Generating Semantically Enriched Mobility Data from Travel Diaries

Maja Schneider*[1][0000−0001−5936−1415], Charini Nanayakkara[2][0000−0002−7603−1845], Matthias Mohn[1], Peter Christen[2][0000−0003−3435−2015], and Erhard Rahm[1][0000−0002−2665−1114]

[1] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
[2] School of Computing, The Australian National University, Canberra, Australia

Abstract. Human mobility data is valuable for many applications, but it can pose a significant privacy risk for individuals. A person's daily movements are closely linked to their sociodemographic characteristics and their points of interest (POI), which can reveal sensitive information about them, such as their religion or educational background. Researching and mitigating these risks requires realistic, semantically rich data, which is often unavailable or lacks semantic features. We introduce ASTRA, an agenda-based approach for generating synthetic mobility data with semantic attributes. Using real travel surveys and census data, ASTRA simulates artificial agents with sociodemographic features that follow daily activity agendas. It maps activities to semantically similar POIs, and projects them onto real-world locations within a user-defined geographical region. To model movement, ASTRA extends the exploration and preferential return (EPR) model with spatio-temporal and semantic constraints as imposed by an agenda. Our evaluation against real check-in data and the EPR baseline model shows that ASTRA can generate realistic mobility data at scale, preserving important characteristics of human movement.

Keywords: Synthetic Data · Human Mobility · Points of Interest · Agenda-Based Modeling

1 Introduction

Driven by the extensive use of location-based services on mobile devices, location data of app users is becoming increasingly relevant. Such data can give valuable insights into human mobility patterns and enable a range of applications, such as urban planning or personal recommendations [26]. However, the collection of their location data bears potentially serious privacy risks for an individual [23]. Human mobility is highly correlated with an individual's sociodemographic characteristics and their *agenda*, which is the schedule of their daily activities [14,22]. These attributes determine a person's points of interest

* Corresponding author. Email: mschneider@informatik.uni-leipzig.de

(POI) which are the places they stay at or visit, for example their home, work, or places of leisure activities. POIs are especially privacy-critical because they can reveal sensitive information about a person, such as their educational background, age or gender [30]. To understand the involved privacy risk and mitigate such risk with location privacy-preserving methods, researchers require suitable data to study questions such as: *Does an individual's geographical trajectory reveal their demographic attributes or even allow to re-identify them? Is there an increased privacy risk for certain demographic groups?*

When data is collected for research purposes, it is usually limited to small regions or certain groups of people, such as students, and can therefore be biased [2,17,27]. A solution to the lack of available data is to generate synthetic mobility data [13]. However, existing methods usually do not include semantic features, such as the location and label of a POI or the sociodemographic details of a person. In addition, real and synthetic mobility data is often modified to protect private user attributes, making such data unsuitable for privacy research. To be useful for privacy research, synthetic human mobility data needs to fulfill certain requirements: The data needs to be *realistic* to reflect actual human movement behavior. It needs to be *representative* for a society, for example a city population, in order to avoid bias and allow for generalizable and fair analyses. To understand privacy implications, it needs to be *explainable* with regard to the privacy-sensitive factors that influence mobility, such as the sociodemographic attributes of a person and the purpose of their movement as represented by their POIs. Finally, to understand privacy concerns in different geographical contexts, the synthesis of mobility data must be *flexible* with respect to the geographical simulation region.

Existing approaches to synthesizing human mobility data generally do not meet all of the above criteria, and in particular lack explainability [13]. To fill this gap, we present ASTRA, our approach to generate **A**genda-based **S**emantic **TRA**jectories. Agenda-based modeling is based on the assumption that a person's trajectory is not random but highly influenced by the activities they pursue and their social characteristics [12,29]. These factors determine which POIs a person visits, while travel can be seen as a by-product of these visits. To model the different sociodemographic groups in a society, ASTRA samples a population of artificial mobile agents with certain sociodemographic features and a daily *agenda* from a data set of real travel diaries. We define an agenda to be a sequence of activities that can be divided into alternating segments of consecutive travel activities or non-travel activities. A non-travel segment refers to an agent's POI and all of its activities are assumed to be carried out at the same location. For example, a non-travel segment with the activities *sleeping*, *eating*, and *reading* can be linked to the label of a POI where these activities can be executed, such as an *apartment*, *hotel*, or *tent*. Travel segments connect two non-travel segments and indicate the main mode of transportation (like *walking* or *driving*) that the agent uses to move between the respective locations.

