GeoLife+: Large-Scale Simulated Trajectory Datasets Calibrated to the GeoLife Dataset

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

Richard Yang
richard.yang@emory.edu
Emory University, Atlanta, USA

Andreas Züfle
azufle@emory.edu
Emory University, Atlanta, USA

Abstract

Analyzing individual human trajectory data helps our understanding of human mobility and finds many commercial and academic applications. There are two main approaches to accessing trajectory data for research: one involves using real-world datasets like GeoLife, while the other employs simulations to synthesize data. Real-world data provides insights from real human activities, but such data is generally sparse due to voluntary participation. Conversely, simulated data can be more comprehensive but may capture unrealistic human behavior. In this Data and Resource paper, we combine the benefit of both by leveraging the statistical features of real-world data and the comprehensiveness of simulated data. Specifically, we extract features from the real-world GeoLife dataset such as the average number of individual daily trips, average radius of gyration, and maximum and minimum trip distances. We calibrate the Pattern of Life Simulation, a realistic simulation of human mobility, to reproduce these features. Therefore, we use a genetic algorithm to calibrate the parameters of the simulation to mimic the GeoLife features. For this calibration, we simulated numerous random simulation settings, measured the similarity of generated trajectories to GeoLife, and iteratively (over many generations) combined parameter settings of trajectory datasets most similar to GeoLife. Using the calibrated simulation, we simulate large trajectory datasets that we call GeoLife$^+$, where $^+$ denotes the Kleene Plus, indicating unlimited replication with at least one occurrence. We provide simulated GeoLife$^+$ data with 182, 1k, and 5k over 5 years, 10k, and 50k over a year and 100k users over 6 months of simulation lifetime.

CCS Concepts

• **Information systems → Geographic information systems** ; **Location based services**.

Keywords

GeoLife, Patterns of Life, Simulation, Realistic Trajectory Datasets

ACM Reference Format:
Hossein Amiri, Richard Yang, and Andreas Züfle. 2024. GeoLife+: Large-Scale Simulated Trajectory Datasets Calibrated to the GeoLife Dataset. In *7th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation (GeoSim'24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3681770.3698573

GeoSim'24, October 29-November 1, 2024, Atlanta, GA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1149-7/24/10
https://doi.org/10.1145/3681770.3698573

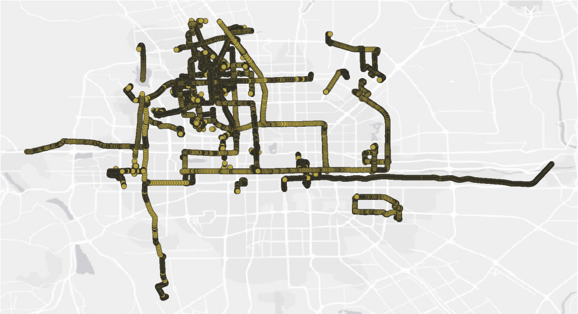

Figure 1: GeoLife dataset visualization for the busiest day in Beijing, China

1 Introduction

Trajectory data [1, 3] is essential for analyzing human behavior [20, 28] and mobility analysis [18], traffic analysis [6], providing insights that are valuable for applications like outlier detection [2, 25, 26], tracking and modeling the spread of infectious diseases [2, 13], and urban planning [10]. Analyzing location data derived from real human mobility can lead to better-informed decisions, but real-world data is difficult to obtain due to location privacy [11, 14, 21]. Consequently, openly available trajectory datasets are collected by usually small numbers of volunteers. The largest and most commonly used trajectory dataset is the GeoLife Dataset [27] which captures 182 individuals in Beijing, China over more than five years. However, in GeoLife only 45 users have more than 100 staypoints, indicating that the dataset is sparse and patterns focuses primarily on its major users. In addition, the maximum number of GeoLife users active on any given day is only 25 in Beijing. Figure 1 shows the trajectory of these 25 users on one of the busiest GeoLife days. We see that these trajectories are distributed very sparsely over the Beijing area. Given the population of Beijing of more than 20 million people, it appears impossible to find representative patterns of human mobility from no more than 25 users. This makes it very challenging to employ GeoLife for the many applications.

