Massive Trajectory Data Based on Patterns of Life

Hossein Amiri
Emory University, USA
hossein.amiri@emory.edu

Shiyang Ruan
George Mason University, USA
sruan@gmu.edu

Joon-Seok Kim
Oak Ridge National Laboratory, USA
kimj1@ornl.gov

Hyunjee Jin
Siemens, USA
hyunjee.jin@siemens.com

Hamdi Kavak
George Mason University, USA
hkavak@gmu.edu

Andrew Crooks
University at Buffalo, USA
atcrooks@buffalo.edu

Dieter Pfoser
George Mason University, USA
dpfoser@gmu.edu

Carola Wenk
Tulane University, USA
cwenk@tulane.edu

Andreas Züfle
Emory University, USA
azufle@emory.edu

ABSTRACT

Individual human location trajectory and check-in data have been the driving force for human mobility research in recent years. However, existing human mobility datasets are very limited in size and representativeness. For example, one of the largest and most commonly used datasets of individual human location trajectories, GeoLife, captures fewer than two hundred individuals. To help fill this gap, this Data and Resources paper leverages an existing data generator based on fine-grained simulation of individual human patterns of life to produce large-scale trajectory, check-in, and social network data. In this simulation, individual human agents commute between their home and work locations, visit restaurants to eat, and visit recreational sites to meet friends. We provide large datasets of months of simulated trajectories for two example regions in the United States: San Francisco and New Orleans. In addition to making the datasets available, we also provide instructions on how the simulation can be used to re-generate data, thus allowing researchers to generate the data locally without downloading prohibitively large files.

ACM Reference Format:
Hossein Amiri, Shiyang Ruan, Joon-Seok Kim, Hyunjee Jin, Hamdi Kavak, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. 2023. Massive Trajectory Data Based on Patterns of Life (Data and Resources Paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625592

1 INTRODUCTION

Managing, analyzing, understanding, querying, mining, and predicting human mobility is of great significance for a variety of applications ranging from disaster response [6], environmental sustainability [4], public health [13], urban planning [16], traffic management [17], and activity-based intelligence [3] to name but a few. Datasets that enable these applications mainly comprise location *trajectory data* and *location-based social network (LBSN) data*. Trajectory data captures the location of individuals at a relatively high frequency, such as at 1Hz (one location per second per individual). LBSN data captures both 1) so-called check-ins, that is, the location of individuals only when a point of interest (POI) such as a restaurant is visited as well as 2) a social network between individuals. However, publicly available real-world trajectory and LBSN data sets exhibit certain weaknesses:

Non-Representativeness: For trajectory data, the most commonly used data set is the GeoLife dataset [19] which captures the locations of 178 individuals in Beijing, China. It becomes very challenging to infer patterns of human behavior from such a small sample as it is not a representative sample of the large population of Beijing. In all existing check-in datasets that capture individuals' check-ins, the vast majority of users have less than ten check-ins [11]. Past research analyses of LBSN data concludes that "Researchers working with LBSN data sets are often confronted by themselves or others with doubts regarding the quality or the potential of their data sets." and that "it is reasonable to be skeptical"[9].

Small size: Existing data sets of human mobility are small [7]. They tend to only cover a short period of time, a small number of users, or a small number of check-ins. Using such small datasets, it becomes difficult to assess the scalability of methods and algorithms to an entire population of an area and to a long time duration.

Privacy concerns: Individual human mobility data is considered Personal Identifiable Information (PII) as it allows one to trace an individual's identity. Acquiring, storing, and publishing individual human mobility data require the consent of individuals. Even if such consent is given, users may later revoke this consent, for instance, by deleting their LBSN account. This limits, for good reasons, our ability to acquire additional individual-level human mobility data.

No ground-truth: It is a big challenge, in existing human mobility data, to infer the underlying behaviors that lead to an individual's decision to visit a place. For example, did an individual visit a restaurant to eat by themselves? To meet a friend? To have a business meal? Or to work at the restaurant? What preferences lead an individual to choose one grocery store over another? But without knowing the underlying human behavior that led to the observed mobility, it is difficult to confidently infer patterns of human behavior and to predict future mobility.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

To overcome these weaknesses, we developed a LBSN simulation capable of creating multiple artificial but socially plausible, large-scale trajectory and LBSN data sets in [7]. These large and dense data sets allow the broader social and data science research communities to test human mobility-based hypotheses without encountering issues pertaining to data representativeness, data sparsity, privacy concerns, and lack of ground-truth. These datasets enable investigation of research questions that is not currently possible given the limitation of existing real-world data sets. In addition to making these datasets available, this Data and Resources paper provides the underlying simulation framework enriched by instructions on how to regenerate the datasets locally (thus avoiding the transfer bottleneck for very large datasets) and how to generate new datasets for new study regions.

