https://doi.org/10.1038/s43588-024-00650-3

Enhancing human mobility research with open and standardized datasets

Takahiro Yabe, Massimiliano Luca, Kota Tsubouchi, Bruno Lepri, Marta C. Gonzalez & Esteban Moro

Human mobility research intersects with various disciplines, with profound implications for urban planning, transportation engineering, public health, disaster management, and economic analysis. Here, we discuss the urgent need for open and standardized datasets in the field, including current challenges and lessons from other computational science domains, and propose collaborative efforts to enhance the validity and reproducibility of human mobility research.

The emergence of human mobility data and research

The proliferation of large-scale, passively collected location data from mobile devices has enabled researchers to gain valuable insights into various societal phenomena[1]. In particular, research into the science of human mobility has become increasingly critical thanks to its interdisciplinary effects in various fields, including urban planning, transportation engineering, public health, disaster management, and economic analysis[2]. Researchers in the computational social science, complex systems, and behavioral science communities have used such granular mobility data to uncover universal laws and theories governing individual and collective human behavior[3]. Moreover, computer science researchers have focused on developing computational and machine learning models capable of predicting complex behavior patterns in urban environments. Prominent papers include pattern-based and deep learning approaches to next-location prediction and physics-inspired approaches to flow prediction and generation[4].

Regardless of the research problem of interest, human mobility datasets often come with substantial limitations. Existing publicly available datasets are often small, limited to specific transport modes, or geographically restricted, owing to the lack of open-source and large-scale human mobility datasets caused by privacy concerns[5]. Examples of real-world trajectory datasets include the widely used GeoLife[6], T-Drive trajectory dataset[7], the NYC Taxi and Limousine Commission dataset[8], and the Gowalla dataset[9], and although such datasets are valuable in conducting large-scale experiments on human mobility prediction, the lack of metropolitan-scale and longitudinal open-source datasets of individuals has been one of the key barriers hindering the progress of human mobility model development. The lack of open data also perpetuates gatekeeping, where researchers without access to exclusive datasets are excluded from this research area, raising equity concerns in science. Moreover, even in the case where researchers may access processed mobility datasets, privacy concerns limit access to raw and open data sources. This means that even the datasets that are publicly available are often pre-processed without using standardized procedures. It is possible to obtain a completely different dataset just by slightly changing a parameter in the data pre-processing pipeline, for instance, by changing the spatial and temporal definition of a stop location. This makes it difficult to conduct fair performance comparisons across different methods[10].

Challenges in human mobility data and research

Human mobility datasets are produced from raw geolocation data through a series of pre-processing steps, the details of which are often not disclosed to those outside the research team that conducted the analysis, as shown in Fig. 1. Pre-processing steps are conducted (1) to de-noise the data and remove GPS drifts, (2) to correct for any potential biases in the mobility data, (3) to enrich the data's semantic information, and (4) to comply with privacy standards. Location data, often originally collected for marketing and business rather than research purposes, typically contain various biases. These include, but are not limited to, demographic biases (such as age, income and race), geographic biases (such as urban versus rural areas, developed versus developing countries), and behavioral biases, where observations may be more frequent during certain activities, such as checking into points of interest (POIs)[11].

Moreover, to enrich the data's semantic information for further analyses, various pre-processing steps are applied to the dataset, including user cutoff and selection, stop detection, privacy enhancement, attribution of points of interest and other contexts, and transport mode estimation. Each of these steps requires the selection of multiple parameters by the data analyst. For example, to detect a stop within a mobility trajectory, data scientists need to define arbitrary hyperparameters such as the minimum number of minutes spent at the stop and the maximum movement distance allowed from the stop centroid. With several hyperparameters needed for each pre-processing step, a slight change in the selection of these parameters could result in a very different processed human mobility dataset.

The complexity of human mobility data processing makes it difficult for data users, including researchers and analysts, to keep track of all of the decisions that were made during the pre-processing steps. Moreover, thanks to the proprietary nature of raw and processed human mobility datasets, disclosing the details of the pre-processing methods may not be sufficient to grasp the full characteristics of the human mobility data with which the downstream tasks were conducted. This lack of transparency about the quality of processed human mobility datasets raises critical issues in human mobility research, including the lack of replicability, generalizability, and comparability of method performance. Researchers may claim state-of-the-art prediction results