Due to these limitations, researchers often employ simulated trajectory data: it offers a richer and cleaner dataset without privacy concerns. However, many existing trajectory simulations have agents travel to uniformly random destinations [5, 8] or use parametric distributions to generate trip destinations [4, 15]. While such trajectories are useful for evaluating database storage and retrieval solutions for trajectories, the randomness of mobility does not allow to improve our understanding of human mobility. Towards a more realistic simulation of human mobility, the Pattern of Life Simulation (POL) [12, 30] simulates the physiological, safety, love, and esteem needs of individual agents to create trips with a purpose such as going home to sleep, going to work to make money, or going to a restaurant to eat.

In this work, we leverage the Patterns of Life simulation to generate a set of datasets that are statistically similar to GeoLife's active user data. This approach allows us to 1) maintain the socially plausible agent behavior inherited from the Patterns of Life simulation, 2) while having descriptive statistics of the trajectory dataset closely reflecting the original GeoLife data, and 3) having a substantially higher density of data points compared to GeoLife. To identify the optimal parameter configuration for the Patterns of Life Simulation to generate GeoLife-like data, we developed a genetic algorithm that initially explores random simulation parameters and iteratively "crosses" parameter settings having high similarity to GeoLife. Our primary contributions in this Data & Resource paper is as follows:

- We provide multiple trajectory datasets collected over 5 years of simulation periods for 182, 1k, and 5k agents, 1 year for 10k and 50k agents, and 6 months for 100k agents. These datasets sized at hundreds of gigabyte of trajectories are available at OSF link provided on https://github.com/onspatial/geolife-plus
- A genetic algorithm used to calibrate parameters based on the GeoLife dataset, with instructions on how to apply various statistical methods to generate additional datasets. The source code can be accesses on the GitHub repository.
- The simulation parameters resulting from this geometric algorithm and used for the generation of the aforementioned datasets. Together with our documentation to re-run the simulation, this allows users to re-generate the data without having to download large datasets and allows users to create even larger datasets having more agent or having longer simulation periods. Parameters, configurations, and documentations are available on GitHub.

The remainder of the paper is organized as follows: In Section 2, we review the existing trajectory datasets, both real and synthetic. In Section 3, we explain our methodology, including data generation, the genetic algorithm, and the steps to create the dataset. In Section 4, we describe and analyze the generated data. In Section 5 regeneration process is described. Finally, we summarize our findings and draw conclusions in Section 6.

2 Existing Trajectory Datasets

Trajectory datasets have been widely studied previously. Trajectories can be recorded for humans, animals, vehicles, etc., using sensors in real-world settings [22, 27] or synthesizing the data using simulations [3], generative models [24] and dataset enhancement [9, 28, 29]. In this section, we describe the most commonly used real-world and synthesized individual human trajectories.

2.1 Real World Human Trajectory Data

The GeoLife GPS trajectory dataset [27], collected by Microsoft Research Asia, is one of the most commonly used real-world trajectory datasets. It encompasses trajectories of 180 users in Beijing, China, over a period of more than four years (from April 2007 to October 2011). The dataset captures a wide range of movements, not only routine activities like commuting to work or returning home, but also leisure activities such as shopping, sightseeing, dining, hiking, and cycling. Although the dataset is of high quality and fidelity, its relatively small size of only 180 users makes it challenging to infer broad mobility patterns, particularly in a large urban area like Beijing. A large-scale real-world human mobility dataset was introduced recently in [22]. The data were collected from mobile phones on a metropolitan scale and encompasses observations from 100,000 individuals over a period of 75 days. Data collection was conducted with user consent, and to ensure privacy, all data was anonymized. To further protect privacy, the data is structured into a spatial grid having spatial cells measuring 500 meters by 500 meters, and data is recorded at 30-minute intervals. This relatively low spatial granularity allows understanding high-level human mobility, but can't be used to infer visit patterns at individual places of interest. In addition, numerous trajectory datasets do not pertain to individual humans, but to vehicles such as taxi trajectories in Beijing, China [23] and San Francisco [19], and bus trajectories in Rio de Janeiro, Brazil [7]. While such datasets can be used to understand traffic patterns based on travel speeds, it is difficult to infer human mobility patterns from such data, as a taxi may capture a different and independent human passenger each trip.

