

HumoNet: A Framework for Realistic Modeling and Simulation of Human Mobility Network

Joon-Seok Kim
Gautam Malviya Thakur
Licia Amichi
Annetta Burger
Chathika Gunaratne
Joseph Tuccillo
Taylor Hauser
Joseph Bentley

Oak Ridge National Laboratory
Oak Ridge, TN, USA
{kimj1,thakurg,amichil,burgerag,
gunaratnecs,tuccillojv,hauserrt,bentleyjd}@ornl.gov

Kevin Sparks
Debraj De
Chance Brown
Elizabeth McBride
Jesse McGaha
James Gaboardi
Xiuling Nie
Steven Carter Christopher

Oak Ridge National Laboratory
Oak Ridge, TN, USA
{sparkska,ded1,browncc,mcbrideec,
mcgahajr,gaboardijd,niex,christophesc}@ornl.gov

Abstract—Understanding, analyzing, and predicting human mobility and dynamics are valuable to solving pressing problems, developing effective plans, and prescribing timely remedies. As a computational approach, realistic human mobility simulations allow us to understand, analyze, and predict complex systems, including human societies. Accurate simulations rely on (1) the model that captures interactions and behaviors of myriad entities in our society and (2) the mapping of model instances to real-world entities. Taking this into account, this paper introduces the **Human Mobility Network simulation framework (HumoNet)**, an integrated patterns of life (POL) simulation framework that leverages real-world data layers including transportation networks, points of interest, populations, popularity, and human trajectories. HumoNet is a data informed model in which agents are equipped with activities, locomotion, and planning capabilities. To simulate realistic kinematic maneuvers of individuals in transportation networks, HumoNet harnesses a microscopic traffic simulator that provides interaction among vehicles and traffic objects. In this paper, we describe the framework, outline our methodologies, and discuss the data processing and challenges of each data layer. Through experiments, we demonstrate that our simulations capture key features of human mobility by comparing them to the literature and real data using standard measures of human mobility (i.e., the radius of gyration, number of locations visited, level of exploration) and metrics scoring (i.e., Jensen-Shannon divergence). We envision that the synthetic data

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA). Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<https://www.energy.gov/doe-public-access-plan>). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOE, or the U.S. Government.

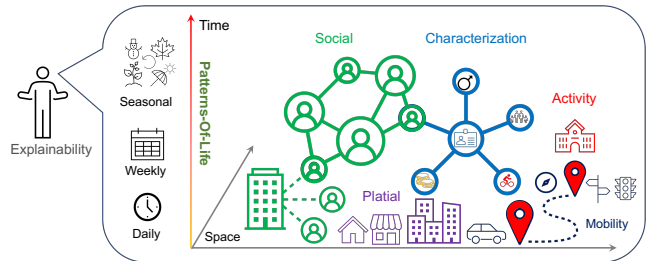


Fig. 1: Integrated Patterns of Life Framework

produced by HumoNet will serve as a benchmark for analyzing epidemics, deploying EV charging networks, and validating AI/ML tasks such as location prediction.

Index Terms—simulation, data-driven, patterns-of-life, population, point-of-interest, microscopic traffic simulation

I. INTRODUCTION

Human mobility is an emergent feature of the complex dynamics in human society, and observable mobility data provide external evidence of social causal factors that are necessary for understanding, predicting, and analyzing human systems. These observations underpin the knowledge crucial for solving today's socioeconomic and environmental issues. For instance, the dynamics of urbanization can be observed through human mobility data as population densities change across spatial landscapes [1]. Also, social phenomena emerge as a result of changes in demographics, infrastructure development, transportation, and socioeconomic activities, which are reflected in patterns of human mobility. Thus, the analysis of human mobility aids decision making and improves understanding of complex societal challenges.

Modeling and simulation (M&S) is one of the computational approaches used in various applications, including transportation, urban planning, market analysis, and disaster response. One of the strengths of M&S is explainability.

This means that M&S allows us to understand how the dynamics in a complex system result in a set of outcomes. This is accomplished by decomposing the complex system into subsystems that are digestible for processing, analysis, and validation. In transportation models, for example, it is important to capture drivers’ driving behaviors & patterns, types of vehicles, and road environments, such as the existence of lane-level road networks, traffic signs, and traffic lights. By providing information on these subsystems and their interactions, analysis of the model can illuminate how and why a system responds to specific events. Thus M&S provides the capability to design, implement, and monitor interventions [2] and support problem-solving and decision-making for practitioners and policy-makers. To that end, M&S can effectively be used for modeling human mobility patterns because it enables the detailed examination and prediction of human movement patterns through these complex systems.

The field of human mobility M&S is confronted with numerous challenges inevitably emerging from characterizations of a complex adaptive system. Critically, high-fidelity, realistic trajectory data are important observations for determining causal factors in these human systems, such as transportation, energy, urban planning, climate, disaster management, and security, but these data sets are not comprehensive and suffer from gaps and a lack of similarity. Of greater difficulty is identifying the core variables that represent the social causal drivers and their effects in a human mobility model. Accurate simulations depend on sufficiently capturing the behavior and interactions of large populations of heterogeneous actors and mapping their behavior to real-world observations. In our framework, we attempt to address these limitations with innovative methodologies for addressing the scarcity of high-fidelity data, fully characterizing the interactions of social factors that lead to emergent movement patterns, and mapping model outputs to the literature and real world data for validation.

In this paper, we propose a **Human Mobility Network** simulation framework or *HumoNet*, that provides capabilities to simulate (1) integrated individual or composite patterns of life, (2) multi-contextual models at the point of interest (POI) level, such as population demographics and human activity, (3) interactions among agents and (4) realistic human mobility within the built environment. Figure 1 illustrates core components and features of HumoNet. HumoNet is an explainable and data-driven model that captures emerging complex behaviors from diverse mobility patterns within multiple time-scales: daily, weekly, and seasonal.

In this study, we summarize our contributions as follows:

- Real world data conflation: synthetic population, road networks, points of interest, and trajectories.
- Patterns of life (POL) that maintain individual privacy.
- Micro-traffic simulations emerging from POL behavior.
- Simulation output validation with literature and real data.

The remainder of this paper is organized as follows. In Section II, we survey related studies and position our work. We introduce a workflow of how we process data for sim-

ulations in Section III. Section IV describes the architecture of HumoNet and its components. Section V reveals results of experiments of output to demonstrate realism of the simulation. In Section VI, we discuss open challenges for the improvement of simulations. Finally, we conclude our work and discuss future directions in Section VII.

II. RELATED WORK

This section describes a survey of related studies regarding human mobility, simulation, and location privacy.