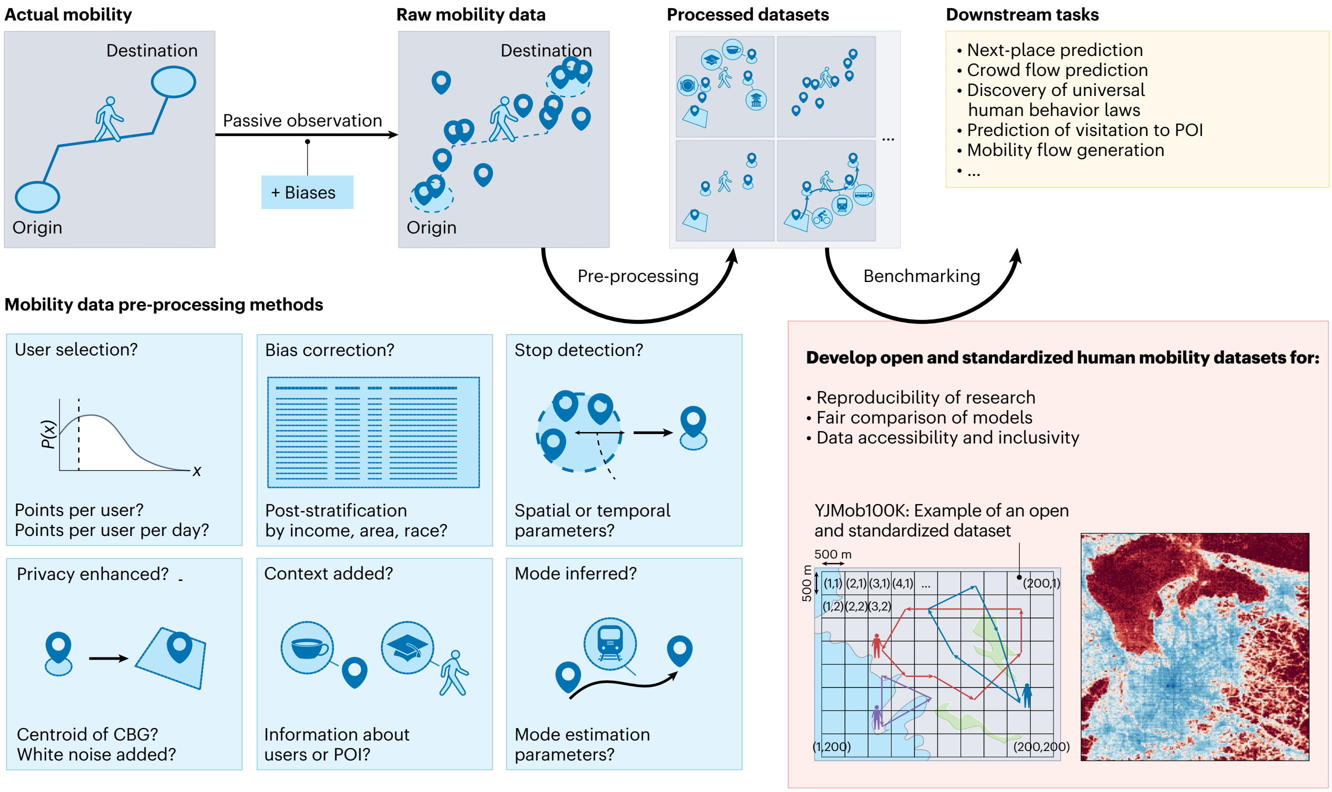
Actual mobility
Destination
Origin
Passive observation
+ Biases
Raw mobility data
Destination
Origin
Processed datasets
...
Downstream tasks
• Next-place prediction
• Crowd flow prediction
• Discovery of universal human behavior laws
• Prediction of visitation to POI
• Mobility flow generation
• ...
Pre-processing
Benchmarking
Mobility data pre-processing methods
User selection?
P(x)
x
Points per user?
Points per user per day?
Bias correction?
Post-stratification by income, area, race?
Stop detection?
Spatial or temporal parameters?
Privacy enhanced?
Centroid of CBG?
White noise added?
Context added?
Information about users or POI?
Mode inferred?
Mode estimation parameters?
Develop open and standardized human mobility datasets for:
• Reproducibility of research
• Fair comparison of models
• Data accessibility and inclusivity
YJMob100K: Example of an open and standardized dataset
500 m
500 m
(1,1) (2,1) (3,1) (4,1) ...
(200,1)
(1,2) (2,2) (3,2)
(1,200)
(200,200)

Fig. 1 | Human mobility data are produced from raw mobile phone location data through a series of complex pre-processing steps, and are used for interdisciplinary downstream research tasks. We argue for the need for fit-for-purpose and standardized human mobility benchmark datasets for reproducible, fair and inclusive human mobility research. CBG, census block group; POI, point of interest.

on specific datasets, potentially leading to overfitting and loss of generalizability. To address the lack of transparency on the validity of mobility data, several data companies (such as Unacast, Safegraph, and Cuebiq), as well as scientific papers[12] have evaluated and reported the accuracy of human mobility datasets through comparisons with available external data, such as the American Community Survey and visitation patterns to stadiums and factory facilities.

Learning from other scientific domains

Being able to compare the performances of different techniques and methods fairly is essential to the design of more efficient and effective methods. In the early days of machine learning and deep learning research, the ImageNet dataset hosted competitions to test the performance of computer vision algorithms, and has been widely recognized for driving innovation in deep learning research[13]. There has been an explosion of domain-specific or task-specific benchmark datasets in recent years, for example, in computer vision, natural language processing, speech processing, graph machine learning, molecular machine learning, atmospheric sciences, urban three-dimensional point clouds, and multi-modal machine learning. These datasets have been utilized by a large number of researchers, contributing to the collective development of fair and reproducible methods and techniques. In particular, such benchmark datasets guarantee by design that all the algorithms are tested using the same data settings, such as using the same training and test sets, as well as pre-processing procedures.

To date, however, there is no benchmark dataset of human mobility to serve state-of-the-art data-intensive research on human mobility. The proprietary nature of human mobility datasets, as opposed to text and image data that can be scraped from the open web, has been a major barrier to the development of benchmark mobility datasets.

Towards addressing the data challenges in human mobility

Although there is no standard open benchmark dataset in human mobility research, one approach that has been explored more recently to overcome this limitation is the development of machine learning models that generate synthetic privacy-preserving human mobility datasets. Earlier attempts have used models such as recurrent neural networks and long short-term memory to model human mobility sequences. More recent works have proposed diffusion model-based and attention-based modeling approaches, and have shown substantial improvement in the emulation accuracy and practicality of these datasets[14]. With the rapid innovations in large language models and foundation models, we anticipate substantial progress.

Another approach to bridging this gap is to develop privacy-safe, anonymized large-scale mobility datasets in collaboration with private data companies. An example is our work with Yahoo Japan Corporation (now called LY Corporation), where we created an open-source and anonymized dataset of human mobility trajectories from mobile phone location data, tailored for next-place prediction tasks, named YJMob100K[15]. To ensure that the YJMob100K dataset can be widely used as a dataset to create benchmarks for human mobility research, we minimized the number of arbitrary pre-processing decisions and kept the data as close to the raw form as possible. In this way, researchers will have the flexibility to apply their desired pre-processing techniques and procedures for their application and problem settings. Initial feedback from data users has been overall positive, but some users have shared concerns about how the data is messy and noisy in its current form. For example, there is much heterogeneity in data quantity and quality across individual users, which is a typical bias inherent in human mobility data. This was a design decision we made to increase the flexibility that researchers would have in pre-processing, also with the hope that the YJMob100K dataset will raise awareness and bring more attention to challenges regarding the pre-processing steps for mobility data.

Data challenges are effective ways to bring together the research community and raise awareness of both the new dataset and societal challenges that may be solved using it. In the past, numerous data challenges have been organized together with telecommunication companies in the human mobility research domain. Examples include two Data for Development challenges using mobile phone data provided by Orange in the Ivory Coast and Senegal, Telecom Italia's Big Data Challenge, the Future Cities Challenge by Foursquare at NetMob 2019, the NetMob 2023 Data Challenge with Orange, and the Data4Refugees co-organized by Türk Telekom, Bogazici University and Tübitak in collaboration with Fondazione Bruno Kessler, MIT Media Lab, the Data-Pop Alliance, UNHCR, IOM, and UNICEF in 2018. In each of these data challenges, 50 to 100 teams applied the datasets to different problem contexts. Such early efforts have successfully standardized the production and accessibility of mobility datasets. To further promote the use of the YJMob100K dataset, we hosted a human mobility prediction data challenge (HuMob Challenge 2023) using the YJMob100K dataset (https://connection.mit.edu/humob-challenge-2023). The workshop was held in conjunction with ACM's SIGSPATIAL conference in 2023, attracting more than 85 teams to participate in the data challenge. The challenge brought together a community of over 200 human mobility researchers from academia and public agencies such as the World Bank, with expertise ranging from urban planning to computer science. Beyond the data challenge, we hope that the YJMob100K dataset will serve as a tool for fostering fair, reproducible, and accessible human mobility research. The YJMob100K dataset was used as a dataset for the next place-prediction task in the HuMob Data Challenge, but the next step should be to build benchmark datasets for other human mobility-related tasks and questions that require different considerations.

Open challenges and call to action

The development of the YJMob100K dataset was an attempt to create a human mobility benchmark dataset for the mobility prediction research community, but we recognize several key challenges. First, we need to define criteria for 'fit-for-purpose' benchmarking datasets to foster consensus within the research community. There must be community consensus on data specification metrics and industry standards for pre-processing, including but not limited to the steps outlined in Fig. 1. Second, different research tasks require different types of data. For example, the YJMob100K dataset may be suitable for human mobility prediction tasks, but because the data do not include any information about the specific city or POIs in each grid cell, it may be less suitable for urban science research investigating contextual information about human mobility. For crowd-flow prediction tasks, benchmark datasets do not require individual-level data and could be spatially and temporally aggregated. Instead, having more contextual information about space (for instance, actual longitude and latitude information) and time (for instance, actual dates, times, and event information) could be beneficial for such tasks. In addition to developing human mobility benchmark datasets for different downstream tasks, the research community could benefit from geographical, social, economic, temporal, and contextual diversity. Studies have revealed how the performance of models and methods could depend greatly on the heterogeneity of these contexts, such as urban versus rural areas, high-income versus low-income regions, and normal times versus during disaster events[1,2]. Thus, we argue that a collection of 'fit-for-purpose' benchmark datasets, which are tailored to specific research domains, communities, and socio-spatial-temporal contexts are necessary and should be developed through a bottom-up effort by the respective research communities.

We have argued that science and research in numerous domains would benefit from the development of open and standardized human mobility datasets to complement proprietary alternatives. Scientists, researchers, and practitioners need to advocate for the use of open-source human mobility data. They should collaboratively develop open-access benchmark datasets tailored to various research needs, as well as convene as a research community to establish and standardize data specifications and pre-processing protocols, ensuring consistency and reliability in data usage across studies. Such collective efforts will not only enhance the validity and reproducibility of human mobility research but also democratize access to high-quality data, paving the way for more inclusive and effective scientific inquiry.

Takahiro Yabe[1,2] ✉, **Massimiliano Luca**[3,4], **Kota Tsubouchi**[5], **Bruno Lepri**[3], **Marta C. Gonzalez**[6] **& Esteban Moro**[7,8]

[1]Center for Urban Science and Progress and Department of Technology Management and Innovation, Tandon School of Engineering, New York University, Brooklyn, NY, USA. [2]Center for Spatial Information Science, University of Tokyo, Kashiwa, Chiba, Japan. [3]Mobile and Social Computing Lab, Fondazione Bruno Kessler, Trento, Italy. [4]Instituto de Fisica Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Palma de Mallorca, Spain. [5]LY Research, LY Corporation, Chiyoda, Tokyo, Japan. [6]Department of Civil and Environmental Engineering and Department of City and Regional Planning, University of California, Berkeley, Berkeley, CA, USA. [7]Network Science Institute and Department of Physics, Northeastern University, Boston, MA, USA. [8]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA.

✉e-mail: takahiroyabe@nyu.edu

Published online: 2 July 2024

References

1. Blondel, V. D., Decuyper, A. & Krings, G. *EPJ Data Sci.* **4**, 10 (2015).
2. Yabe, T., Jones, N. K., Rao, P. S. C., Gonzalez, M. C. & Ukkusuri, S. V. *Comput. Environ. Urban Syst.* **94**, 101777 (2022).
3. Pappalardo, L., Manley, E., Sekara, V. & Alessandretti, L. *Nat. Comput. Sci.* **3**, 588–600 (2023).
4. Luca, M., Barlacchi, G., Lepri, B. & Pappalardo, L. *ACM Comput. Surv.* **55**, 7 (2021).
5. De Montjoye, Y. A. et al. *Sci. Data* **5**, 180286 (2018).
6. Zheng, Y., Xie, X. & Ma, W. Y. *IEEE Data Eng. Bull.* **33**, 32–39 (2010).
7. Yuan, J. et al. in *Proc. 18th SIGSPATIAL Int. Conf. Advances in Geographic Information Systems* 99–108 (Association for Computing Machinery, 2010).
8. TLC trip record data. *New York City Taxi and Limousine Commission* https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (2023).

9. Cho, E., Myers, S. A. & Leskovec, J. in *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* 1082–1090 (Association for Computing Machinery, 2011).
10. Smolak, K., Siła-Nowicka, K., Delvenne, J. C., Wierzbiński, M. & Rohm, W. *Sci. Rep.* **11**, 15177 (2021).
11. Coston, A. et al. in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency* 173–184 (Association for Computing Machinery, 2021).
12. Moro, E., Calacci, D., Dong, X. & Pentland, A. *Nat. Commun.* **12**, 4633 (2021).
13. Deng, J. et al. in *2009 IEEE Conf. Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
14. Zhu, Y., Ye, Y., Wu, Y., Zhao, X. & Yu, J. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)* (eds Oh, A. et al.) 22961–22977 (Curran Associates, Inc., 2023).
15. Yabe, T. et al. *Sci. Data* **11**, 397 (2024).

Acknowledgements

M.L. and B.L. acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. E.M. acknowledges support by the National Science Foundation (grants 2218748 and 2420945).

Author contributions

All authors developed the idea. T.Y. prepared the original draft. All authors reviewed and edited the paper.

Competing interests

The authors declare no competing interests.