2.2 Synthetically Generated Trajectories

Deep generative models have been proposed recently in [24] to addressing the scarcity of large real datasets. Their "End-to-End Trajectory Generation with Spatiotemporal Validity Constraints" (EETG) framework significantly improves trajectory synthesis, showcasing the similarity of generated trajectories to real-world trajectories. Towards dataset enhancement, a generic optimization-based approach was introduced recently [28], which utilizes both position and velocity data as a baseline to enhance driving trajectory data. In [29] a diffusion model is utilized to synthesize the spatial-temporal behavior of the original dataset accurately. This model learns complex spatial-temporal motion patterns and emulates the geographical distribution and statistical properties of real-world trajectories. However, the goal of these generated datasets is to mimic kinematic trajectory features rather than creating human mobility patterns.

The Patterns of Life Simulation [12] allows generating city-level human mobility data. The simulation uses data from OpenStreetMap to model agents moving between various locations like home, work, restaurants, and recreational sites. Agents' activities are guided by Maslowian needs [17], including basic physiological needs, financial needs, and social needs, which drive their decisions and interactions. A detailed description of the simulation can be found in [30]. In [3], the Patterns of Life Simulation was used to generate a massive trajectory dataset. The dataset comprises over 1.5 terabytes of simulated data, which includes more than 22,360,320,024 trajectory locations, over 423,609,129 check-ins, and more than 1,736,701,154 social links. The Patterns of Life Simulation has a very large number of parameters to define the agents' needs and consequently, their behavior. While the default parameters used for this dataset are realistic, we do not have a clear understanding how these parameters may differ at different places around the world. In this paper, we fill this gap by calibrating the Patterns of Life Simulation to GeoLife to find near-optimal parameter settings to reproduce features observed in GeoLife data for Beijing.

3 Simulation Calibration

In this section, we explain our approach to generate the GeoLife* dataset. Specifically, we describe the similarity function we calculated to assess the similarity of the original GeoLife dataset and the generated datasets in Section 3.1. Based on this similarity function, we describe the genetic algorithm developed to calibrate simulation parameters in Section 3.2, and we provide the results of this simulation calibration in Section 3.3.

3.1 Trajectory Data Similarity

As GeoLife trajectories travel outside of Beijing (and even of China), we restricted the trajectories to the Beijing area with the bounding box of [39.748, 116, 165, 40.038, 116.628]. Within this region, we first extracted staypoints using the algorithm described in [16] to find the places that users visit and delineate the trips connecting these places. For this calibration, we only use GeoLife users at least 100 staypoints. This filter yielded approximately 12,000 staypoints for 45 users. For these users, we computed the average distance per trip (ADT) as $3692.13m$, the average distance per agent per day(ADA) as $4474.59m$, the maximum trip distance (MXD) as $30262m$ and the median trip distance (MDD) as $3349.75m$. We applied the same metrics to the simulated data. For each set of simulated data, we used the formula in Equation 1 to score the similarity of the simulated check-ins to the GeoLife check-ins. A score closer to 1 indicates greater similarity between the simulated and GeoLife check-ins.

$$Similarity(G,P) = 1 - \frac{1}{|M|}\sum_{k\in M}\frac{|k(P)-k(G)|}{k(G)} \tag{1}$$

where G is the GeoLife dataset, P is a set of simulated trajectories, $M = \{ADT, ADA, MXD, MDD\}$ represents the set of metrics, $k(P)$ represents the results of metric k on the Patterns of Life dataset, and $k(P)$ is the result of metric k on GeoLife.

3.2 Genetic Algorithm for Calibration

Initially, we manually identified 63 simulation parameters we deemed relevant in defining the agents' behavior. These parameters included factors such as the number of agents' interests, the maximum allowed rental salary ratio, and the agents' walking speed. The full list of all parameters used for calibration can be found on our GitHub repository. Our goal was to find values for these parameters that yield a simulation of Beijing that most closely mimics the trajectory metrics observed on GeoLife.

We initialize the algorithm using n ("layer size") simulation runs using randomly chosen (within manually chosen ranges deemed plausible) parameter values. For each trajectory dataset by a simulation, we used Equation 1 to find parameter settings yielding the top simulations most similar to GeoLife. These simulations initialized the genetic algorithm in which these parameter settings ("parents") were combined into in five different ways to create new parameter settings ("child") by choosing, for each parameter, at random one of the following: 1) the maximum parameter value of the selected parents, 2) the minimum parameter value of the selected parents, 3) the mean parameter value of the selected parents, 4) random combinations of the values from the selected parents, or 5) a new parameter value chosen at random ("mutation"). This random combination of attributes was repeated until n children were generated. For each of these n children, a corresponding simulation (with the selected parameter setting) was run. Then, Equation 1 was again used to find the children yielding trajectory data most similar to GeoLife. Using these children, this process is repeated to create a new layers ("generation") of simulations. This process of creating new generations of simulations is repeated indefinitely until manually stopped. The parameter settings (across all simulated generations) yielding the most similar trajectories to GeoLife is the chosen as the result of the calibration step. The source code of this genetic algorithm can be found on https://github.com/onspatial/geolife-plus.

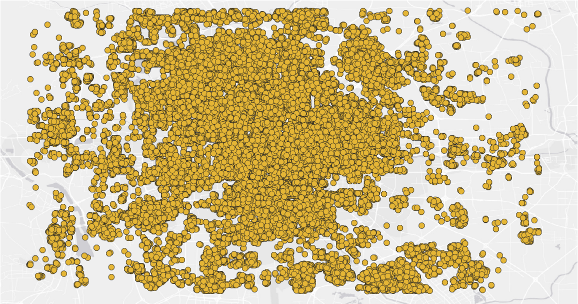

Figure 2: Staypoints of a simulation day with 50k agents.

3.3 GeoLife Calibration Results

Running the genetic algorithm with layer sizes of 8, 32, 64 and 128 yielded 10, 15, 32, and 78 configurations, respectively, that achieved similarity scores (using Equation 1) higher than 0.8. We retained the top 10 configurations for subsequent dataset scaling. These configurations are accessible at https://github.com/onspatial/geolife-plus/blob/main/restults/params.top.json. We utilize the *params.top.json* file in the subsequent phase to generate datasets and evaluate their similarities to the original GeoLife dataset across varying numbers of agents and simulation setups.

4 Dataset Description

Figure 2 shows the location of staypoints observed in a single day of GeoLife* with 50,000 agents. This figure omits the trajectories of agents between staypoint. We observe that agents visit a location in Beijing across the entire city, giving a much more representative coverage than a day of GeoLife shown in Figure 1.

The datasets generated in this study, as detailed in Table 1 and Table 2, capture a range of simulated scenarios based on the original GeoLife dataset, scaled to various agent populations and time periods. Table 1 shows the size (data size, # staypoints, # GPS updates) for our generated GeoLife* datasets and the original GeoLife dataset. Running GeoLife* with the same number of agents and duration (182 users, 5 Years), we observe that GeoLife* is approximately 100 times larger. This is because in GeoLife* all 182 agents are fully observed every day, whereas users in GeoLife become inactive for long periods. The number of GPS Updates in GeoLife* is only about five times larger than GeoLife. That's because GeoLife* uses five-minute frequency location updates (to minimize storage cost), whereas GeoLife users are observed every 1–5 seconds. We note that the sampling frequency of the Patterns of Life simulation can be changed as a parameter. But we see that by simulating 1k agents for five simulation years, the dataset already grows to 180GB even at a 1/300Hz sampling frequency.

Table 2 additional shows the metrics we used to calibrate the simulation for each of the generated datasets and the resulting similarity score defined in Equation 1. As the number of agents increases, we observe that the average distance per agent per day (ADA) decreases. We explain this due to the simulation creating more recreational sites to accommodate the large number of agents. Thus, agent are able to find recreational sites closer to their home. At the same time, the agent's work location (which is usually further than recreational sites and, thus, defined the agents' radii of gyration) remains similar as the number of agents increases.