A. Simulations in Human Mobility

Silverman et al. [3] describes two generations of POL models: the first generation, in which agents are modeled using a Markovian approach with rigid activity schedules that are simplistic and inflexible, yet computationally efficient and easily scalable, and the second generation, in which agent models are more flexible and exhibit diverse activities, network relations, and situated behaviors. The models provide metrics for POL goodness of fit along the dimensions of activities in daily life, social skills, and cognition. The authors introduce StateSim, a meso-scale human behavior model that includes both first- and second-generation POL agents scaled to simulations of 100,000 agents in urban defense scenarios. Kim et al. [4] model urban POL simulations in which manifested circadian rhythms and weekly patterns are governed by the first three levels of Maslow’s [5] hierarchy of needs: (1) physiological, (2) safety, and (3) belongingness and love. Amiri et al. extended their work to generate trajectory data [6] for four areas of interest. However, their models are not calibrated to real world mobility data. Pesavento et al. [7] proposed a data-driven, agent-based simulation approach, utilizing SafeGraph data to extract POI visit patterns of people within the census block group level. The underlying kinematic mobility in the study is not based on microscopic traffic simulations.

Machine learning approaches to characterizing POL in simulation can be categorized at the macro-level with focus on pedestrian flows based on density or POI occupancy over time, or at the micro-level in which the model describes the next locations or trajectories of individual agents over time [8]. Meso-level approaches also exist where the agent may represent a group of individuals. In transportation, Kashiyama et al. [9] proposed an Open PFLOW simulation based on open data including National Census and trip survey data in Japan. While Open PFLOW is based on meso-level traffic simulations, HumoNet supports high-fidelity, microscopic models including car-following and lane change models, reporting agents’ location every second all of which are driven by individual agent POL.

B. Human Mobility and Privacy

The availability of fine-grained and large-scale mobility datasets has been compared to the microscope invention, triggering extensive research on human mobility [10], [11]. Following the availability of these data, the scientific community has witnessed numerous mobility models developed in an attempt to reproduce high-fidelity, individual movements [8],

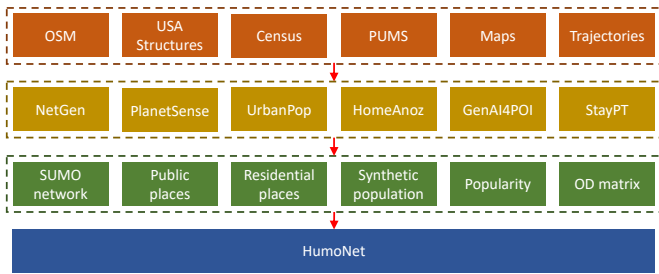


Fig. 2: Data processing for HumoNet

[12], [13]. An individual’s trajectory is typically considered an ordered sequence of visited geographical coordinates. Traditional models seek to reproduce these sequences of visits using one or several scaling properties that characterize human mobility [8], [12], [13]. However, in some cases this approach can reflect a person’s real trajectory, raising privacy concerns due to the possibility of tracing these patterns back to personal identities.

As input to data to models, methods for mobility data anonymization has historically and legally achieved a balance between protecting individual privacy and making statistical data valuable. However, the existence of a panoply of anonymization algorithms, pseudonymization techniques, and standard de-identification methods are not sufficient to prevent users from being re-identified or tracked [11], [14] from model data output. *Synthetic mobility data* generation is an emerging method that provides a viable solution to this dilemma. Learning to simulate and synthesize high-quality mobility trajectories contributes to urban behavior discovery and practitioner decision-making without personal privacy disclosure [10]. Instead of sharing real-world data, synthetic data provide trajectories and guarantee their representativeness by reproducing statistical features observed in the original data [10], [11], [14].

III. DATA SOURCES AND PROCESSING

This section introduces an overview of data processing for HumoNet (see Figure 2). The performance of realistic simulations depends on the quality of input data as well as the model itself. All real data layers are noisy and incomplete, and data processing is a necessary step for ensuring data quality. We note that the data processing for each data type and integrated layer is sophisticated in a complex system model. Because of the limited space in this paper, we will focus on a high-level data workflow description that summarizes the key ideas of our data processing. The top layer of Figure 2 shows the main data sources and types for data processing. In the next layer of the figure, six yellow boxes (i.e., NetGen, PlanetSense, UrbanPop, HomeAnoz, GenAI4POI, and StayPT) are independent data processing modules. The following subsections elaborate on how each module processes the data and produces output for HumoNet.

A. NetGen: High-fidelity road network

NetGen is a module that processes transportation network data so that HumoNet can run detailed lane-level traffic simulations. NetGen supports OpenStreetMap (OSM), which

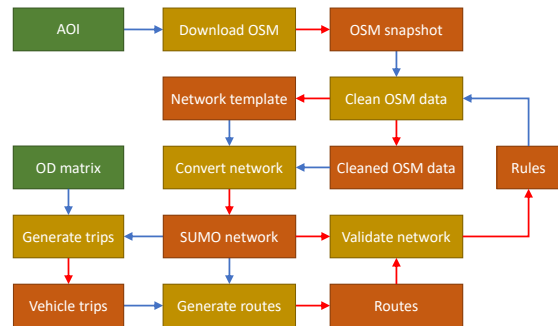


Fig. 3: Network data processing

is a free, open geographic database based on crowdsourcing and is one of the most common data sources for road networks and POIs. Similar to other datasets, these data are evolving and imperfect. OSM’s flexible schema supports the contributions of volunteers who map features for their own purposes with variations in attribute management styles. The Overpass API allows users to download certain types of data for their purpose. In our dataset, we found that low priority road types (e.g., service roads) are more error-prone (e.g., connectivity, one way, etc.) and missing attributes (e.g., the number of lanes, max speed).

Figure 3 shows the workflow of NetGen. First, NetGen downloads raw OSM data and saves it as a snapshot. In many applications, it is common to select a subset of the network within a given area of interest (AOI), entailing a clip of the network. In Section VI, we discuss issues related to sub-networks including clipping. Taking the issues (e.g., network connectivity) into account, NetGen cleans the OSM snapshot and generates cleaned data and a network template that feeds network conversion. The network template is used to handle instances with missing attributes.

Next, the network is integrated with the underlying traffic simulator of HumoNet, *Simulation of Urban MObility* (SUMO) [15]. To convert cleaned data to the SUMO network, we use `netconvert`, a command line application and part of the SUMO ecosystem. Conversion to SUMO may generate new errors, so we validate the converted SUMO network in two steps: checking full network connectivity and checking specific paths in the network by generating routes for a given set of OD matrices. During this process, we may find ambiguous and invalid cases. Without ground truth data that can be obtained in field tests, it is challenging to develop an algorithm to consistently apply in all cases, so we generate data-specific rules that can be applied as part of an iterative cleaning process.