ASTRA uses an embedding approach [21] to map the activities of non-travel segments to semantically similar POI labels. The segments are then projected

Fig. 1. An agent's agenda is mapped by ASTRA to a sequence of POI labels using a semantic similarity score. We use Cosine similarity, ranging from –1 (low similarity) to 1 (high similarity). The POI label sequence is further processed into a POI trajectory by selecting the GPS coordinates of actual POIs on a geographical map.

to the GPS coordinates of actual POIs with the respective label on a real map, creating a *POI trajectory*. To select a POI, ASTRA uses an extension of the exploration and preferential return model (EPR) [24] that takes into account spatio-temporal travel constraints and individual and collective semantic constraints, imposed by the agent's agenda and the map environment. Fig. 1 illustrates the mapping from an agent's agenda to a POI trajectory. The example highlights how certain POIs can potentially disclose private attributes, such as the educational background. Note that while the aim of ASTRA is to synthesize realistic trajectory data to allow researchers to study privacy, its intention is not to reconstruct an actual person's trajectory or to re-identify them.

Our contributions are as follows: (1) We extend the EPR model [24] by three spatio-temporal and semantic constraints to reflect that human mobility is driven by people's daily activities. (2) We present ASTRA, a new approach for synthesizing human mobility data that includes not only GPS locations, but also semantic features, such as the labels of POIs and the sociodemographic features of the data producers. The POI trajectories generated with ASTRA are realistic, representative, explainable, and flexible. (3) We evaluate ASTRA against a real data set of POI trajectories and against the baseline EPR model. Our results show that ASTRA can generate mobility data that is realistic with respect to spatial, temporal, and semantic measures of mobility.

2 Background and Related Work

Traditional mobility models focus on reproducing spatial and temporal aspects of human mobility. Many approaches create synthetic trajectories by sampling cer-

tain features of a trajectory, such as start and end locations, speed, or trip length, from distributions that are extracted from real mobility data sets [11,15,18]. Agenda-based models in turn rely on travel survey data to extract semantic mobility patterns and demographics, assigning them to artificial agents that move in a predefined geographical area in a rule-based manner [1,25,29]. Transitions between consecutive trajectory points are often sampled based on Markov models [3,20] or random walk models [6,7].

To overcome the limitations in existing random walk methods to not sufficiently reproduce certain aspects of human mobility, the exploration and preferential return model (EPR) [24] introduced two fundamental mechanisms to describe human mobility. The first, *exploration*, describes that individuals visit new locations during their day but the likelihood to do so decreases over time. The second mechanism, *preferential return*, reflects the observation that individuals are more likely to return to previously visited locations where the likelihood to choose a location is proportional to its visitation frequency.

Several extensions have been proposed to account for inaccuracies in EPR modeling. Empirical observations indicate that the likelihood to choose a location as a return does not only depend on its visitation frequency. Instead, individuals tend to return to recently visited locations, even if these have not been visited frequently. The concept of *recency* [4] thus introduces a rank-based approach that considers for each potential return location a frequency rank and a recency rank that are considered with a certain likelihood when choosing a return location.

The original EPR model also does not account for the observation that human mobility is influenced not only by individual motivations but by collective behavior, without which some mobility metrics cannot be properly modeled. The density-EPR [19] model incorporates such collective forces during the exploration phase by introducing the *gravity model*. Instead of following an equal distribution, a destination location is selected using a gravity score. This score is calculated based on the relevance of start and destination locations and their geographical distance from each other. The relevance of a location can, for example, be measured by its population density or another measure of popularity.

The density-EPR approach however ignores the semantic relevance of a location in the context of a typical activity schedule of a person with certain sociodemographic characteristics. Therefore, ASTRA incorporates such knowledge by combining density-EPR with an agenda-based modeling approach.

3 Agenda-Based Semantic Trajectory Generation

We now describe our approach for generating agenda-based semantic trajectories, ASTRA. Our goal is to generate realistic and representative POI trajectories for a population of artificial agents in an arbitrary simulation region. A POI trajectory consists of the GPS locations and labels of the POIs that an agent visits throughout a day. A key aspect of ASTRA is that it associates agents and their POI trajectories with corresponding sociodemographic features. We first

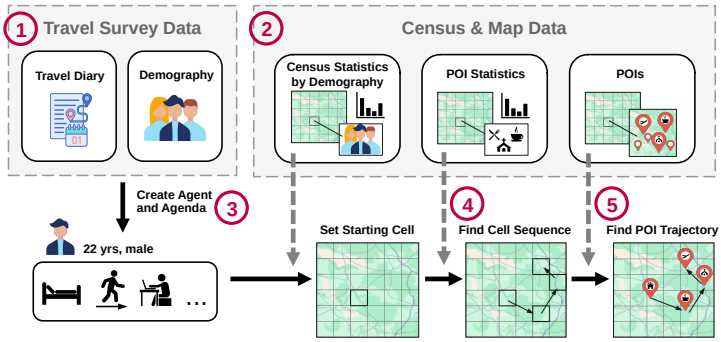


Fig. 2. The data processing pipeline of ASTRA

give an overview of ASTRA before we discuss its individual steps in the following subsections.

Fig. 2 outlines the data flow in ASTRA. (1) We use daily travel diaries from a travel survey data set, which includes sociodemographic attributes. (2) We then define the geographical simulation region, which we partition with a spatial tessellation into cells to reduce complexity and obtain census and POI data for each cell. (3) We create a user-defined population of artificial agents and assign each agent a starting cell, age and gender, proportionally to the actual population distribution in the cell. Each agent is randomly assigned a real travel diary and sociodemographic attributes from a person with matching age and gender. Diaries are preprocessed into agendas consisting of travel segments and non-travel segments. (4) ASTRA maps non-travel segments to cells according to the labels of their available POIs, while taking into account the spatio-temporal travel restrictions of travel segments. (5) Each segment is then assigned a precise POI within its cell to create a POI trajectory. Fig. 3 shows how a travel diary is translated into a POI trajectory with ASTRA.

3.1 Semantic Extension of EPR

The original EPR model [24] and its extensions do not capture the semantic context of movement. To make synthetic trajectory data more explainable in this regard, we propose an agenda-based semantic extension of density-EPR [19]. This means that we take into account both the spatio-temporal constraints of an agenda's travel segments, and the semantic constraints as imposed by an agenda's non-travel segments and the environment, in which the agent moves.

Spatio-temporal Travel Constraints: In density-EPR, the target location of exploration is chosen based on the gravity model (see Sect. 2). This allows an agent to pick any destination from a set of candidate locations with a certain probability. This approach, however, is agnostic of the actual travel constraints given by an agenda. In ASTRA, we restrict an agent's movement in the exploration phase to those locations that are reachable in the travel time of

Input Data: Travel Diary

Start	End	Activity	Travel Mode
00:00	08:00	sleeping	-
08:00	08:30	eating	-
08:30	09:30	reading	-
09:30	10:00	travel	walk
10:00	13:30	working	-
13:30	13:45	travel	walk
13:45	20:00	education	-
20:00	20:10	travel	drive
20:10	21:00	eating	-
21:00	23:55	sleeping	-

Output Data: POI Trajectory

Start	End	Lat	Lon	POI Label	POI Name	Travel Mode to POI
00:00	09:30	40,743	-73,985	apartment	Prism at Park Avenue South Apartments	-
10:00	13:30	40,746	-73,992	office	Micro Office Solutions	walk
13:45	20:00	40,741	-73,987	university	Princely International University	walk
20:10	23:55	40,737	-73,991	apartment	The Decker Building	drive

Fig. 3. Example of how a travel diary is converted by ASTRA into a POI trajectory

the current travel segment considering the specified travel mode. A relaxation of this constraint is allowed if there are no locations reachable that the agent has not yet visited. Note that in the return mode these constraints are not imposed and the agent can pick any of their previously visited locations.

Individual Semantic Constraints: To model the semantic dimension of mobility, we incorporate an agent's daily activities into density-EPR. To do so, we map each of the agenda's non-travel segments to a suitable POI label and restrict the agent's movement to locations that have this label. To find a suitable mapping, potential POI label candidates and the agenda segment's activities are embedded based on a textual description using a sentence transformer approach (SBERT) [21] and compared with a *semantic similarity score* S_{sem}.

Collective Semantic Constraints: The gravity score S_{grv} in density-EPR uses the relevance of a location to describe the influence of a collective force on an agent's movement. We measure the relevance of a cell in the tessellation by its population count, where we additionally consider the semantic relevance of a cell by introducing a *POI label frequency score* S_{frq}. This score aims to favour POIs that are not geographically isolated with regard to their label. For example, if an agent wants to visit a POI with the label *restaurant*, the likelihood to pick a cell that lies in a city's food district where many restaurants are located should be higher than the likelihood to pick a cell that contains only one restaurant.

Scoring of Locations: While the spatio-temporal constraints decide which locations are allowed to be selected by an agent, the final choice of a location is done in proportion to a weighted combination $S = \alpha S_{sem} + \beta S_{frq} + \gamma S_{grv}$, where $\alpha + \beta + \gamma = 1$.

3.2 Data Sources

We now describe the input data required for ASTRA, consisting of travel survey data, census data, and map data. The data sets discussed below are examples that can be replaced by other data sets of the same type.

Travel Survey Data: The basis of ASTRA are mobility diaries of real persons together with a set of their demographic features. Such data can be

obtained, for example, from travel survey data sets. In our work we use the Multinational Time Use Study (MTUS) data set [8,9], which is collected at random households in over 20 countries. For each survey respondent, this data set contains sociodemographic features (such as age, gender, ethnicity, or citizen status), their educational and employment background (including employment status, occupation, education level, and income), and details about their social life (like marital status, number of children, or care taking obligations). The respondents' characteristics also include information about their household, such as household size and income, or car or house ownership. As shown in Fig. 3, respondents also provide detailed one-day travel diaries, recording each activity of their day along with its start time, duration, and the travel mode used.

Census and Map Data: To allow for data abstraction, the simulation region is modeled with a spatial *tessellation* that partitions the region into non-overlapping subregions. This can be implemented, for example, using a grid, dividing the region into equally sized square cells. The cell width w of the grid is a user parameter and should be chosen to find a good balance between computational complexity and modeling accuracy. For each cell in the tessellation, census and map data is obtained.

- *Sociodemographic Features*: To create a realistic spatial distribution of the start locations of agents, and to measure the relevance of a cell as a start location for an agent, distributions of sociodemographic features are required. We determine each cell's population numbers by *age* and *gender* groups based on census information obtained from WorldPop[3].
- *Points of Interest*: To allow ASTRA to map activities to real POIs in the simulation area, the POIs of each cell along with their labels are queried from OpenStreetMap (OSM) using OSMnx [5]. We consider only POI labels that are a subcategory of one of the OSM tags *amenity*, *building*, *office*, *shop*, *tourism*, *leisure*, and *sport*, to account for places where people usually spend their time. In each cell, POIs are grouped by their label and counted. This information will serve as the basis for calculating the POI label frequency score S_{frq} as one of the three scores required for the semantic mapping as we described in Sect. 3.1.
- *Travel Times*: For a certain selection of travel modes, such as *car*, *bike* or *walk*, the travel times between pairwise cells are calculated. The calculation is based on the OSM street network which is represented as a directed graph between OSM nodes. Nodes are identified via their respective OSM identifiers and connected via edges, that are weighted with travel speeds for the respective travel modes. Travel times are calculated based on the shortest path between the centroids of the cells.

3.3 Generating an Agent Population and Their Agendas

After the input data for ASTRA is prepared, a user-defined number of artificial agents n is created and distributed onto the cells of the tessellation in proportion

[3] www.worldpop.org

to the actual population size. Each agent is randomly assigned an age and gender following the actual distribution of these features in their cell's geographic area. An agent is further assigned a travel diary and sociodemographic features of a person from the travel diary data set. For this, a person is chosen randomly from all persons with matching gender and (approximate) age. The travel diary is converted into an agenda by applying a non-overlapping sliding window. For each window the longest activity is extracted. Consecutive windows are grouped into travel and non-travel segments according to the nature of their activities.

3.4 Creating POI Trajectories

The goal of the next step is to project each agent's agenda onto the simulation region. ASTRA uses density-EPR with the recency extension described in Sect. 2 and applies the spatio-temporal and semantic extension introduced in Sect. 3.1. This means that in the exploration phase we use the gravity model together with our semantic extension, while in the return phase with a certain likelihood the recency-based approach is chosen.

During exploration, an agenda segment is first mapped to a cell in the tessellation before a specific POI in the chosen cell is derived later on. Candidate cells are obtained according to the spatio-temporal travel time constraints. For each candidate cell, S_{sem} and S_{frq} are calculated. Furthermore, S_{grv} is calculated based on the travel distance between the centroids of the previous and the candidate cell, and by measuring the relevance of a cell by its population size. All three scores are normalized and combined into a single score S by weighting the three components according to a user-specific weight assignment (see Sect. 3.1). A POI label and cell are then chosen randomly from the top-k highest scoring cells, where $k \geq 0$ is a user-specific parameter.

To pick a specific POI from the chosen cell, only POIs with a matching label are considered. These POI candidates are evaluated based on a single score. We again use the gravity score as we described in Sect. 2, but calculate it in a different way than before. The calculation of gravity is based on the distance between the centroid of the previous cell's POIs and the POI candidate, and the relevance of a POI candidate is measured by the spatial density of POIs in its surroundings. Following this approach, POIs that are in proximity to other POIs are preferred over those that are spatially isolated. This reflects the increased popularity of POIs in lively areas, such as a store in a shopping center, or a bar in a nightlife district. To realize this calculation, each cell is further divided into subcells using a grid. The relative number of POIs in each subcell in relation to the entire cell then indicates the relevance of the POIs in a subcell.

During the return phase, a POI must be selected that was previously visited by the agent. In order to ensure the greatest possible semantic integrity with the agenda, we ignore the spatio-temporal travel restrictions in this case. ASTRA evaluates each of the POI candidates according to their frequency and recency scores. In addition, ASTRA allows a return to a POI only if its semantic similarity with the segment is greater or equal to a user-defined threshold: $S_{sem} \geq \tau_{sem}$. Otherwise, the return is rejected and an exploration is executed instead.

4 Evaluation

We now evaluate the modeling capabilities of ASTRA[4]. We first assess the semantic mapping quality from agenda segments to POI labels. Then, we compare the spatial, temporal, and semantic characteristics of the generated POI trajectories of ASTRA with a real mobility data set and the traditional density-EPR model, instantiated by the DITRAS framework [20] using default parameters.

4.1 Experimental Setup

We first describe the baseline data set and configurations of ASTRA as used in our experiments.

Baseline Data Set: We compare ASTRA with the Foursquare data set [26], a real mobility data set collected in New York City, USA, between April 2012 and February 2013. It contains 227,428 check-ins (POIs) from 1,083 users who logged their visits to a POI with an app, indicating the POI's label, GPS location, and timestamp. Due to the nature of the app usage, the check-in history of a user can be incomplete and reflect only a part of their daily POIs [28]. Also, because such apps are more likely to be used by young adolescents to connect with their friends, this data set is likely biased towards this demographic group. Even though data sets collected via social media apps usually have these limitations, only they provide the POI information required for comparison with ASTRA.

Simulation Parameters: The simulation region of our experiments consists of the convex hull of all users' POIs in the Foursquare data set. We choose a tessellation cell width of $w = 1{,}500\,\text{m}$. With ASTRA, we create a population of $n = 10{,}000$ agents along with their POI trajectories. We create the same number of agents and trajectories using DITRAS [20].

Travel Diary Data Basis: We run ASTRA based on an extract of 190,088 MTUS [8,9] travel diaries from the USA between January 2003 and December 2019. This is to reflect travel behavior that is not influenced by the Covid-19 pandemic. The selected data is used by ASTRA as the basis for assigning a real diary to an artificial agent. The travel diaries are divided into agendas with a sliding window. The window size should be chosen appropriately to not miss any travel activities in the preprocessing, but at the same to keep computation time to a minimum. As more than 90 % of non-travel activities in the MTUS extract have a duration of at least 5 min, we choose this value as the window size. The data set contains travel modes *car*, *bike*, and *walk*, for each of which the travel times between all cells in the simulation area are calculated.

Scoring: Because the level of the semantic similarity score S_{sem} influences the number of POI candidates that an agent can choose from, this might impact the quality of the generated POI trajectories. We therefore test the three semantic similarity score weights of $\alpha = 0.9$, 0.6, and 0.3 and redistribute the remaining weight equally across the two other score components, gravity and POI

[4] ASTRA is available open-source at https://github.com/majaschneider/ASTRA.

label frequency score: $\beta = \gamma = \frac{1-\alpha}{2}$. After the combined score S is calculated, we select a POI label and its cell from the top-$k = 5$ highest ranking scores to introduce variability in the data generation process. In the return phase, we test the three minimum similarity score thresholds of $\tau_{sem} = 0.0$, 0.3, and 0.5.

4.2 Evaluation of the Semantic Mapping

The aim of the semantic mapping is to find the most similar POI label for an agenda segment and its activities. Using MTUS and OSM, this involves the mapping of 60 different non-travel activities to 665 unique POI labels.

To calculate the similarity of agenda segments and POI labels, we first embed their textual representation with the sentence transformer SBERT[5]. This model was specifically designed for semantic text comparisons and has been trained on 215 million (question, answer)-pairs. It converts texts into a 384-dimensional vector space and allows semantic nearest neighbor search based on Cosine similarity, which we use as the semantic similarity score.

POIs and activities usually only have a short identifier, or label. Due to SBERT's training data, it can be advantageous to provide more context and convert both identifiers and labels, into a longer textual description so that they resemble a question and an answer. We empirically test which of the following two options achieves better semantic mapping quality: *Label*: Only the identifiers and labels of activities and POIs are used for mapping, or *Desc*: A more comprehensive description of activities and POIs is used for mapping.

To realize the option *Desc*, we use descriptions of POI labels from OSM to obtain an answer-like text. For example, the POI label "Food Court" is converted into "An area with several different restaurant counters and a common dining area. Often found in shopping centers, airports, etc.". Similarly, we manually phrase activity descriptions as questions, in some cases indicating a location. For example, the activity "Food preparation, cooking" is converted to "What is a usual place for personal cooking and food preparation? (Often at apartment, detached house or building)". Optionally, a large language model can be used to help automate this task, with a subsequent manual verification.

In the case that an agenda segment contains multiple activities, several strategies are possible to map it to a POI, varying the impact of certain activities onto the mapping based on their significance. For example, a segment consisting of three activities 6 h *sleeping*, 15 min *eating*, and 15 min *reading*, can probably be linked to a home location due to the long duration of the sleeping activity. If the duration is ignored, the segment could potentially also be mapped to a restaurant or café. We therefore study different strategies: *Avg*: All activities of an agenda segment are mapped and the scores are averaged for the same POI. *Dur*: All activities are mapped and all single scores are averaged using a weighting according to their duration. *DurSqrd*: This is the same strategy as *Dur*, but the weights are squared in order to test an increased impact of the duration. *Long*: Only the activity with the longest duration is mapped.

[5] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

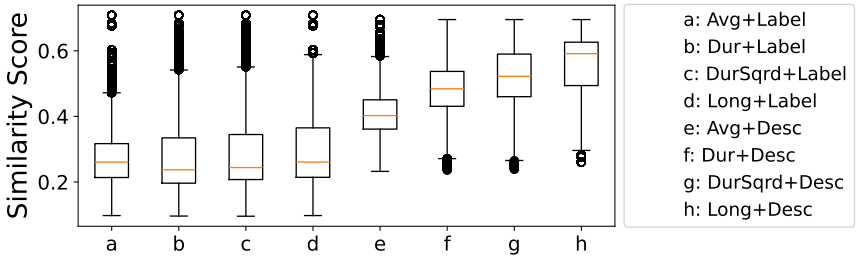


Fig. 4. The semantic similarity scores (based on Cosine similarity) for different mapping approaches between agenda segments and their five respective most similar POIs

To find the best mapping strategy, we use MTUS agenda segments with one of the 400 most frequent activity patterns, covering 75 % of all segments. This subset is chosen to focus only on the most influential activity patterns.

Fig. 4 shows the semantic similarity scores of the top five mappings between each of the agenda segments and all POIs, distinguished by the mapping strategies. The figure shows that the similarity score and thus the mapping quality can be increased by using a comprehensive description for POIs and activities. Furthermore, higher scores can be achieved with the mapping strategies *Long* and *DurSqrd*, which assign more weight to longer activities in the embedding.

4.3 Evaluation of Spatio-temporal and Semantic Properties

To evaluate how well ASTRA captures spatial, temporal, and semantic properties of mobility, we calculate a set of established mobility metrics [13] over the POI trajectories generated with ASTRA, and compare them against the Foursquare data set, and the synthetic POI trajectories created with the DITRAS baseline model. We provide a visual evaluation with comparative plots for each metric in Fig. 5 and calculate the error of fit between synthetic and real data with the Jensen-Shannon divergence (JSD) in Table 1.

JSD is often used to measure the dissimilarity between two distributions, P and Q, and builds on Kullback-Leibler divergence (KLD) [13]. JSD is defined as $JSD(P||Q) = (KLD(P||M) + KLD(Q||M))/2$, where M is the pointwise mean of P and Q and $KLD(P||Q) = \sum_{x \in X} P(x) \log(P(x)/Q(x))$. JSD is normalized and values close to zero indicate high similarity between P and Q.

Trip Distance: The distance of a trip measures the geographical Euclidean travel distance between two consecutive POI locations in a POI trajectory. We calculate the trip distance for each trip over all POI trajectories and plot the distribution in Fig 5 (a). ASTRA captures the shape of the distribution of real trip distances in the Foursquare data set while slightly overestimating longer trip distances. ASTRA achieves both visually and quantitatively similar results as DITRAS with an average JSD of 0.240 (ASTRA) versus 0.244 (DITRAS). Both approaches do not capture the high ratio of short trip distances that are prevalent in the Foursquare data.

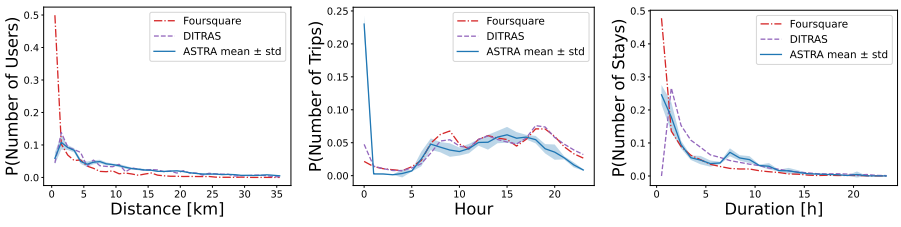
Table 1. Jensen-Shannon divergence between synthetic and real data per metric. Results for ASTRA are averaged over all nine tested configurations (described in Sect. 4.1). The respective best (lower) result in each column is highlighted in bold.

	Trip Distance	Trips per Hour	Stay Duration	Locations per User	Location Frequency	Visits per Location
ASTRA	**0.240**	0.106	**0.065**	0.224	**0.148**	**0.028**
DITRAS	0.244	**0.006**	0.294	**0.032**	0.237	0.600

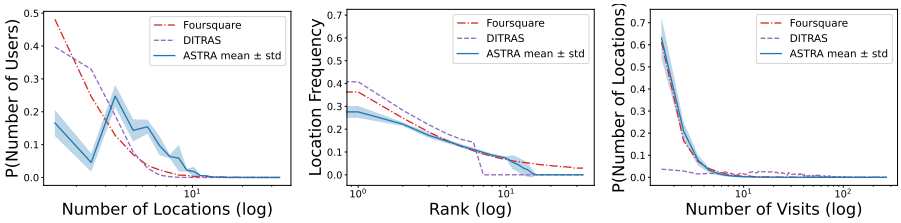
Trips per Hour: The number of trips per hour gives an indication of how realistic the generated data is in terms of its temporal distribution. To calculate this metric, the start times of the agents' visits to a POI are assigned to one hour time windows and counted. We plot the distribution of trips per hour over all POI trajectories over the 24 hours of the day in Fig. 5 (b). All three data sets show multiple peaks reflecting the typical human movement behavior to mainly move at specific times of the day [10]. The peaks coincide with the major travel times, for example when people go to work in the morning, go to lunch around noon, and return to their home in the evening. ASTRA is able to reproduce the overall shape of the real distribution with an average JSD of 0.106 while DITRAS can slightly better reproduce all peaks with a JSD of 0.006. For most of the tested configurations of ASTRA, the peaks are present at the typical moving times, however the most relaxed configuration using $S_{sem} = 0.3$ and $\tau_{sem} = 0.0$ can best reproduce the different peaks. Because ASTRA uses travel diaries that start at midnight, the first POI in each generated POI trajectory has a start time at midnight as well, causing another peak at this time.

Stay Duration: The duration of a stay indicates how much time an individual spends at a POI that they visit. This metric is calculated as the time difference in hours between the start times of consecutive POIs in an agent's POI trajectory. A stay time thus also includes the travel time between two POIs. This is due to the fact that POI trajectories from Foursquare and DITRAS only indicate the start timestamp of a visit to a POI. We plot the distribution of the stay duration over all agents' POI trajectories in Fig. 5 (c). ASTRA can capture the shape of the Foursquare distribution better than DITRAS with a JSD of 0.065 (ASTRA) versus 0.294 (DITRAS). In contrast to the real distribution, ASTRA shows a peak at approximately 8 h. This peak is likely to reflect the time duration that individuals stay at home or at work, whereas this might not be reflected in the Foursquare data set. A possible explanation is that Foursquare users preferably check in at food places, but not as often check in at home or work [16]. The high ratio of short trip distances in the Foursquare data set (see Fig. 5 (a)) also leads to a high ratio of short or zero durations of stay. DITRAS struggles to reproduce this peak, while ASTRA achieves more similar results.

Locations per User: The locations per user is the number of distinct POIs, that an agent visits during the period of observation. We plot the distribution



(a) Trip Distance (b) Trips per Hour (c) Stay Duration



(d) Locations per User (e) Location Frequency (f) Visits per Location

Fig. 5. Distributions of spatial, temporal, and semantic metrics

of locations per user over all agents in Fig. 5 (d). DITRAS reproduces the distribution of this metric in the Foursquare data set closely (JSD of 0.032), while ASTRA is showing a different pattern (JSD of 0.224) and produces POI trajectories with more POIs than Foursquare. The reason is likely that ASTRA is based on real travel surveys which often include POIs related to a person's home or work. Foursquare users, however, do not check in as often at such a type of POI, reducing the overall number of POIs of a user. Studies of travel diary data and mobile phone records show that about 90 % of human trajectories can be explained by fewer than seven regularly visited POIs [22]. ASTRA reproduces this trend while the majority of agents visits up to ten different POIs.

Location Frequency: The location frequency is the probability of an agent to visit a POI given its rank in the POI trajectory. The rank of a POI is its position in a descending order of visit frequency in the POI trajectory. A low rank thus indicates that a POI was visited often. In Fig. 5 (e) we plot the distribution of the location frequency given the rank over all POI trajectories. Both ASTRA and DITRAS can reproduce the shape of the real distribution approximately with a JSD of 0.148 (ASTRA) and 0.237 (DITRAS), but both approaches underestimate the distribution for higher ranks.

Visits per Location: The visits per location count the number of visits by all users to a unique POI location, that is a certain GPS coordinate. We plot the distribution of the number of locations given the number of visits in Fig. 5 (f). ASTRA can better reproduce the shape of the real distribution compared to DITRAS leading to a lower JSD of 0.028 versus 0.6 for DITRAS.

5 Discussion

The results of our experiments indicate that ASTRA is a suitable approach to generate realistic synthetic human mobility data that satisfies important spatial, temporal and semantic mobility measures. While the data quality is comparable or better than the DITRAS baseline, ASTRA additionally provides semantically accurate POI information and sociodemographic details about the generated agents. This information is usually not provided by other mobility models but essential for certain questions of (privacy) research.

ASTRA can flexibly be applied to any simulation region where the required input data is available. Census and map data are generally available for most countries, for example from WorldPop and OSM. However, the quality of the map data might vary per region which can limit ASTRA's applicability. Travel surveys can be obtained for many different countries [8]. While studies suggest that human mobility patterns are generally very similar [22], future work should investigate the question whether the semantic and temporal aspects of human travel patterns are similar in different countries. In this case, travel surveys might be interchangeable, which would eliminate the need for a geographically accurate input data set. Because ASTRA uses an embedding approach for the semantic mapping between activities and POI labels, their scope and formatting are not fixed, and therefore different data sources can be used. However, the preprocessing routines might need to be adapted to new source formats.

To create realistic, fair and unbiased data sets, the quality and size of the input data needs to be considered when choosing a simulation region and the respective input data set. In order for ASTRA to reflect realistic travel patterns, the used travel survey data must not contain exceptional events, such as the Covid-19 pandemic, which could distort results. It also needs to be large enough to ensure the synthetic population has diverse sociodemographic features.

Because ASTRA relies on agendas and needs to map agenda segments with specific activities and time range to a suitable POI, the distribution and availability of POIs in the chosen simulation area has an impact on the mapping quality. ASTRA's parameters control how much the semantic accuracy during mapping can be relaxed. Our experiments indicate that these parameters also influence how accurately the temporal patterns are preserved in the synthetic data while the impact on other analyzed metrics seems to be only small. As future work we therefore aim to test ASTRA's applicability to regions with different POI distributions, for example, smaller regions or urban and rural areas.

ASTRA is scalable to the desired number of artificial agents, because travel diaries can be selected multiple times from the travel survey data set if necessary. This makes the data useful also for other research, for example to study the effects of a growing population on the utilization of a city's transport network.

6 Conclusion

We have presented ASTRA, a novel approach to generate synthetic human mobility data. ASTRA creates POI trajectories that are representative of a society,

following an agenda-based approach based on real travel diaries. A major advantage of ASTRA is that its trajectories are explainable, because they retain semantic information about the contained POIs, and they can be linked to the sociodemographic features of their artificial agent. Because the data is not modified by privacy-preserving mechanisms, unlike other synthetic data sets, it is suitable for researching privacy-relevant questions regarding human mobility. ASTRA creates trajectories flexibly in any simulation region and is scalable with regard to the number of created trajectories. Our experiments show that ASTRA can create realistic mobility data that satisfies important metrics of mobility.

Acknowledgments. The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research, "Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project id: ScaDS.AI. This work was partially funded by Universities Australia and the German Academic Exchange Service (DAAD) grant 57701258.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Adenaw, L., Bachmeier, Q.: Generating Activity-Based Mobility Plans from Trip-Based Models and Mobility Surveys. Applied Sciences **12**(17), 8456 (2022)
2. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing **7**(6), 643–659 (2011)
3. Anda, C., Ordonez Medina, S.A., Axhausen, K.W.: Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. Transportation Research Part C: Emerging Technologies **128**, 103118 (2021)
4. Barbosa, H., De Lima-Neto, F.B., Evsukoff, A., Menezes, R.: The effect of recency to human mobility. EPJ Data Science **4**(1), 21 (2015)
5. Boeing, G.: Modeling and Analyzing Urban Networks and Amenities with OSMnx. Working paper. https://geoffboeing.com/publications/osmnx-paper/ (2024)
6. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature **439**(7075), 462–465 (2006)
7. Dandekar, A., Bressan, S., Abdessalem, T., Wu, H., Ng, W.S.: Trajectory simulation in communities of commuters. In: 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS). pp. 39–42 (2016)
8. Fisher, K., Gershuny, J., Flood, S.M., Backman, D., Vega-Rapun, M., Lamote, J., Sayer, L.C.: Multinational Time Use Study Extract System: Version 1.4 [dataset]. Minneapolis, MN: IPUMS (2022)
9. Gershuny, J., Vega-Rapun, M., Lamote, J.: Multinational Time Use Study [dataset] Centre for Time Use Research, UCL IOE, University College London [www.timeuse.org/mtus/] (2020)
10. González, M.C., Hidalgo, C.A., Barabási, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
11. Gursoy, M.E., Liu, L., Truex, S., Yu, L.: Differentially private and utility-aware publication of trajectory data. IEEE Transactions on Mobile Computing **18**(10), 2315–2329 (2019)

12. Hägerstraand, T.: What about people in regional science? Papers in Regional Science **24**(1), 7–21 (1970)
13. Kapp, A., Hansmeyer, J., Mihaljević, H.: Generative Models for Synthetic Urban Mobility Data: A Systematic Literature Review. ACM Computing Surveys **56**(4), 1–37 (2024)
14. Khan, M.T., Hussain, D.J., Tufail, M.: Activity-Based Tour Generation Model with Peshawar as Case Study. CONSTRUCTII Journal **6**(2), 12–16 (2023)
15. Li, J., Chen, W., Liu, A., Li, Z., Zhao, L.: FTS: A feature-preserving trajectory synthesis model. GeoInformatica **22**(1), 49–70 (2018)
16. Li, Y., Steiner, M., Wang, L., Zhang, Z.L.: Exploring Venue Popularity in Foursquare. Proceedings - IEEE INFOCOM pp. 3357–3362 (2013)
17. Madan, A., Cebrian, M., Moturu, S., Farrahi, K., Pentland, A.S.: Sensing the "Health State" of a Community. IEEE Pervasive Computing **11**(4), 36–45 (2012)
18. Mir, D.J., Isaacman, S., Caceres, R., Martonosi, M., Wright, R.N.: DP-WHERE: Differentially private modeling of human mobility. In: 2013 IEEE International Conference on Big Data. pp. 580–588 (2013)
19. Pappalardo, L., Rinzivillo, S., Simini, F.: Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. