, and for *Dataset 2*, 25,000 users were selected.

Privacy Policy

Yahoo Japan Corporation (YJ) has developed its own privacy policy and requires users to read and agree to its privacy policy before using any of the services provided by YJ. Furthermore, because location data is highly sensitive for the users, users were asked to sign an additional consent form specific to the collection and usage of location data when they used apps that collect location information. The additional consent explains the frequency and accuracy of location information collection, and also the purpose and how the data will be used. Moreover, YJ implemented strict restrictions in the analysis procedure. The methodology for handling the data and for obtaining user consent for this study were supervised by an advisory board composed of external experts. YJ also ensured that research institutions other than YJ that participate in this study (including co-investigators) do not have direct access to the data. Although external research institutions were allowed to analyze aggregated data, the actual raw data were kept within YJ, and any analysis performed on raw data were performed within servers administered by YJ.

Data Records

Table 1 shows an example of the dataset provided. Each record refers to an observation of an individual:

- `user ID` is the unique identifier of the mobile phone user (type: integer)
- `day` is the masked date of the observation. It may take a value between 0 and 74 for both Dataset 1 and Dataset 2 (type: integer).

[2] https://sigspatial2023.sigspatial.org/

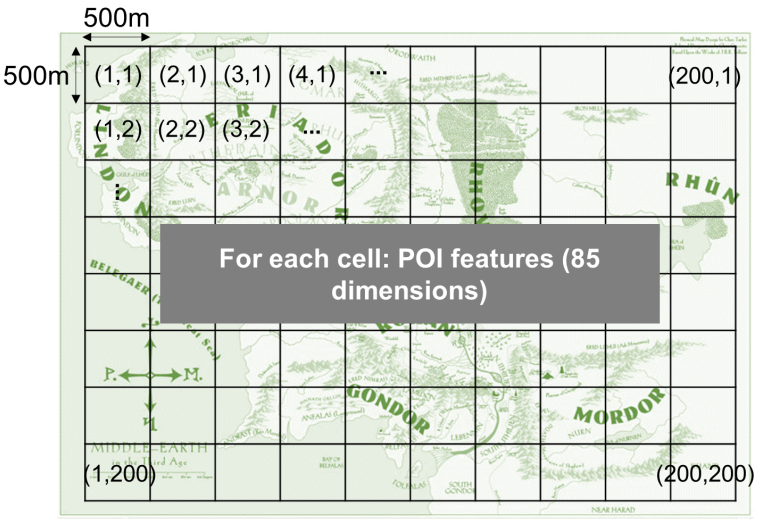

Figure 1. Spatial layout of the target city and the grid cells. Each grid cell is approximately 500 meters x 500 meters, and the target area spans 200 x 200 grid cells.

user ID	day	timeslot	x	y
1	1	13	10	13
1	1	18	11	15
1	1	24	11	17
1	1	27	12	19
		...		
2	3	15	31	19
2	3	28	35	33
2	4	12	35	36
		...		

Table 1. Example of dataframe and the columns in the human mobility trajectory dataset.

- `timeslot` is the timestamp of the observation discretized into 30 minute intervals. It may take a value between 0 and 47, where 0 indicates between 0AM and 0:30AM, and 13 would indicate the timeslot between 6:30AM and 7:00AM.
- `x,y` are the coordinates of the observed location mapped onto the 500 meter discretized grid cell. It may take a value between (1,1) and (200,200). Details are shown in Figure 1.

Basic Statistics of the Data

To provide guidance for data users, we have computed the basic descriptive statistics of *Dataset 1*. The total number of records are 111,535,175, with exactly 100,000 unique users (numbered 0 to 99,999), across 75 days (numbered 0 to 74), in 48 different 30 minute timesteps (numbered 0 to 47). Figure 2 shows the histogram of the number of pings per user ID (left) and the number of unique cells visited per user ID. Both plots show a skewed distribution, where a small fraction of the users are observed many times (i.e., more than 2000 pings, at 100 unique cells). Figure 3 shows the histogram of the number of pings per user ID (left) and the number of unique users visited to each grid cell. Note that the x-axis in both plots are log-scaled. Both plots show a bimodal distribution, where a large fraction of the cells are visited very few times (less than 10 pings or unique users) while another mode can be observed at around 10000 pings and 1000 unique users visited. This highlights the mix of urban and rural areas in the target region.

Figure 4 shows the temporal dynamics of the number of pings and unique users per day (from day 0 to 74) in Dataset 1. The patterns show temporal regularity, showing clear patterns of weekdays and weekends. There is an anomaly on day 27, however this is due to a data collection issue. The unique number of users observed each day fluctuates more, showing a decrease near days 40 to 50 and an increase from day 60 onwards. Figure 5 shows the temporal dynamics of the number of

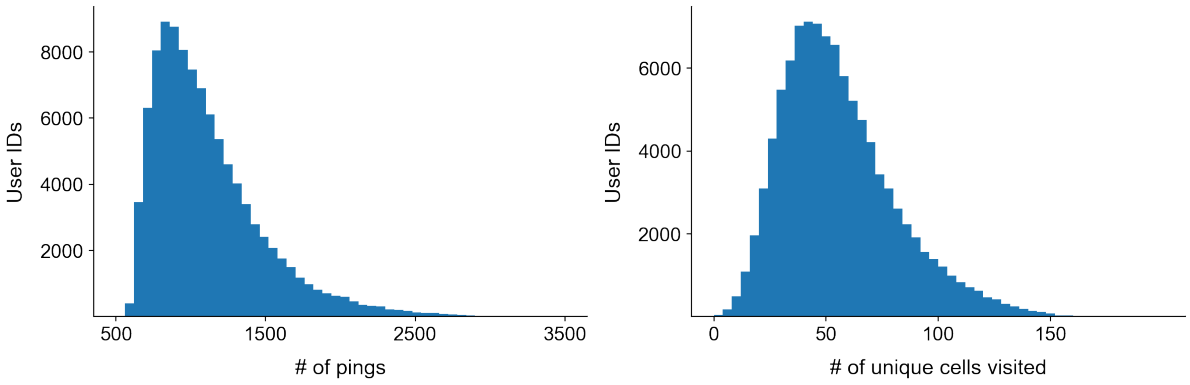

Figure 2. Histograms of the number of GPS location data pings and number of unique cells visited per user, across the 75 day period stored in Dataset 1.

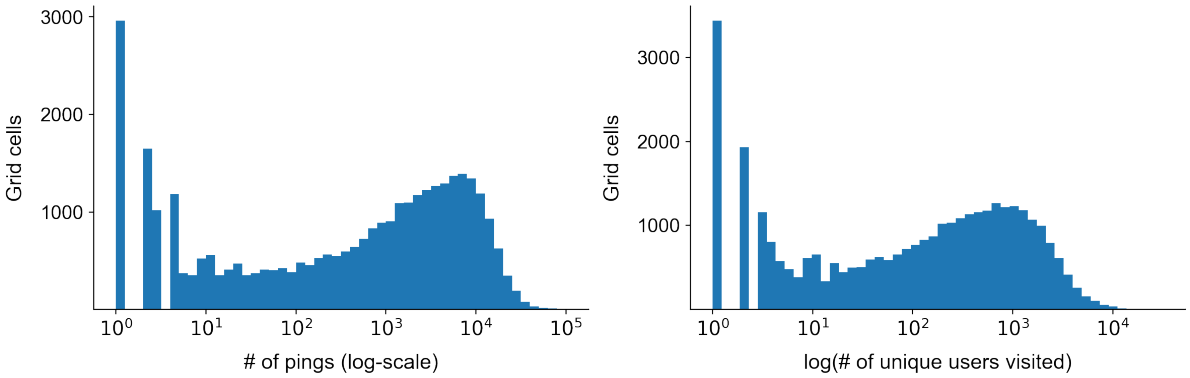

Figure 3. Histograms of the number of GPS location data pings and number of unique users visited per grid cell, across the 75 day period stored in Dataset 1.

pings and unique users per timeslot (from timeslot 0 to 47) aggregated across all days observed in Dataset 1. The patterns show temporal regularity, showing clear morning and daytime peaks. The unique number of users observed between timeslot 12 (6AM) and timeslot 40 (8PM) is stable at around 100,000, showing a high observability during those time periods. Figure 6 shows a 2-dimensional histogram of the number of pings and the number of observed unique users across the 75 days. Note that the scales are log-scaled. The patterns show clear urban (blue) and rural (red) areas.