B. UrbanPop: Synthetic population

To generate agents for human mobility modeling, HumoNet leverages UrbanPop, a toolkit for spatial microsimulation that includes population synthesis and activity modeling capabilities [16]. UrbanPop provides agents that are characterized across many subjects, including demographics, economic status, worker and student characteristics, housing, and mobility modes, and are resolved to realistic residential locations.

UrbanPop creates synthetic populations of individuals whose aggregate characteristics faithfully represent census block group level (600 – 3000 people) profiles published in the American Community Survey (ACS) Summary File (SF). Synthetic populations may be generated for any location in the United States.

UrbanPop employs the Penalized Maximum-Entropy Dasy-metric Modeling (P-MEDM) algorithm to allocate longform records from the ACS Public-Use Microdata Sample (PUMS) from large Public Use Microdata Areas (PUMAs) containing 100,000 people or more to the finer spatial scale of census block groups [17]. P-MEDM seeks to reproduce estimates within the 90% Margin of Error (MOE) for constraining variables (e.g., age, sex, industry/occupation of worker, commute mode) from the ACS SF using the PUMS. Leveraging the MOEs allows for the generation of many plausible synthetic populations without ever directly identifying individuals with these characteristics within a block group’s true population.

Agents in the synthetic population are assigned to realistic residential locations by conflating dwelling type labels pulled from the PUMS (single-family detached, single-family attached, multi-unit from 2 to 50 or more dwellings, mobile homes, informal housing, group quarters) with residential and group quarters labels from the Federal Emergency Management Agency’s (FEMA) USA Structures dataset [18]. Dwelling counts for multi-unit USA Structures are computed using Population Density Tables (PDT) [19] estimates of people per 1000 square feet in multifamily urban residential in the United States and dividing these counts by average household size in the block group. This process is also probabilistic, based on an agent’s “choice” of a dwelling based on the degree of match between a block group’s available residential structures match and a synthetic household’s dwelling type. Therefore, the residential assignment process does not leverage any personally identifiable information such as names, addresses, or property data.

C. PlanetSense: Points of interest

We use the PlanetSense data [20], that collects POIs from multiple open sources, social media, and passive and participatory sensors (IoTs, traffic cameras, detectors). Each record contains a unique POI ID, latitude, longitude, and three different levels of semantical classification. PlanetSense POI data is comprised of a variety of sources and formatted into a common schema, with the goal of providing a comprehensive data set of nonresidential places. Within PlanetSense, POIs are represented by a point geometry and include extensive attribution including name, category, address, and, when available, supplementary information such as hours of operation.

The data are further enriched through native language translation, reverse-geocoding, de-duplication, and a multi-tiered system of categorization. Address information is supplemented through reverse-geocoding using OSM Nominatim. POI names are translated into English. SONET is a semantic ontological network graph database created by PlanetSense in order to provide a uniform method of categorization across multiple

data sets [21], [22]. Whenever category information is lacking, predictive analysis is used to assign a POI type based on the POI’s name. Data de-duplication is performed by analyzing the spatial and semantic similarity of POIs and removing co-referent data.

D. GenAI4POI: Generative AI for Places of Interest

Public places are common origins and destinations of trips. Visits to places can be captured from multiple sources, such as check-in or navigation destinations. Google Maps popular times data is obtained through the PlanetSense database [20], a framework for compiling geospatial intelligence from real-time streaming and spatio-temporal analytics of open source data. However, data coverage in most regions of interest is often limited, and the popularity of places of interest must often be inferred. To address this issue, the GenAI4POI tool is utilized to infer the weekly popularity information for places of interest with no popularity data. GenAI4POI consists of a generative adversarial network (GAN) that has been trained to infer hourly popularity information of places based on features including the category for the use of the place and the region it is located in. Given this information it is able to produce a typical week of hourly popularity for the place of interest. Specifically, GenAI4POI is able to output a vector of 168 hours of the week with values of the range $[0, 1]$ representing the normalized popularity of the place, where 0 represents the place is empty and 1 indicates the place is at its full capacity. The GAN is able to infer popularity based on varying category resolution, i.e. low-resolution categories, e.g. whether a place could be considered as used for business, residential, or administrative purposes; or higher-resolution categories, e.g. if a place is a restaurant, nightclub, or park, and may be used as features instead.

E. StayPT: Stay points for origin-destination pairs

Location data are captured through mobile devices such as smart gadgets with Global Positioning System (GPS). Due to multiple types of errors (e.g., sensor, sampling, etc.), trajectory data can be biased and sparse. StayPT focuses on extracting semantic locations from raw trajectories. We consider the stay point p to stand for a geographical region where a user stays over a certain time interval [23], [24]. We use two scale parameters to extract the stay points: a distance threshold ϵ and a temporal threshold δ .

Let $L = \{(p_m, t_m), \dots, (p_{m+n}, t_{m+n})\}$ be a sub-sequence from m -th to $(m+n)$ -th in the mobility trajectory \mathcal{D}_u of the user u , where t is time, $p = (x, y)$ is a point location, $m \geq 0$ and $n > 1$. If $\forall m < i \leq m+n$, $\text{dist}(p_i, p_{i+1}) \leq \epsilon$ and $t_{m+n} - t_m \leq \delta$, the sequence L is viewed as a single stay point. In this study, we set $n = 6$ from empirical data exploration.

To examine users’ daily mobility behavior, we convert their mobility trajectories into sequences of stay points and select only individuals with at least one stay point per day.

Given the mobility trace \mathcal{D}_u of an individual u that consists of timestamped stay point centroids, we use the PlanetSense POI data for the semantic enrichment of the daytime period and the UrbanPop synthetic data for night time.

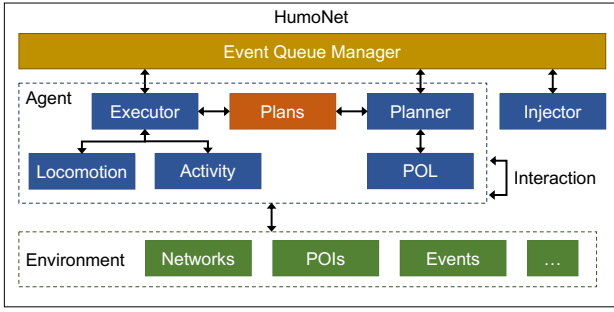


Fig. 4: Conceptual Diagram of HumoNet Architecture

- Inferring home locations: Following state-of-the-art practice, we extract $\mathcal{D}_{\text{night}}$ sequence, that contains only records occurring during the nighttime. Let p_i be the most frequent coordinates $\mathcal{D}_{\text{night}}$. Using the UrbanPop data, we find the nearest home location to p_i , and assign it to the user u as being a home location [25].
- Inferring the semantics of other locations: We extract $\mathcal{D}_{\text{day}} = \mathcal{D}_u \setminus \mathcal{D}_{\text{night}}$. For each (t_i, p_i) , we assign the closest POI from the PlanetSense data.