Additionally, we present the dataset titled "59k-5yrs," a carefully curated compilation of the top 300 most similar generated datasets

	Staypoints		GPS Updates		
Dataset	Size	Number	Size	Number	Time
GeoLife	1.7 MB	19K	1.7GB	28M	N/A
182-5yrs	178 MB	2.2M	30GB	120M	1.72h
1k-5yrs	0.6GB	8M	180GB	600M	8.2h
5k-5yrs	3.1GB	39M	NP	NP	18.46h
10k-1yr	1.2GB	15.5M	NP	NP	11.75h
50k-1yr	16GB	200M	NP	NP	127.45h
100k-6mo	16GB	200M	NP	NP	139.42h
59k-5yrs	41GB	575M	NP	NP	NA

Table 1: Specification of GeoLife scaled datasets. Each dataset name follows the format [#agents]-[simulated time]. For example, "1k-5yrs" refers to a dataset where 1,000 agents were simulated for 5 years. (K: Thousnd, M: Million, B: Billion, T: Trillion). NP: Not provided because the size was very large and logging the data was storage and time-consuming

Dataset	ADT	ADA	MXD	MDD	Score
GeoLife	3692.13	4474.59	30262	3349.75	1
182-5yrs	4217.12	4100.51	29013.0	3346.0	0.93
1k-5yrs	3557.37	3445.18	34365.0	2815.0	0.85
5k-5yrs	1217.29	1209.29	29202.0	885.0	0.45
10k-1yr	909.34	910.11	29679.0	667.0	0.40
50k-1yr	475.63	473.67	31978.0	369.0	0.32
100k-6mo	427.96	425.79	38612.0	326.0	0.25
59k-5yrs	5073.91	4884.97	39886.0	4165.0	0.74

Table 2: Geo-statistics of the generated datasets.

from the "182-5yrs" series. These datasets have been selected based on their high similarity, resulted in a total similarity score of 0.74 to the original Geolife dataset. The "59k-5yrs" dataset, with a total size of 41GB, contains approximately 575 million staypoints.

5 Data Sharing and Re-Generation

We're sharing the datasets sized smaller than 100GB on `OSF.io`. Links to these datasets can be found on our GitHub repository. For datasets larger than 100GB, our GitHub documentation provides instructions how to run the simulation and locally re-generate the data. This will allow researchers to generate even larger (in terms of the number of agents or simulation duration) datasets. For this purpose, Table 1 also shows the wall-close time the simulation took to run on a compact desktop machine having a 2.40Ghz i5-1135G7 processor with eight cores and 16GB of main memory running Linux Fedora.

6 Conclusions

In this paper, we have presented a novel approach to generate large-scale synthetic geospatial datasets by simulating the pattern of life of individuals with the consideration of real-world constraints. We have demonstrated the effectiveness of our approach by generating a set of synthetic datasets based on the GeoLife dataset. We have shown that the generated datasets exhibit similar statistical properties to the original dataset, and can be used for various geospatial data analysis tasks. We have also provided detailed instructions on how to reproduce the generated datasets, and have made the code and data available on GitHub. With the provided datasets and code, researchers can easily generate large-scale synthetic geospatial datasets for their research purposes and evaluate the performance of their algorithms on realistic data.

References

[1] Hossein Amiri, Will Kohn, et al. 2024. The Patterns of Life Human Mobility Simulation. arXiv:2410.00185

[2] Hossein Amiri, Ruochen Kong, and Andreas Zufle. 2024. Urban Anomalies: A Simulated Human Mobility Dataset with Injected Anomalies. arXiv:2410.01844

[3] Hossein Amiri, Shiyang Ruan, et al. 2023. Massive Trajectory Data Based on Patterns of Life. In *SIGSPATIAL'23*. ACM, 1–4.

[4] Nikos Armenatzoglou, Stavros Papadopoulos, and Dimitris Papadias. 2013. A general framework for geo-social query processing. *Proc. of the VLDB Endowment* 6, 10 (2013), 913–924.

[5] Thomas Brinkhoff. 2002. A framework for generating network-based moving objects. *GeoInformatica* 6, 2 (2002), 153–180.

[6] Wenqiang Chen, Tao Wang, et al. 2022. Lane-based Distance-Velocity model for evaluating pedestrian-vehicle interaction at non-signalized locations. *Accident Analysis & Prevention* 176 (2022), 106810.

[7] Daniel Dias and Luis Henrique Maciel Kosmalski Costa. 2018. CRAWDAD dataset coppe-ufrj/RioBuses (v. 2018-03-19).

[8] Christian Düntgen, Thomas Behr, and Ralf Hartmut Güting. 2009. BerlinMOD: a benchmark for moving object databases. *The VLDB Journal* 18 (2009), 1335–1368.

[9] Xiangwang Hu, Zuduo Zheng, Danjue Chen, et al. 2022. Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies* 134 (2022), 103490.

[10] Sibren Isaacman, Richard Becker, et al. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 239–252.

[11] Ali Khoshgozaran, Cyrus Shahabi, and Houtan Shirani-Mehr. 2011. Location privacy: going beyond K-anonymity, cloaking and anonymizers. *Knowledge and Information Systems* 26 (2011), 435–465.

[12] Joon-Seok Kim, Hyunjee Jin, Hamdi Kavak, et al. 2020. Location-based social network data generation based on patterns of life. In *MDM*. IEEE, 158–167.

[13] Will Kohn, Hossein Amiri, and Andreas Züfle. 2023. EPIPOL: An Epidemiological Patterns of Life Simulation (Demonstration Paper). In *SIGSPATIAL SpatialEpi'23 Workshop*. ACM, 13–16.

[14] John Krumm. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13 (2009), 391–399.

[15] Justin J Levandoski, Mohamed Sarwat, Ahmed Eldawy, and Mohamed F Mokbel. 2012. Lars: A location-aware recommender system. In *ICDE*. IEEE, 450–461.

[16] Quannan Li, Yu Zheng, Xing Xie, et al. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. 1–10.

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

[18] Mohamed Mokbel, Mahmoud Sakr, , et al. 2024. Mobility Data Science: Perspectives and Challenges. *ACM Transactions on Spatial Algorithms and Systems* (2024).

[19] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD dataset epfl/mobility (v. 2009-02-24).

[20] Eran Toch, Boaz Lerner, Eyal Ben-Zion, and Irad Ben-Gal. 2019. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* 58 (2019), 501–523.

[21] Yonghui Xiao and Li Xiong. 2015. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1298–1309.

[22] Takahiro Yabe, Kota Tsubouchi, et al. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data* 11, 1 (2024), 397.

[23] Jing Yuan, Yu Zheng, Chengyang Zhang, et al. 2010. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. 99–108.

[24] Liming Zhang, Liang Zhao, and Dieter Pfoser. 2022. Factorized deep generative models for end-to-end trajectory generation with spatiotemporal validity constraints. In *SIGSPATIAL '22*. 1–12.

[25] Zheng Zhang, Hossein Amiri, et al. 2023. Large Language Models for Spatial Trajectory Patterns Mining. arXiv:2310.04942

[26] Zheng Zhang, Hossein Amiri, et al. 2024. Transferable Unsupervised Outlier Detection Framework for Human Semantic Trajectories. arXiv:2410.00054

[27] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. *GeoLife GPS trajectory dataset - User Guide* (geolife gps trajectories 1.1 ed.).

[28] Feng Zhu, Cheng Chang, Zhiheng Li, Boqi Li, and Li Li. 2024. A generic optimization-based enhancement method for trajectory data: Two plus one. *Accident Analysis & Prevention* 200 (2024), 107532.

[29] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. 2024. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. *Advances in Neural Information Processing Systems* 36 (2024).

[30] Andreas Züfle, Carola Wenk, , et al. 2023. Urban life: a model of people and places. *Computational and Mathematical Organization Theory* 29, 1 (2023), 20–51.