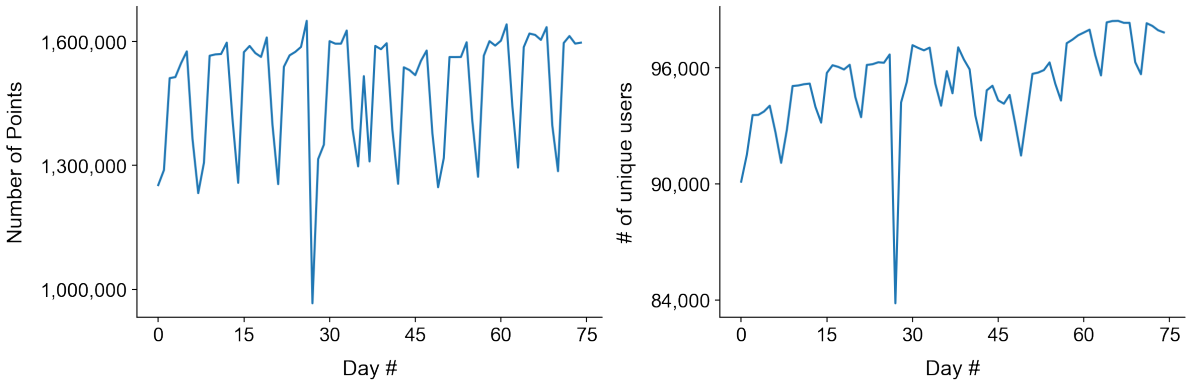

Figure 4. Temporal dynamics of the number of pings and unique users per day (from day 0 to 74) in Dataset 1.

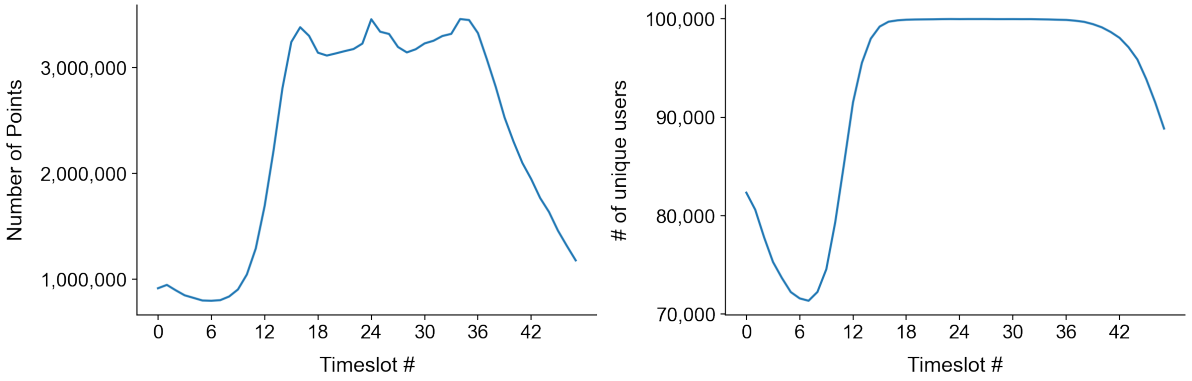

Figure 5. Temporal dynamics of the number of pings and unique users per timeslot (from timeslot 0 to 47) in Dataset 1.

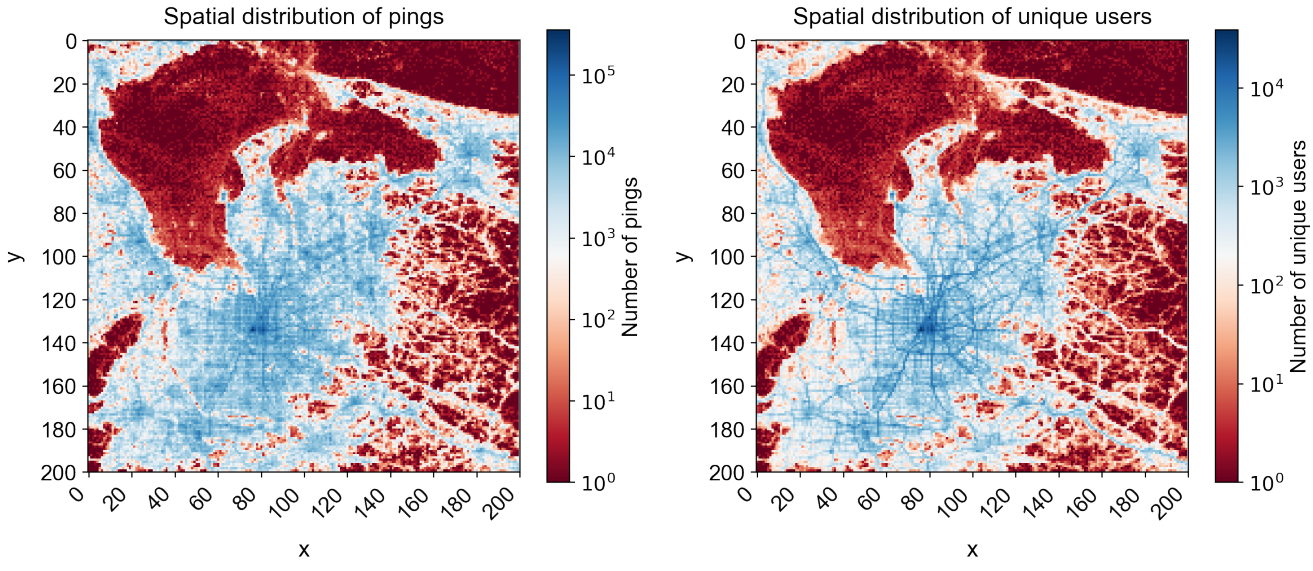

Figure 6. 2-dimensional histogram of the number of pings and the number of observed unique users across the 75 days. Note that the scales are log-scaled. The patterns show clear urban (blue) and rural (red) areas.

Human Mobility Prediction Data Challenge

The challenge takes place in a mid-sized and highly populated metropolitan area, somewhere in Japan. The area is divided into 500 meters x 500 meters cells, which span a 200 x 200 grid, as shown in Figure 1. The human mobility datasets ('task1_dataset.csv.gz' and 'task2_dataset.csv.gz') contain the movement of a total of 100,000 individuals across a 90 day period, discretized into 30-minute intervals and 500 meter grid cells. The first dataset contains the movement of a 75 day business-as-usual period, while the second dataset contains the movement of a 75 day period during an emergency with unusual behavior.

There are 2 tasks in the Human Mobility Prediction Challenge, as shown in Figure 7. In task 1, participants are provided with the full time series data (75 days) for 80,000 individuals, and partial (only 60 days) time series movement data for the remaining 20,000 individuals ('task1_dataset.csv.gz'). Given the provided data, Task 1 of the challenge is to predict the movement patterns of the individuals in the 20,000 individuals during days 60-74. Task 2 is similar task but uses a smaller dataset of 25,000 individuals in total, 2,500 of which have the locations during days 60-74 masked and need to be predicted ('task2_dataset.csv.gz').

While the name or location of the city is not disclosed, the participants are provided with points-of-interest (POIs; e.g., restaurants, parks) data for each grid cell (85 dimensional vector) as supplementary information (which is optional for use in the challenge) ('cell_POIcat.csv.gz'). For more details, see https://connection.mit.edu/humob-challenge-2023.

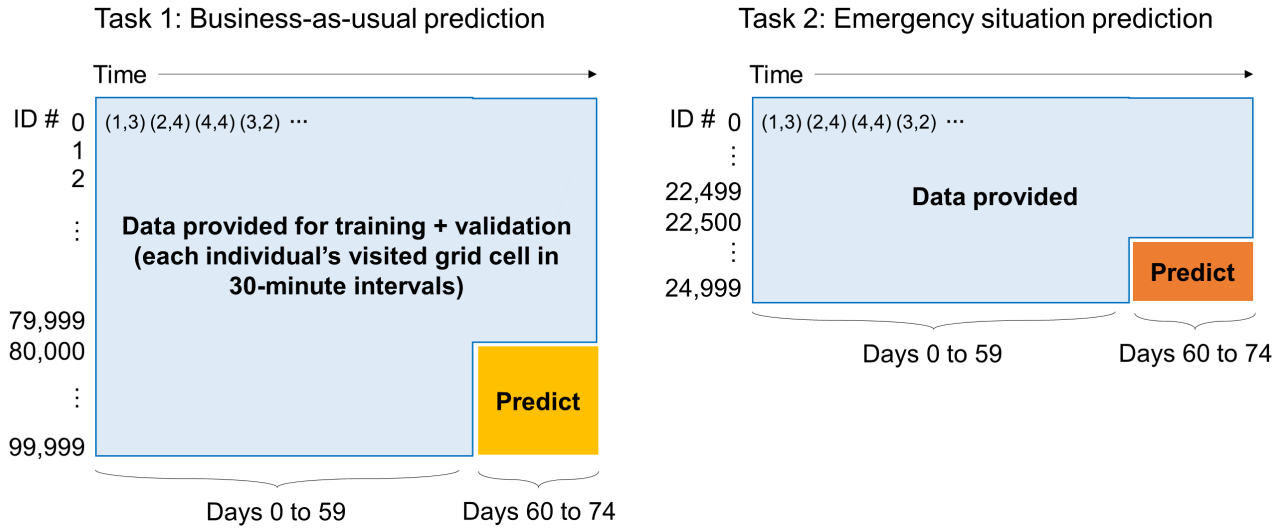

Figure 7. 2 tasks in the Human Mobility Prediction Challenge. In task 1, participants predict the movement of a subset (20,000) of the individuals for days 60 to 74 during a business-as-usual period. In task 2, participants predict the movement of a subset (2,500) of the individuals for days 60 to 74 during an emergency situation.

Provided Datasets and Tasks

The data challenge participants will be provided with 3 datasets – HuMob datasets #1 and #2 (which are derived from the original human mobility dataset), and the POI dataset which may be used to supplement the prediction of human mobility.

The data may be downloaded from https://zenodo.org/record/8111993. For teams to be granted access to the data, teams should request access via the Zenodo website by providing the name and email address of the lead investigator, and the following information in the 'Justification' box:

- full list of members (name, institution, email address)
- team name (alphabets and numbers only, keep it $\leq$ 10 characters)

Upon approval by the organizing team, the data will be available for download. If you do not receive the data approval within 24 business day hours, please contact humob2023@gmail.com with your information.

Participants shall not carry out activities that involve unethical usage of the data, including attempts at re-identifying data subjects, harming individuals, or damaging companies. Participants will be allowed to submit full versions of their works to venues of their choice, upon approval by the organizers.

HuMob dataset #1 (`task1_dataset.csv.gz`)
Contains the movement of 100,000 individuals in total during a business-as-usual scenario. 80,000 individuals' movements are provided completely (for 75 days), and the remaining 20,000 individuals' movements for days 61 to 75 are masked as `999`. The challenge is to use the provided data to predict the masked cell coordinates (i.e., replace the '`999`'s). See Table 1 for the data format.

HuMob dataset #2 (`task2_dataset.csv.gz`)
Contains the movement of 25,000 individuals in total during a business-as-usual scenario. 22,500 individuals' movements are provided completely (for 75 days), and the remaining 2,500 individuals' movements for days 61 to 75 are masked as `999`. Similar to task 1, the challenge is to use the provided data to predict the masked movement coordinates (i.e., replace the '`999`'s). See Table 1 for the data format.

POI dataset (`cell_POIcat.csv.gz`)
To aid the prediction task, we have prepared an auxiliary dataset that provides the count of different points-of-interest categories in each grid cell (e.g., restaurants, cafes, schools). However, to maintain anonymity of the location, we are not able to provide the actual category name that corresponds to each dimension. Therefore, each cell has a 85 dimensional vector, as shown in Table 2.

x	y	POI category (dim)	# of POIs
1	1	13	10
1	1	18	11
1	1	24	11
1	1	27	12
		...	
2	2	15	31
2	2	28	35
2	2	12	35
		...	

Table 2. Example of dataframe and the columns in the POI category dataset. First two columns show the x and y coordinates of the grid cell, third column denotes the dimension of the POI category (between 1 and 85), and the fourth column shows how many POIs of the POI category dimension located in the grid cell.

Evaluation Metrics
Two metrics will be used to measure the accuracy of the predicted mobility trajectories:

- Dynamic Time Warping (DTW)[14], for evaluating the similarity of trajectories as a whole, with step-by-step alignment.
- GEO-BLEU[15], a metric with a stronger focus on local features, as in similarity measures for natural language sentences. Python implementation for the `GEOBLEU` metric can be found at https://github.com/yahoojapan/geobleu.

Submissions will be ranked for each metric, and the top 10 teams will be decided based on the two rankings. We recommend the teams to try to optimize for both metrics.

Submission Procedure and Rules
- Prediction results for Tasks 1 and 2 should be uploaded to an online storage (e.g., Dropbox, Box, Google Drive, etc.) and the download links should be sent to `humob2023@gmail.com`.
- The attached files should be named as: `teamnumber_{task1,task2}_humob.csv.gz`. For example, team number 5 submitting their solutions for task 1 should submit their prediction as `5_task1_humob.csv.gz`.
- Only 1 submission per team would be evaluated. The final submission before the deadline (September 15th 23:59 AOE) will be considered as the final submission.
- The format of the submission should include the same 5 columns as the original dataset (user ID, day, timeslot, x, y). Separate the columns using commas (,) and include no redundant spaces, and save the file using the `csv.gz` format.
- *Only send the data for the predicted users*. For Task 1, only users #80000 to #99,999, and for Task 2, only users #22500 to #24999.

Important Dates

The top 10 teams with the best predictions will be invited to submit a final report with details of the methods and to present their work at the HuMob 2023 Workshop held in conjunction with ACM SIGSPATIAL 2023 in Hamburg, Germany on November 13th, 2023. We have prizes for the top 3 participants!

- June 15, 2023: data challenge announcement
- July 10, 2023: data open at https://zenodo.org/record/8111993
- September 15, 2023: submission deadline for predictions
- September 22, 2023: notification of top contestants
- October 14, 2023: submission deadline of workshop papers for top 10 teams
- October 20, 2023: camera-ready submission
- November 13, 2023: presentation in the workshop

Organizing Team

The team members are: Dr. Takahiro Yabe, MIT; Dr. Kota Tsubouchi, Yahoo Japan Corporation; Toru Shimizu, Yahoo Japan Corporation; Professor Yoshihide Sekimoto, University of Tokyo; Professor Kaoru Sezaki, University of Tokyo; Professor Esteban Moro, MIT; Professor Alex 'Sandy' Pentland, MIT. For general questions about the challenge: humob2023@gmail.com

Code availability

The dataset can be downloaded from https://zenodo.org/record/8111993, and details about the Data Challenge can be found in https://connection.mit.edu/humob-challenge-2023. Python implementation for the GEOBLEU metric can be found at https://github.com/yahoojapan/geobleu.

References

1. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
2. Jiang, S., Ferreira, J. & Gonzalez, M. C. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data* **3**, 208–219 (2017).
3. Yabe, T., Jones, N. K., Rao, P. S. C., Gonzalez, M. C. & Ukkusuri, S. V. Mobile phone location data for disasters: A review from natural hazards and epidemics. *Comput. Environ. Urban Syst.* **94**, 101777 (2022).
4. Oliver, N. *et al.* Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Sci. Adv.* **6**, eabc0764 (2020).
5. Ratti, C., Frenchman, D., Pulselli, R. M. & Williams, S. Mobile landscapes: using location data from cell phones for urban analysis. *Environ. Plan. B: Plan. Des.* **33**, 727–748 (2006).
6. Sekimoto, Y., Shibasaki, R., Kanasugi, H., Usui, T. & Shimazaki, Y. Pflow: Reconstructing people flow recycling large-scale social survey data. *IEEE Pervasive Comput.* **10**, 27–35 (2011).
7. Blondel, V. D., Decuyper, A. & Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **4**, 1–55 (2015).
8. Luca, M., Barlacchi, G., Lepri, B. & Pappalardo, L. A survey on deep learning for human mobility. *ACM Comput. Surv. (CSUR)* **55**, 1–44 (2021).
9. De Montjoye, Y.-A. *et al.* On the privacy-conscientious use of mobile phone data. *Sci. Data* **5**, 1–6 (2018).
10. Zheng, Y., Wang, L., Zhang, R., Xie, X. & Ma, W.-Y. Geolife: Managing and understanding your past life over maps. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*, 211–212 (IEEE, 2008).
11. Yuan, J. *et al.* T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 99–108 (2010).
12. Kashiyama, T., Pang, Y. & Sekimoto, Y. Open pflow: Creation and evaluation of an open dataset for typical people mass movement in urban areas. *Transp. Res. Part C: Emerg. Technol.* **85**, 249–267 (2017).

13. Kashiyama, T., Pang, Y., Sekimoto, Y. & Yabe, T. Pseudo-pflow: Development of nationwide synthetic open dataset for people movement based on limited travel survey and open statistical data. *arXiv preprint arXiv:2205.00657* (2022).

14. Senin, P. Dynamic time warping algorithm review. *Inf. Comput. Sci. Dep. Univ. Hawaii at Manoa Honolulu, USA* **855**, 40 (2008).

15. Shimizu, T., Tsubouchi, K. & Yabe, T. Geo-bleu: similarity measure for geospatial sequences. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4 (2022).

Author contributions statement

T.Y. and K.T developed and computed the mobility indices. T.S. developed the evaluation metric code. All authors wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.