F. HomeAnoz: Anonymized home location assignment

Home locations are the critical pseudo-identifier for individuals, among other visits in mobility data. In case any trajectory data includes home locations, we apply anonymization of home locations preserving location, k -anonymity, and reciprocity [26]. First, we order all households' home locations identified by UrbanPop using Hilbert Order to meet the reciprocity condition of every user. Then, we assign home locations to corresponding buckets so that each bucket contains at least k home locations. For preserving locality based on network distance, we apply cell ordering using a hierarchical-graph [26]. Using these further anonymized home location, we randomly assign agent home locations within reasonable proximity. This privacy-enhancing technology (PET) is a pre-processing step, and we discuss an additional simulation-level privacy preserving technique in Section IV.

IV. HUMONET

This section describes the HumoNet architecture. Figure 4 illustrates the conceptual architecture of HumoNet, which consists of four main groups, the Agent logic, Query Manager, Injector, and Environment. In Agent logic, each agent has their POL, which describes their intended activity within a certain time frame. The Planner converts the POL into actionable items, i.e., Plans that consider the Environment and include constraints. Plans specifies what time the agent would do a certain activity and how to get to a certain location. Then Executor implements Plans by performing Activity and Locomotion. Activation of each component for each agent is considered an event. Query Manager is in charge of handling events and ensuring synchronization. Injector plays a role in inserting external events into a simulation. Intervention can be external events and can refer to an anomaly or a policy that does not originate from the targeting agent. We explain more details of key components in subsequent sections.

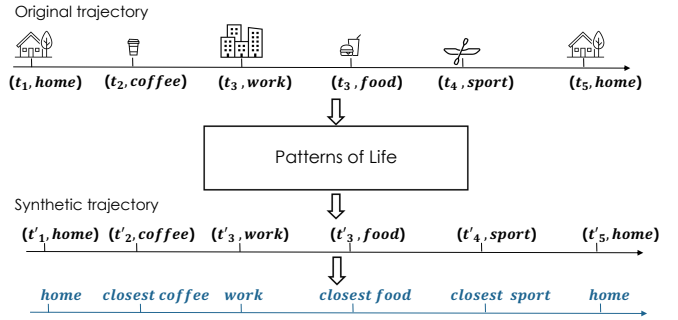


Fig. 5: Synthetic trajectories in POL.

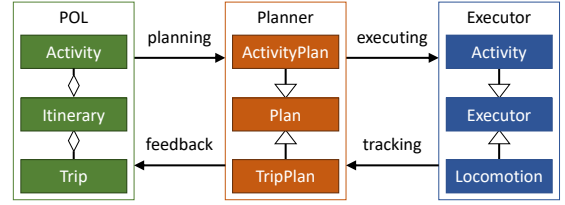


Fig. 6: Flow among itineraries, plans and executions

A. Patterns of Life (POL)

The proposed POL approach exploits semantic knowledge of POIs to ensure synthetic trajectories' veracity while preserving privacy. We analyze the semantic patterns of individuals' visits instead of their geographical coordinates. For instance, the trajectory of the individual in Figure 5, consists of home, coffee shop, work, fast-food, park, work, sport, restaurant, then home. In the first step, we extract some statistical properties that describe individuals' semantic preferences. We then construct a synthetic trajectory by referring to the nearest POI holding the candidate semantics.

For example, suppose a user u_1 is at a location l_1 having the semantic *restaurant*. According to the statistical characteristics observed within the real-world data, this user is more susceptible to going to a location holding the semantic *shop*. Hence, the semantic *shop* is inferred as the next type of place the user would go to. For the geographical selection, our POL module suggests the nearest location from l_1 holding the candidate semantic, i.e., *shop*. Thus, we do not use geolocation to track a user's location but instead reflect their visitation behavior for a sequence of semantic types while safeguarding their privacy.

B. Planner

Planner is in charge of creating executable plans from itineraries that consist of activities and trips (see Figure 6). Plans are information that describes 5W1H (i.e., who, what, when, where, why, and how) of actions, which is compatible with activity-based mobility intervention (ABMI) [2]. We explain why this is important in the following subsections. Plans are divided into two types: ActivityPlan and TripPlan. The planned activities and trips are performed by Executor.

We track execution and which plans are completed, being executed, or ongoing. Figure 7 shows an example of how plans are structured. Each agent has a current plan that consists of a series of ActivityPlan and TripPlan. TripPlan

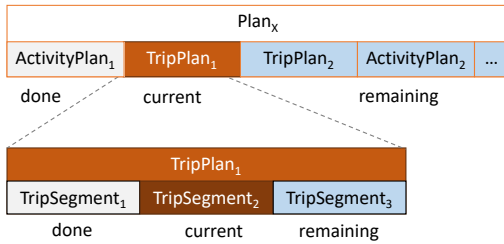


Fig. 7: Example of tracking plans and executions

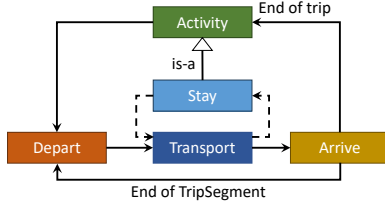


Fig. 8: Transport state transition diagram

is divided further into *TripSegments* to specify details, including routes and modes of transportation. *TripSegment* is a trip with a single mode of transportation or *leg*. For instance, if agent X travels from POI P_1 to POI P_2 , X may walk from P_1 to road R_1 , drive from R_1 to road R_2 , and then walk from R_2 to P_2 . In this journey, $\overrightarrow{P_1R_1}$, $\overrightarrow{R_1R_2}$, and $\overrightarrow{R_2P_2}$ are *TripSegments*. We note that any one activity may or may not entail a trip from one place to another, i.e., moving from one place to another. In Figure 7, *TripSegment*₂ of *TripPlan*₁ is being executed. Execution of plans can be input by POL to make a feedback loop.

C. Locomotion

Agents in HumoNet have the capability to move from one place to another, referred to as locomotion. HumoNet supports multi-modal locomotion. For vehicles, HumoNet employs SUMO [15]. In this study, we assume that each agent has a privately owned vehicle (POV), and their vehicle moves along the road network, which is part of *Environment*. Due to the lack of available sidewalk data, pedestrians are free-space walking from locations within the network to POIs, i.e., assuming no obstacles in their way. If an agent needs to move beyond the network (e.g., POIs in *Environment*), the agent will move by *Walking*.

Figure 8 shows a conceptual diagram of the transport state transitions within HumoNet. When a trip of one agent starts, a transport state is *Depart*. After *Depart*, the agent is moving in the *Transport* state. *Stay* is an intentional activity when the agent has no intention to move. For instance, if a car stops during movement, it is not considered as *Stay*. *Arrive* is a state in which the agent arrives at a destination, transitions from the *Transport* state, and stops moving.

