

# Building a Global-Scale Trajectory Dataset from OpenStreetMap

Yuanshao Zhu  
City University of Hong Kong  
Hong Kong, China  
yaso.zhu@my.cityu.edu.hk

## Abstract

GPS trajectory data serves as an essential resource for analyzing human mobility, optimizing transportation systems, and supporting data-driven decision-making across multiple disciplines. However, the limitations of existing trajectory datasets have restricted the community from thriving. First, data accessibility locks out wider use and collaboration due to proprietary restrictions or privacy concerns. Second, these datasets typically cover narrow geographic areas, which undermines the generalizability and reliability of their findings. Additionally, existing datasets suffer from low data quality and inconsistencies, which pose obstacles to contemporary trajectory analysis research. To address these limitations, we introduce the WorldTrace dataset, which contains millions of trajectories with billions of points across the world. The dataset also provides a suite of data types, including detailed GPS coordinates and trip metadata, offering rich contextual information for comprehensive analysis. Moreover, rigorous data processing and validation procedures ensure the accuracy and reliability of the dataset. We believe that the comprehensive, high-quality, and global nature of WorldTrace offers new perspectives for trajectory data analysis and drives innovation across various research fields.

## CCS Concepts

• **Information systems** → **Spatial-temporal systems**; *Geographic information systems*; *Global positioning systems*.; • **Applied computing** → *Transportation*.

## Keywords

GPS Trajectory, Spatio-Temporal Data Mining, Urban Computing

### ACM Reference Format:

Yuanshao Zhu. 2025. Building a Global-Scale Trajectory Dataset from OpenStreetMap. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'25)*. ACM, New York, NY, USA, 3 pages. <https://doi.org/XX.XXX>

## 1 Introduction

Trajectory data, encompassing the sequence of time-stamped geographic coordinates traversed by an entity over time, serves as a critical record of movement patterns. This data offers valuable

insights into human mobility at both spatial and temporal scales, empowering researchers and practitioners across diverse fields [1, 5]. Analysis of trajectory data permits us a deeper understanding of the transportation system, informs strategic planning for urban development, and refines traffic forecasting models [11]. Beyond transportation, trajectory data proves instrumental in epidemiology, environmental science, and socioeconomic studies, where mobility patterns help uncover broader trends and behaviors [10].

What measures can be taken to advance trajectory-based research further? Empirically, comprehensive, large-scale, and high-quality datasets are essential for the progress of any research community. For instance, datasets such as ImageNet for computer vision [2], GLUE for natural language processing [8], and LargeST for spatio-temporal data mining [4] have significantly advanced their respective fields. However, recent efforts in trajectory-based research have predominantly focused on developing algorithms and models using advanced machine learning techniques, neglecting the critical need for comprehensive trajectory datasets [10, 14]. Consequently, there is an urgent need for a publicly available trajectory dataset with sufficient breadth, volume, granularity, and multimodal integration to enable researchers to fully exploit the potential of trajectory data analysis.

With these limitations in mind, we introduce WorldTrace, a public, large-scale, and high-quality GPS trajectory dataset. WorldTrace contains millions GPS trajectories collected worldwide between 2021 and 2023, aiming to advance trajectory analysis and human mobility research. The raw data, sourced from OpenStreetMap [7], undergoes rigorous data calibration and map matching to ensure high accuracy and reliability. Specifically, WorldTrace distinguishes itself from previous GPS trajectory datasets through its diverse geographic scope, superior quality, and data consistency. Firstly, WorldTrace provides samples on a global scale, overcoming geographical limitations and ensuring applicability to various research contexts. Secondly, the dataset features a high-resolution sampling at one-second intervals, enabling a finer-grained understanding of movement patterns and behaviors. Most importantly, WorldTrace extends beyond basic GPS coordinates to include rich information types, such as GPS coordinates, trip metadata, corresponding geographical information. We believe that the comprehensive and global nature of WorldTrace offers novel perspectives for trajectory analysis and supports more comprehensive research endeavors. By making WorldTrace publicly accessible, we aim to foster open collaboration and drive innovation across various research fields.