In the remainder of this paper and beginning in Section 2, we provide an overview of existing trajectory and LBSN datasets and generation tools. In Section 3, we briefly sketch the functionality of the Patters of Life simulation described in [7, 20]. We then present the generated datasets by explaining their structure and providing descriptive statistics in Section 4. Then, Section 5 provides instructions to regenerate the datasets and to help users to create new datasets for new study areas and Section 6 concludes the paper.

The two resources offered by this Data and Resources paper are:

- **Simulated datasets of human mobility** that ran more than 927 hours to capture more than 22,360,320,024 trajectory locations, more than 423,609,129 check-ins, and more than 1,736,701,154 social links. The total size of the provided datasets exceeds 1,528 GB. Datasets are shared at https://osf.io/gbhm8/ and documented at https://github.com/azufle/pol.
- **Source code of the simulation** found at https://github.com/azufle/pol. It includes documentation and parameter files to re-run the simulation locally and regenerate the aforementioned datasets (or even larger datasets) locally without having to transfer large datasets over the web.

2 PRIOR DATASETS AND SIMULATORS

Location Trajectory Data: The most commonly used real-world trajectory data set is the Geolife GPS trajectory dataset [19], which was collected and shared by Microsoft Research Asia. This dataset captures detailed trajectories of 178 users in Beijing, China, over a period of over four years (from April 2007 to October 2011). Furthermore, the dataset captures a broad range of users' movements, including not only life routines like going home and going to work but also some entertainment and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. While this dataset is excellent in terms of quality and fidelity, it is, unfortunately, very small. It is difficult to infer broad mobility patterns from a set of only 178 users, especially in a large city such as Beijing.

The second dataset is the T-Drive trajectory data sample [18], which was (also) collected and shared by Microsoft Research Asia and captures one-week trajectories of 10,357 taxis in Bejing, China. While the number of individuals captured in this dataset is much larger than in GeoLife, T-Drive captures the trajectories of taxis, not individuals. Thus, consecutive trajectories of the same taxi may not correspond to the same passenger. While useful for applications such as traffic prediction, this dataset is very limited in terms of providing insights into individual human mobility and behavior, as it is impossible to understand the sequence of places that individuals have visited.

LBSN Data: A summary of existing LBSN datasets having up to 2.7M users and up to 90M check-ins is given in [7]. It has been shown that after removing users with less than 15 check-ins and removing locations with less than ten visitors, more than half of the check-ins are eliminated [10]. Further, these datasets are captured over years, thus capturing less than 50K check-ins per day. Since these datasets are collected globally, this leaves only hundreds of check-ins per city per day.

Synthetic Data: The problem of using sparse and noisy real-world LBSN data has already been identified in previous work (e.g., [1, 2, 8]). However, none of these works have proposed a way to obtain plausible check-in data. For example, [1, 2, 8] generated user-location check-ins at random using parametric distributions without considering the semantics of the movement. While [15] created additional check-ins by replication of Gowalla and Brightkite data, thus creating more data for run-time evaluation purposes but without creating more information.

One could, therefore, question whether the experimental results of existing works on LBSN and trajectory data may be considered conclusive, both in terms of scalability and effectiveness due to a lack of large-scale available data sets [9]. This paper aims to close this gap by proposing the means to generate large-scale and ground-truth based synthetic data sets through simulation, described in the following.

3 PATTERNS OF LIFE SIMULATION

Our approach for the generation of massive and realistic human mobility data (trajectories, check-ins, and social networks) is based on a socially plausible agent-based simulation called Urban Life [20]. Urban Life is an agent-based city-level simulation in which each agent represents a simulated human in the real-world that follows socially plausible patterns of life. The simulation allows to leverage real-city environment data (road network, buildings, apartments) leveraging a pipeline to extract data from OpenStreetMap (OSM) detailed in Section 5.4. Agents in Urban Life commute between their home and work locations. They go to restaurants to eat, and they go to recreational sites to meet friends and socialize. A social network that captures friendship, family, and co-worker relationships evolves as agents interact with each other over time.

Agent behavior is driven by Maslowian needs [12] such as physiological needs (shelter, food), financial needs (money), and love needs (friends, family). These needs drive the decision-making of agents that lead to behavior to satisfy the needs, leading to an emerging behavior in which agents find a balance between spending time and making money, meeting friends, and satisfying other needs. An in-depth description of the Urban Life simulation can be found in [20] and the Java-based source code of the simulation can be found on GitHub at https://github.com/gmuggs/pol.

4 DATASET DESCRIPTION

Here, we provide summary statistics for a number of simulated LBSN and trajectory datasets that we provide for benchmarking. The datasets, as well as additional documentation, can be found at OSF (https://osf.io/gbhm8/). Details on how to regenerate these datasets and how to leverage the Patterns of Life simulation to new study regions can be found in Section 5.

Table 1: Simulated Location-Based Social Network and Trajectory Data

Settings			Check-Ins		Social Links		Trajectory Points		Runtime
Map	#Agents	#Days	Count	Size (MB)	Count	Size (MB)	Count	Size (GB)	(hours)
GMU	1,000	450	2,442,934	175	9,768,020	269	129,600,000	8.8	3.16
GMU	1,000	3,630	19,109,109	1,433	85,828,108	2,457	1,045,440,000	71	22.33
GMU	1,000	7,230	37,952,610	2,764	171,906,382	4,812	2,082,240,000	141	46.12
GMU	3,000	450	7,369,221	532	30,203,582	872	388,800,000	27	11.66
GMU	5,000	450	12,314,442	890	56,073,810	1,637	648,000,000	45	31.96
GMU	10,000	450	24,168,718	1,843	123,572,434	3,788	1,296,000,000	89	49.57
NOLA	1,000	450	2,473,475	176	9,529,160	262	129,600,000	8.8	3.15
NOLA	1,000	3,630	19,325,948	1,433	81,377,850	2,252	1,045,440,000	71	23.96
NOLA	1,000	7,230	38,370,713	2,764	162,645,920	4,505	2,082,240,000	141	47.22
NOLA	3,000	450	7,666,739	552	27,204,936	785	388,800,000	27	14.80
NOLA	5,000	450	12,839,165	926	49,217,712	1,432	648,000,000	45	21.62
NOLA	10,000	450	25,731,468	1,945	97,549,684	2,969	1296,000,000	89	53.87
ATL	1,000	450	2,481,774	176	9,169,470	252	129,600,000	8.8	6.25
ATL	1,000	3,630	19,404,487	1,433	80,419,112	2,252	1,045,440,000	71	47.50
ATL	1,000	7,230	38,526,948	2464	161,632,490	4,505	2,082,240,000	141	91.71
ATL	3,000	450	7,610,274	545	27,333,644	789	388,800,000	27	22.73
ATL	5,000	450	12,816,339	919	44,525,826	1,297	648,000,000	45	39.57
ATL	10,000	450	25,389,727	1,945	102,295,850	3174	1,296,000,000	8.8	65.93
SFCO	1,000	450	2,507,220	178	8,663,638	238	129600000	8.8	7.22
SFCO	1,000	3,630	19,662,213	1,433	74,305,732	20,448	1,045,440,000	71	54.86
SFCO	1,000	7,230	39,063,687	2,867	149,298,832	4,198	2,082,240,000	141	103.7
SFCO	3,000	450	7,651,106	547	26,694,044	771	388,800,000	27	28.54
SFCO	5,000	450	12,938,737	927	50,027,578	1,454	648,000,000	45	53.89
SFCO	10,000	450	25,792,051	1,945	97,457,318	2,867	2,082,240,000	141	85.51
Total	84000	50640	423609105	30812	1736701132	68285	23146560000	1499	936.83

Table 1 provides an overview of the generated datasets. As regions of interest for the simulation, we used four suburban and urban regions, including 1) the George Mason University Campus area, Fairfax, Virginia, 2) the French Quarter of New Orleans, Louisiana, 3) San Francisco, California, and 4) Atlanta, Georgia. For each study region we obtained road networks and building from OpenStreetMap [5]. Detailed instructions how to obtain OSM data for use in the Patterns of Life simulation (for any region of interest in the world) can be found in Section 5.4. For each of the four study regions, we run the simulation with 1K, 3K, 5K, and 10K agents for 15 months of simulation time. The first simulated month is always used as a "simulation warm-up period" and not reported as data. We also provide simulations for 10 years and 20 years, having 1K agents for each of the four regions of interest. For each dataset, three datasets are provided: 1) Check-ins, and 2) social network links are described in [7]. In addition, 3) trajectory information is provided consisting of three primary columns: simulationTime, location, and agentId. One tuple of data is provided per agent per five-minute tick. The simulation parameters allow us to provide data at more frequent intervals. However, a frequency of 1Hz would increase the (already prohibitively large) data size by a factor of 300 without including substantially more information. Location is provided using geographical coordinates.

For each of the resulting 24 datasets Table 1 shows the resulting number of check-ins, social network links (which may change over simulation time), and trajectory points.

Table 1 also reports the corresponding dataset sizes. Due to dataset larges becoming prohibitively large (to share and download) we do not provide datasets with a higher number of agents or a longer simulation time. We note, however, that instead of downloading these datasets, researchers may also run the simulation locally to reproduce the datasets, as described in the following section. This approach allows us to generate datasets of arbitrary size by scaling the number of agents and the simulation time without a data transfer bottleneck.

5 DATA RE-GENERATION INSTRUCTIONS

In this section, we outline the procedure involved in executing the simulation and producing the dataset. Specifically, we demonstrate how to create an identical dataset, as well as generating a novel dataset for a different geographical region, such as a new city or area.

5.1 Running the Simulation

The simulation Java code can be obtained via cloning from our GitHub repository (https://github.com/azufle/pol), with the project dependencies managed through Maven. Detailed instructions on setting up the simulation model and software environment are provided in the documentation of the GitHub repository.

5.2 Simulation Parameterization

The simulation includes a file named 'parameters.properties', which encompasses the default properties, including the number of agents, the map to be used, and the simulation duration. For each of datasets provided in Table 1 we provide a separate 'parameters.properties'

to allow exact reproduction of each of the datasets. Researchers are encouraged to change the parameters to simulate more agents, for a longer duration, or for a different region of interest which can be prepared using OpenStreetMap as detailed in Section 5.4. Many of the parameters that pertain to the behavior of agents and their physiological, financial, and love needs are explained in more detail in [20].

5.3 Dataset Preprocessing

Upon the completion of the simulation, generated data can be found in the 'logs' folder. Trajectory data can be extracted from the file 'AgentStateTable.tsv' which documents the state (including location) of each agent at every time step. Trajectory information is obtained by projecting the attributes simulationTime, location, and agentId. Additionally, it's important to note that the simulation includes a "warm-up period" of one month of simulation time. During this period, agents may behave differently as their social network has not yet formed and their simulated patterns of life have not yet converged. Thus, it is recommended to discard the first 30 days of simulation data.

5.4 New Map Creation

We provide our datasets for four cities across the United States. However, researchers may wish to simulate different study regions across the world. Here, we provide brief instructions to simulate and study regions by leveraging OpenStreetMap (OSM) data. Detailed instructions can be found in our Github repository (https://github.com/azufle/pol) in a separate file 'documentation/map.md'. The process begins with the extraction of data from open map services, such as Overpass Turbo[14]. Three geographic files are necessary for each simulation: buildings.shp, buildingUnits.shp, and walkways.shp. The buildings.shp file contains building footprints, the buildingUnits.shp file holds details about building units, and the walkways.shp file provides information on the transportation network. Each of these files requires additional attributes for simulation purposes. In the buildings.shp file, building type helps distinguish between residential and non-residential buildings. Road segments bear a 'function class' (fclass) tag, helping classify buildings as residential or non-residential. The buildingUnits.shp file, derived from building footprints, contains points representing possible locations for homes and workplaces. These can be obtained from OSM or created based on the size of a building. The attribute 'building' in this file maps a unit to the corresponding building.

6 CONCLUSIONS

In this Data and Resources paper, we provide very large sets of simulated individual-level trajectory and location-based social network data. Our datasets are orders of magnitude larger than existing real-world datasets. We also share the source code of the simulation along with parametrization files to allow the community to regenerate the data locally without the bottleneck of transferring hundreds of GB of data. This aspect helps to include researchers from countries that may not have as ubiquitous access to high-speed internet. By providing these datasets, we address the limitations inherent in publicly available real-world trajectory and location-based social network (LBSN) data sets, which include issues of data representativeness, data sparsity, privacy concerns, and a lack of ground truth. While our datasets are simulated, they are based on real-world map data (roads, buildings, units) and based on socially plausible patterns of life following robust social theories of human behavior. In making these datasets available and providing the means for the generation of new datasets, we hope to facilitate more extensive and reliable research in human mobility data science. Our datasets will enable scalability evaluation for human mobility data science and they will enable new research towards activity-based intelligence [3] by providing large mobility datasets that enriched with a ground truth of underlying behavior. For example, the simulated data may enable research towards the classification of different types of agents (adults, children, retirees) based on their mobility or may allow researching the early-detection of abnormal trajectories such as agents that exhibit a changed behavior due to exposure to an infectious disease. For future work, an important aspect will be adding realistic movement on top of the realistic behavior as, currently, all agents move at a constant velocity on a shortest path between their current location and their destination. In addition, allowing agents to have interactions outside of places of interest (on the road network and via telecommunication) will be an additional step towards more realistic behavior.

REFERENCES

[1] N. Armenatzoglou, R. Ahuja, and D. Papadias. Geo-social ranking: functions and query processing. *VLDB Journal*, 24(6):783–799, 2015.

[2] N. Armenatzoglou, S. Papadopoulos, and D. Papadias. A general framework for geo-social query processing. *Proc. of the VLDB Endowment*, 6(10):913–924, 2013.

[3] P. Biltgen, T. Bacastow, T. Kaye, and J. Young. Activity-based intelligence: Understanding patterns-of-life. *USGIFs State & Future of GEOINT Report*, 2017.

[4] C. Guo, B. Yang, O. Andersen, C. S. Jensen, and K. Torp. Ecomark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. *GeoInformatica*, 19:567–599, 2015.

[5] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

[6] B. Hong, B. J. Bonczak, A. Gupta, and C. E. Kontokosta. Measuring inequality in community resilience to natural disasters using large-scale mobility data. *Nature communications*, 12(1):1870, 2021.

[7] J. S. Kim, H. Jin, H. Kavak, O. C. Rouly, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle. Location-based social network data generation based on patterns of life. In *MDM*, pages 158–167, 2020.

[8] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *ICDE*, pages 450–461. IEEE, 2012.

[9] M. Li, R. Westerholt, H. Fan, and A. Zipf. Assessing spatiotemporal predictability of lbsn: a case study of three foursquare datasets. *GeoInformatica*, pages 1–21, 2016.

[10] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *CIKM*, pages 733–738. ACM, 2013.

[11] Y. Liu, T.-A. N. Pham, G. Cong, and Q. Yuan. An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proc. VLDB Endowment*, 10(10):1010–1021, 2017.

[12] A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

[13] M. F. Mokbel, L. Xiong, and D. Zeinalipour-Yazti. Introduction to the special issue on contact tracing. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 8(2):1–2, 2022.

[14] OpenStreetMap contributors. Overpass turbo. https://overpass-turbo.eu/, 2023.

[15] M. A. Saleem, X. Xie, and T. B. Pedersen. Scalable processing of location-based social networking queries. In *MDM*, volume 1, pages 132–141. IEEE, 2016.

[16] D. Wang, Y. Fu, P. Wang, B. Huang, and C.-T. Lu. Reimagining city configuration: Automated urban planning via adversarial learning. In *ACM SIGSPATIAL*, pages 497–506, 2020.

[17] H. Yuan and G. Li. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering*, 6:63–85, 2021.

[18] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *ACM SIGSPATIAL*, pages 99–108. ACM, 2010.

[19] Y. Zheng, X. Xie, W.-Y. Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[20] A. Züfle, C. Wenk, D. Pfoser, A. Crooks, J.-S. Kim, H. Kavak, U. Manzoor, and H. Jin. Urban life: a model of people and places. *Computational and Mathematical Organization Theory*, 29(1):20–51, 2023.