D. Injector

This subsection describes how HumoNet can support *activity-based mobility intervention* (ABMI) from outside of simulations, which is an important modeling requirement for policy and decision making. Kim et al. [2] define ABMI as the action of intervening in the activities of a person or a group of

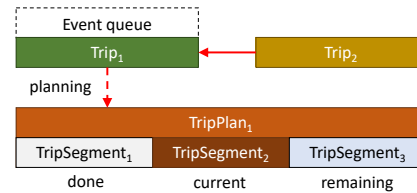


Fig. 9: Example of interruption by intervention

TABLE I: Handling conflicts of plan

Current trip	Interrupting trip	Interrupting trip	
		Normal	Controlled
Normal	Normal	Append	Replace/Insert
Controlled	Controlled	Ignore	Append

people such that their mobility is changed from their planned activity, and it is a type of intervention that may change the location of an agent by activity that needs to achieve certain goals.

Injector is a module that allows us to inject run-time interventions in experimental settings [27]. One of the challenges in run-time injection is how to handle a conflict between an existing plan and a newly inserted activity. Figure 9 illustrates a case of conflict where *Trip*₁ is an existing trip in an event queue. After planning, *Trip*₁ becomes *TripPlan*₁ consisting of *TripSegment*₁, *TripSegment*₂, and *TripSegment*₃. When a trip interruption is injected into an existing trip plan, the challenge is how to effectively handle the remaining segments, i.e., *TripSegment*₂ and *TripSegment*₃, and interrupting *Trip*₂. Unless specific instructions to handle such conflicts are given, it is up to the controlled agent's decision and is the running simulation's responsibility to perform reasonable actions.

To resolve conflicts across agents in a consistent manner, we use a decision table and explanation of each decision, as shown in Table I. There are two types of trips: *normal* and *controlled*. A normal trip is a planned trip based on our simulation model while a controlled trip is a trip inserted by external parties. A key idea is that controlled trips have a higher priority than normal. We define four ways to handle conflicts: *append*, *replace*, *insert*, and *ignore*. Figure 10 shows the summary of the explanation of each operation and example cases before and after handling. If the priorities of a current trip and interrupting trip are the same, then we *append* the interrupting trip to the current trip. One example for this is that agent X is supposed to depart for a gym after visiting cafe at 2pm, but due to traffic jam X has not arrived yet. In this case, the agent will leave the cafe as soon as they arrive at the gym since going to the gym is an interrupting trip. Although it may not always be realistic, it is reasonable because their priorities are the same.

Suppose a current trip is normal and an interrupting trip is controlled. In that case, we *replace* the interrupting trip with the existing trip or *insert* the interrupting trip into the current trip. The choice between using *replace* or *insert* is dependent on the length of the remainder of the planned activity that will take place at the original destination. Let us assume that agent Y is on the way to work at 12pm and will leave work at 5pm.

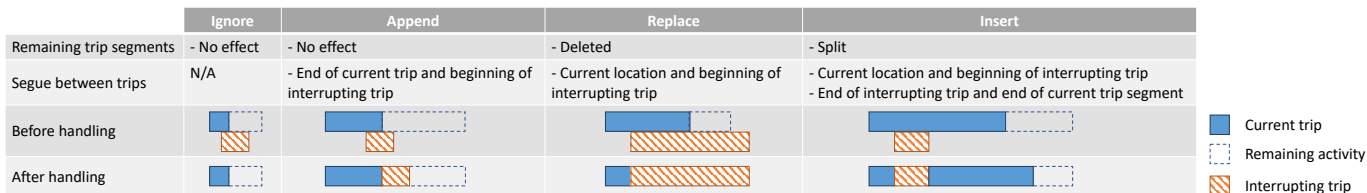


Fig. 10: Trip conflicts and strategy to resolve

If the interrupting trip is to stop by a cafe at 12:05pm, we want Y to keep the original destination in the plan because the time remains for the work activity. Thus, we split the existing trip segment into two parts and *insert* the interrupting trip between them. On the other hand, if the interrupting trip takes too long to perform the remaining activity, we *replace* the existing trip with the interrupting trip. We *ignore* an interrupting trip if a current trip is *controlled* and the interrupting trip is normal.

V. EXPERIMENTS AND RESULTS

To assess the realism of HumoNet’s mobility patterns as derived from POL, we conduct a set of experiments using mobility measures and divergence scoring. For the experiments HumoNet simulations were run on Google Cloud Compute Engine in which the machine type VM instance is n2-standard-64 with 64 CPU cores and 252GB RAM, and the Operating System (OS) is Ubuntu 20.04. The core modules of HumoNet are implemented in C++. The experiments test the following: (1) does HumoNet realistically characterize mobility patterns as measured by agent visits and exploration and (2) how similar are POLs to those found in real data. In the following section, we describe the simulation input data and reference datasets for this study. Then, we describe the metrics used for similarity testing in the evaluation and report the findings of our analyses.

A. Simulation input data

- **AOI:** We selected Knoxville Metropolitan Statistical Area (MSA), TN, USA, as area of interest (AOI).
- **Synthetic population:** Approximately 781,672 agents are listed under UrbanPop for 23,062 distinct households within the city of Knoxville.
- **The number of agents:** For this study, we use two different agent population datasets, with an $n = 20K$ and $n = 50K$.
- **Trajectories:** Commercial data consists of the trajectories of 170,470 users within the city of Knoxville. The data collection includes 12,368,934 records for a duration of 15 days within the month of January 2023. The frequency of sampling is of the order of a few seconds.
- **POIs:** We model 55,034 POIs within the given AOI. Figure 11 shows the distribution of each category.

Reference datasets are used to calculate probability distributions and are compared with HumoNet data to determine the degree of divergence between microsimulation data and a standard set of benchmarks. These datasets include real-world trajectory data and aggregate data from the literature:

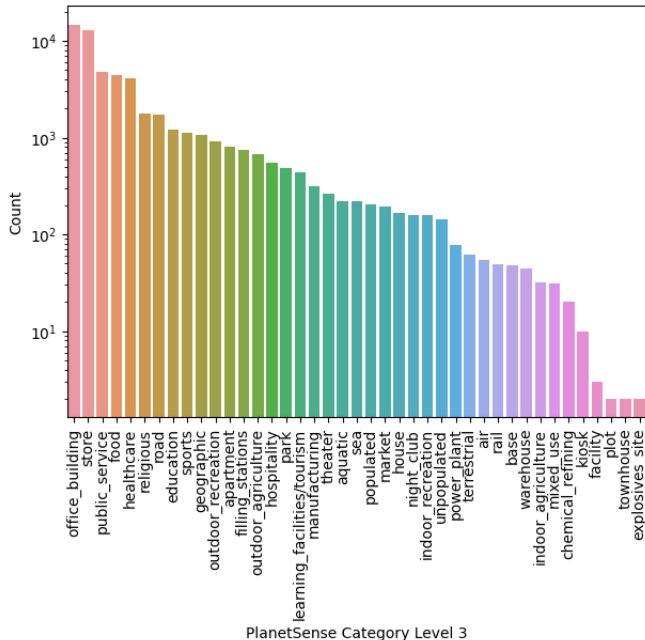


Fig. 11: Number of POIs for each category in Knoxville MSA

- **GeoLife** [28]: GPS trajectory dataset collected in (Microsoft Research Asia) GeoLife project by 182 users in a period of over three years.
- **Literature:** Data points statistically drawn from well-cited mobility science literature that include number of locations visited (Figure 2 of [29]), radius of gyration (Figure 2 of [30]), and level of exploration (Figure 1 of [13]).