## 2 Related Work

The advancement of recent studies relies heavily on existing datasets, but researchers still face several key limitations that hinder the progression of trajectory data analysis. These datasets vary in availability, geographic coverage, granularity, and overall quality, each with

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*KDD'25, Aug 03–07, 2025, Toronto, ON, Canada*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-XXXX-X/18/06  
<https://doi.org/XX.XXX>



**Figure 1: WorldTrace dataset acquisition and processing workflow.**

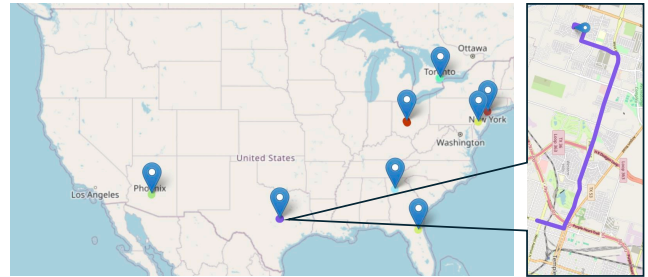
their benefits and constraints. One of the most widely recognized datasets is the GeoLife dataset [13], which has been instrumental in numerous studies. Datasets like Porto [6], T-drive [12], and Electric Vehicle Data [9] are collected through GPS devices mounted on taxi. These datasets provide a broader picture of the activities of multiple individuals but typically offer low or uneven sampling rates, which limits detailed movement pattern analysis. As a synthetic dataset, SynMob offers a uniform sampling rate and an unlimited amount of data [14], but its quality and regions still depend on the original data. Proprietary datasets such as those from Grab-Posisi [3], Taxi-Shanghai, DiDi-Chengdu, and DiDi-Xian offer high-quality and extensive data but are not publicly accessible, limiting their utility for widespread investigation. In particular, the coverage of all existing trajectory datasets is limited to a single city, which hinders broad, global trajectory research.

### 3 Methodology

As shown in Figure 2, the trajectory data used in WorldTrace was collected from the OpenStreetMap (OSM) [7]. This platform, a public sharing project, hosts over 11 million GPS trajectories uploaded by contributors worldwide from 2004 to the present. Here, we focus specifically on car trajectories uploaded by openpilot-based systems such as dragonpilot and sunnypilot between 2021 and 2023. These sources provided over 4 million high-quality raw trajectories in GPX, GIF, and JSON formats, containing geographic coordinates, visual summaries, and metadata. To ensure consistency and enhance usability, the raw data underwent several processing steps: First, we concentrate on GPX and JSON files to obtain the original trajectory sequence and attribute information for that segment of the trip. Then, the raw data was resampled to a one-second interval, reducing redundancy from the original 10Hz sampling rate while preserving essential motion details. Next, we applied filtering to discard trajectories with fewer than 32 points, trip lengths under 100 meters, or implausibly high speeds. Then, map matching was performed to align GPS points to the road network, correcting positional errors caused by noise and enriching trajectories with road-level semantics such as road segment IDs, directions, and speed constraints. Finally, we conducted road information retrieval, obtaining additional attributes like road name, class, and speed limit via OSM APIs or from a locally built OSM database, ensuring efficient large-scale access to map features.

### 4 Results

Upon obtaining and preprocessing the raw trajectory data, we conducted an in-depth analysis to verify the characteristics and quality of the WorldTrace dataset. First, in terms of geographical coverage, this dataset includes trajectory data from more than 70 countries and regions across five major continents, providing unprecedented geographical breadth and diversity. Among them, the trajectory density and number of trajectories in North America, East Asia,



**Figure 2: Sample WorldTrace data from the contiguous USA.**

and parts of Europe significantly exceeded those of other regions, a characteristic consistent with regional development and infrastructure maturity. A detailed list of countries/regions with the highest trajectory counts is provided below, with the United States, China, and Canada leading in data volume, reflecting diverse urban forms, road networks, and mobility cultures. Notably, this demonstrates significant geographical diversity, with trajectory densities varying across urban, suburban, and rural environments. Regarding the spatio-temporal quality of the data, this dataset contains approximately 2.45 million trajectories, comprising approximately 8.8 billion raw GPS points (before normalization). To maintain data consistency, we standardized the sampling interval to 1 second, with the dataset spanning from 2021 to 2023. The average duration of all trajectories is approximately 6 minutes, with an average distance of 5.73 kilometers and an average speed of 48.0 kilometers per hour. To summarize, WorldTrace provides a robust resource for training and evaluating universal trajectory models. Its extensive coverage and comprehensive statistical properties make it suitable for developing the new models that generalize across tasks and geographic regions, thereby addressing the limitations of existing trajectory datasets.

All data follows a strict and comprehensive processing and privacy removal process, and no personal information is included in any traces. Regarding data privacy and open protocols, all raw data is subject to the OpenStreetMap Open Database License (ODbL): <http://opendatacommons.org/licenses/odbl/1.0/>. We will share derivative datasets under the same license terms to comply with the community’s data usage policy. Currently, we provide a data sample for reference purposes <sup>1</sup>.

### 5 Conclusion

In this work, we introduced WorldTrace, a large-scale, high-resolution, and globally distributed GPS trajectory dataset designed to support and advance trajectory-based research. By addressing the limitations of existing datasets—such as restricted access, narrow geographic scope, low sampling rates, and data inconsistency—WorldTrace offers a comprehensive resource with millions of car trajectories collected from over 70 countries between 2021 and 2023. We believe that this open and high-quality nature of WorldTrace will not only facilitate more generalizable and robust trajectory analysis efforts but also foster interdisciplinary research and innovation in areas such as urban computing, transportation modeling, and human mobility analysis with a global perspective.

<sup>1</sup><https://bit.ly/4fkHjBB>

## References

- [1] Wei Chen, Yuxuan Liang, Yuanshao Zhu, Yanchuan Chang, Kang Luo, Haomin Wen, Lei Li, Yanwei Yu, Qingsong Wen, Chao Chen, et al. 2024. Deep Learning for Trajectory Data Management and Mining: A Survey and Beyond. *arXiv preprint arXiv:2403.14151* (2024).
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [3] Xiaocheng Huang, Yifang Yin, Simon Lim, Guanfeng Wang, Bo Hu, Jagannadan Varadarajan, Shaolin Zheng, Ajay Bulusu, and Roger Zimmermann. 2019. Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility*. 1–10.
- [4] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. 2023. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems* 36 (2023), 75354–75371.
- [5] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2021. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–44.
- [6] Wendy Kan Meghan O’Connell, moreiraMatias. 2015. ECML/PKDD 15: Taxi Trajectory Prediction (I). <https://kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i>
- [7] OpenStreetMap Contributors. 2024. OpenStreetMap. <https://www.openstreetmap.org>
- [8] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
- [9] Guang Wang, Xiuyuan Chen, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. Experience: Understanding long-term evolving patterns of shared electric vehicle networks. In *The 25th Annual international conference on mobile computing and networking*. 1–12.
- [10] Sheng Wang, Zhifeng Bao, J Shane Culpepper, and Gao Cong. 2021. A survey on trajectory data management, analytics, and learning. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.
- [11] Yuan Xu, Jiajie Xu, Jing Zhao, Kai Zheng, An Liu, Lei Zhao, and Xiaofang Zhou. 2022. MetaPTP: an adaptive meta-optimized model for personalized spatial trajectory prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2151–2159.
- [12] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (San Jose, California) (GIS '10). Association for Computing Machinery, New York, NY, USA, 99–108.
- [13] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. *Geolife GPS trajectory dataset - User Guide*. <https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/>
- [14] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Jianqiao Yu. 2023. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.