B. Metrics

To evaluate our simulation, we include metrics that assess the realism of agent visit frequency, the distances agent’s travel, and the variation in type of POIs agents visit. At-scale computation of the metrics of interest is performed using the system solution in [31].

- **Number of locations visited (N_l)** [29]: This measure computes the number of unique locations visited by an individual per day. A unique location is a place that an actor visits at least once per day. If an actor stops at home three times in a day, the home will count as just one unique location for that day. As weekday patterns tend to be more consistent with each other compared to weekend days, only weekdays are tested.
- **Radius of gyration (R_g)** [30]: This measure quantifies the spatial extent of an individual’s mobility (i.e., the characteristic distance traveled by an individual). It is presented

TABLE II: JSD scores for each measure

Measure	Ref. dataset	Simulation dataset	
		20K	50K
N_l	GeoLife	0.2590	0.3824
	Lit	0.0215	0.1397
R_g	GeoLife	0.1429	0.0995
	Lit	0.0578	0.0657
L_e	GeoLife	0.0210	0.0155
	Lit	0.0038	0.0025

as a single distance value for an individual, which is then aggregated across the dataset population to generate a single frequency or probability distribution, which will then be compared against other frequency or probability distributions from GeoLife and literature datasets by means of a divergence score for evaluation purposes.

- **Level of exploration (L_e)** [13]: The concentration of an individual’s visits within their top- k visited locations. In other words, this metric examines how an individual’s visit counts are distributed among the places they go.

We rely on the Jensen-Shannon (JS) divergence method in probability theory to compare probability distributions for each metric. JS divergence is a symmetrized and smoothed version of the Kullback–Leibler (KL) divergence. Given two distribution P and Q , JS divergence (JSD) is defined as: $JSD(P||Q) = \frac{1}{2}KLD(P||M) + \frac{1}{2}KLD(Q||M)$, where KLD is KL divergence and $M = \frac{1}{2}(P + Q)$ is a mixture distribution of P and Q . This score is on the scale of $[0, 1]$, where base 2 is used for logarithm calculation in the JS scoring. When the score is 0 or very close to it (i.e., 0.01 or 0.005), it indicates the best case of closeness of matching between $P(x)$ and $Q(x)$. However, a score close to 1 (e.g., 0.8 or 0.9) indicates that distributions $P(x)$ and $Q(x)$ are not matching at all, i.e., they are very dissimilar. Table II shows JSD scores between simulations and references for each measure. In the following subsections, we elaborate on the results of the analysis of the scores with corresponding charts of distribution. For a short notation, let $D_{X,S}$ be the distribution of variable X of data S . For instance, $D_{N_l,GeoLife}$ is distribution of N_l of *GeoLife*. Further, $JSD(D_{R_g,Lit}||D_{R_g,20K})$ is a JSD score of R_g between datasets *Lit* and 20K, which is 0.0578 in Table II.

C. Visiting patterns

Figure 12 shows the frequency of people visiting unique locations over each 24-hour period. Common trends is as N_l increases, the frequency of N_l decreases except for $N_l = 1$. The reason why the frequency of $N_l = 1$ is lower than $N_l = 2$ is that one way trip, namely, people leave/come back their home within 24-hour period, which is counted as 2 visited locations. We note that $D_{N_l,GeoLife}$ has a long tail distribution when compared to others. As a result, $JSD(D_{N_l,GeoLife}||D_{N_l,20K})$ and $JSD(D_{N_l,GeoLife}||D_{N_l,50K})$ are 0.2590 and 0.3824, having a relatively large score. We argue that inference of locations visited impacts the distribution. *GeoLife* has a different sampling mechanism including frequency, and sampling can influence inference of locations visited. $D_{N_l,20K}$ is very similar to $D_{N_l,Lit}$ with $JSD = 0.0215$. We noticed that

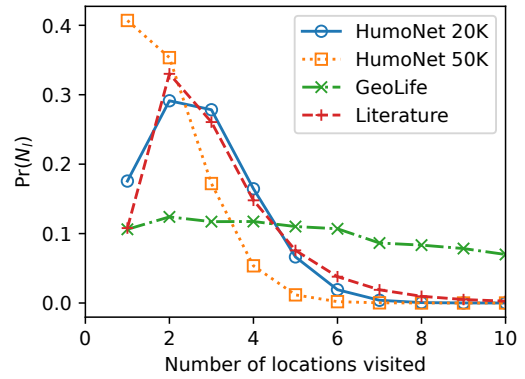


Fig. 12: Distribution of number of locations visited per day

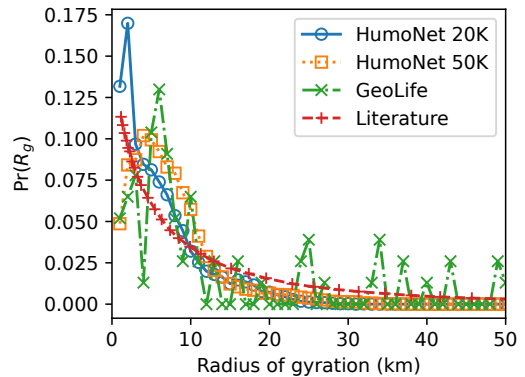


Fig. 13: Distribution of radius of gyration

$D_{N_l,50K}$ is dissimilar to others due to high frequency in $N_l = 1$. One explanation is that simulation 50K has heavier traffic congestion than simulation 20K and more agents in 50K arrive home late after midnight.

D. Exploration patterns

Figure 13 shows a comparison of distributions of the radius of gyration among four datasets. Due to the sample size in *GeoLife*, we observe wide fluctuations of $D_{R_g,GeoLife}$ while the curve of $D_{R_g,Lit}$ is smooth, which is sampled from truncated power law distribution. $JSD(D_{R_g,Lit}, D_{R_g,20K})$ and $JSD(D_{R_g,Lit}, D_{R_g,50K})$ are 0.0578 and 0.0657. We observe differences between HumoNet 20K and HumoNet 50K radius of gyration distributions that can be explained by the upscaling and improved POLs in the 50K version.

Figure 14 shows distributions of level of exploration. The x-axis is the rank- k locations for the entire population that are computed as follows: select the top- k most visited locations of an individual, calculate the total trips by rank, then aggregate the rank-by-rank total to the whole population. The y-axis of the figure is the probability of visits to the rank- k location for the entire population. As we can visually notice in Figure 14, $D_{L_e,20K}$ and $D_{L_e,50K}$ have heavy tail distribution similar to $D_{L_e,GeoLife}$ and $D_{L_e,Lit}$. JSD scores for L_e in Table II are the smallest compared to other measures across the datasets. Agents in HumoNet tend to visit the same places more often compared to reference datasets. Such a gap may be reduced if agents further diversify visit locations.

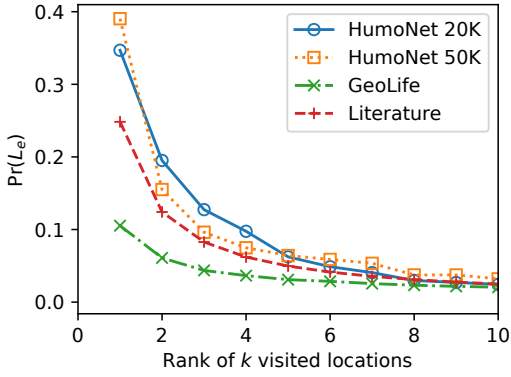


Fig. 14: Distribution of level of exploration

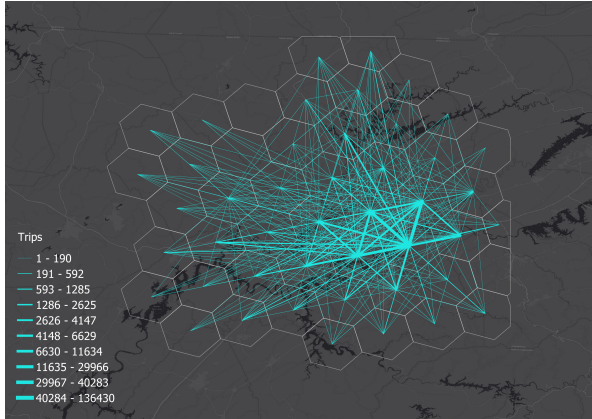


Fig. 15: Trips between H3 hexagons in Knoxville MSA

Figure 15 shows trips of HumoNet 20K between 48 cells of resolution 5 of Uber’s H3 grid (<https://h3geo.org>) intersecting with the Knoxville MSA. The average hexagon area and the average edge length are about 252 km² and 9.85 km, respectively. Thicker lines in the figure depict more trips between cells. Due to the high concentration of public POIs within Knox County and nearby counties, we observe many trips from/to Knox County. Accordingly, we can observe traffic congestion between high traffic demand zones.

VI. DISCUSSION

With our implementation of HumoNet, we found two factors had a significant impact on the performance of the simulations: (1) accuracy of trajectories and (2) running time.

A. Accuracy of simulation trajectories

Simulation accuracy in mobility models should be assessed along two dimensions: models and data. If either dimension is not accurate, we cannot expect realistic output. This subsection focuses on the challenges in data enhancement. As HumoNet takes human mobility into account, infrastructure related to transportation is as crucial as the passage of human movement. Collecting and maintaining transportation networks, including road networks and metro networks, is an ongoing challenge. Depending on the requirements of the model applications,

different levels of detail in network data need to be considered. We consider three aspects to prepare network data for HumoNet: network types, attributes, and granularity.

A network type is closely related to modes of transportation. Thus, it is common for applications to take a subset of an entire transportation network for their purpose. For instance, if an application considers only agents using mass transit, such as subways, the application does not need to use a public road network that trains do not run on. Another example is that a public transportation application might not include private service roads to avoid unnecessary computation. We noticed discrepancies between simulation trajectory and real world movement. In the real world, POVs should be able to access service roads where residents enter their property or community. Public drive roads are open to everyone, while service roads are accessed by a community. These differences lead to unrealistic patterns in the last travel segment before an agent reaches its POI. Thus, we need to consider configurations in selecting data. A challenge in this task is that the crowd-sourced data are noisy and have missing or incorrect values, hence we need a proper way to impute and correct the data.

Another critical problem we faced is related to network connectivity. In routing from one point to another on a network, it is critical if networks are disconnected or weakly connected because we cannot create a path from one to another. Although a network becomes disconnected is plausible in certain circumstances, such as islands or temporarily disconnected networks due to disruptive events (e.g., flooding, earthquake), in our case, the problem was caused by other reasons. One reason is due to how we select networks spatially. Unless selecting all nodes and edges in the entire area, we need to select a sub-graph of the network within AOI. This selection or clipping of a network can cause two types of cases: (1) clipping cuts one-way roads; (2) clipping cuts nodes and edges outside of AOI that are connected to a sub-graph within AOI.

Another reason for network connectivity issues is the way we select network semantics. One-way roads are connected to a parking lot that is not part of the selected network. If a parking lot has its own navigable network, we consider adding the network of the parking lot. However, in some cases, parking lots are regarded as a space rather than a network feature. Noisy data can contribute to the network connectivity problem too. For instance, construction on roads is common, and updating a network should occur. If the network is partially updated, one-way roads can be a dead-end.

B. Simulation time

Some aspects that influence simulation performance exist when we scale HumoNet. The interface between computation nodes (i.e., SUMO and HumoNet core logic) is one of them. On the one hand, TraCI is flexible as a general-purpose interface that SUMO provides, but we noticed that TraCI could be a bottleneck. On the other hand, libsumo is a C++ interface, so we can directly access classes and functions from

our simulation code without packing/unpacking. However, it has limitations to run and access multiple SUMO instances.

Another burden is computation time for routing. Computation time for routing is superlinear to network size and proportional to the number of trips. To reduce repeated computation, we calculate and store routes before running the simulation. By indexing the routes on running simulations, we could reduce trip computation dramatically. Origin-destination (OD) information in POL is used to generate pre-computed routes and index from coordinates to locations on the road network. In this study, we capture and store the locations of every agent with 1HZ: (*id, timestamp, x, y*). Such high-fidelity trajectory data can be used for many applications. However, I/O could be a bottleneck, and the large size of output files is also challenging because it may not fit in local storage.

VII. CONCLUSION

Human mobility contributes to many disciplines, including transportation, energy, urban planning, market analysis, disaster management, and security. In this paper, we proposed HumoNet framework that provides capabilities to simulate (1) integrated patterns of life of individual agents, (2) multi-contextual models at the point of interest (POI) level, (3) interactions among agents, and (4) realistic human mobility patterns. We envision HumoNet will provide explainable and data-driven simulated data that captures emerging complex behaviors from diverse patterns and can be used as a benchmark for validating AI/ML tasks. (e.g. location prediction)

Although we make use of real-world telematics data, we plan to make significant data enhancements, first addressing important gaps around sidewalk data. We propose using privacy-preserving semantic-aware POL trajectory generation techniques invented at ORNL. However, a more advanced, adaptive privacy-preserving mechanism will be developed in future versions that will enable tuning the level of privacy according to requirements for location accuracy preservation. Finally, we would like to develop specific adaptations of the HumoNet simulation for various settings – such as catastrophic modeling, Real-Time or Hardware-in-the-Loop (HIL), scenario modeling, and capable of running on Exascale computing architecture.

REFERENCES

- [1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail *et al.*, “Human mobility: Models and applications,” *Physics Reports*, vol. 734, pp. 1–74, 2018.
- [2] J.-S. Kim, G. M. Thakur, and S. C. Christopher, “A design of activity-based mobility intervention,” in *SSTD’23*, 2023, pp. 131–140.
- [3] B. G. Silverman, G. Bharathy, and N. Weyer, “What is a good pattern of life model? guidance for simulations,” *Simulation*, vol. 95, 2019.
- [4] J.-S. Kim, H. Kavak, C. O. Rouly, H. Jin, A. Crooks, D. Pfoser *et al.*, “Location-based social simulation for prescriptive analytics of disease spread,” *SIGSPATIAL Special*, vol. 12, no. 1, pp. 53–61, 2020.
- [5] A. Maslow, “A theory of human motivation,” *Psychological Review*, vol. 50, no. 4, p. 370, 1943.
- [6] H. Amiri, S. Ruan, J.-S. Kim, H. Jin, H. Kavak, A. Crooks *et al.*, “Massive trajectory data based on patterns of life,” in *ACM SIGSPATIAL GIS’23*, 2023, pp. 1–4.
- [7] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson *et al.*, “Data-driven mobility models for covid-19 simulation,” in *ACM SIGSPATIAL ARIC’22*, 2020, pp. 29–38.
- [8] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, “A survey on deep learning for human mobility,” *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–44, 2021.
- [9] “Open pflow: Creation and evaluation of an open dataset for typical people mass movement in urban areas,” *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 249–267, 2017.
- [10] Y.-A. De Montjoye, S. Gams, V. Blondel, G. Canright, N. De Cordes, S. Deletaille *et al.*, “On the privacy-conscious use of mobile phone data,” *Scientific data*, vol. 5, no. 1, pp. 1–6, 2018.
- [11] M. von Mörner, “Application of call detail records-chances and obstacles,” *Transportation research procedia*, vol. 25, pp. 2233–2241, 2017.
- [12] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González, “The timegeo modeling framework for urban mobility without travel surveys,” *PNAS*, vol. 113, no. 37, pp. E5370–E5378, 2016.
- [13] C. Song, T. Koren, P. Wang, and A.-L. Barabási, “Modelling the scaling properties of human mobility,” *Nature physics*, vol. 6, no. 10, 2010.
- [14] E. Letouzé, P. Vinck, and L. Kammourieh, “The law, politics and ethics of cell phone data analytics,” *Data-Pop Alliance*, 2015.
- [15] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich *et al.*, “Microscopic traffic simulation using sumo,” in *IEEE ITSC’18*, 2018.
- [16] J. Tuccillo, R. Stewart, A. Rose, N. Trombley, J. Moehl, N. Nagle *et al.*, “UrbanPop: A spatial microsimulation framework for exploring demographic influences on human dynamics,” *Applied Geography*, vol. 151, p. 102844, 2023.
- [17] N. N. Nagle, B. P. Battenfield, S. Leyk, and S. Spielman, “Dasymetric modeling and uncertainty,” *Annals of the Association of American Geographers*, vol. 104, no. 1, pp. 80–95, 2014.
- [18] J. V. Tuccillo, “Downscaling Synthetic Populations to Realistic Residential Locations,” *Proceedings of the United States Research Software Engineer Association Conference 2023*, 2024.
- [19] R. Stewart, M. Urban, S. Duchscherer, J. Kaufman, A. Morton, G. Thakur *et al.*, “A bayesian machine learning model for estimating building occupancy from open source data,” *Natural Hazards*, vol. 81, pp. 1929–1956, 2016.
- [20] G. S. Thakur, B. L. Bhaduri, J. O. Piburn, K. M. Sims, R. N. Stewart, and M. L. Urban, “Planetsense: A real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data,” in *ACM SIGSPATIAL GIS’15*, 2015.
- [21] J. Fan, J. Bentley, and G. M. Thakur, “Sonet++: A knowledge graph of geographic categories based on osm tag representation,” 9 2023.
- [22] R. Palumbo, L. Thompson, and G. Thakur, “Sonet: a semantic ontological network graph for managing points of interest data heterogeneity,” in *ACM SIGSPATIAL GeoHumanities ’22*, 2019.
- [23] A. Cuttone, S. Lehmann, and M. C. González, “Understanding predictability and exploration in human mobility,” *EPJ Data Science*, 2018.
- [24] L. Amichi, A. C. Viana, M. Crovella, and A. A. Loureiro, “Understanding individuals’ proclivity for novelty seeking,” in *ACM SIGSPATIAL GIS’20*, ser. SIGSPATIAL ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 314–324.
- [25] L. Amichi, A. C. Viana, M. Crovella, and A. Loureiro, “From movement purpose to perceptive spatial mobility prediction,” in *ACM SIGSPATIAL GIS’21*, 2021, p. 500–511.
- [26] J.-S. Kim and K.-J. Li, “Location k-anonymity in indoor spaces,” *Geoinformatica*, vol. 20, no. 3, pp. 415–451, 2016.
- [27] J.-S. Kim, H. Kavak, U. Manzoor, and A. Züfle, “Advancing simulation experimentation capabilities with runtime interventions,” in *2019 Spring Simulation Conference (SpringSim)*. IEEE, 2019, pp. 1–11.
- [28] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, *Geolife GPS trajectory dataset - User Guide*, July 2011. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/>
- [29] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, “Unravelling daily human mobility motifs,” *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.
- [30] L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti, “Understanding the patterns of car travel,” *The European Physical Journal Special Topics*, vol. 215, pp. 61–73, 2013.
- [31] D. De, G. Malviya Thakur, J. McGaha, C. Brown, X. Nie, T. Thomas *et al.*, “Dicer: Data intensive computing environment and runtime for evaluating unprecedented scale of geospatial-temporal human mobility data,” in *IEEE MDM’24*, 2024.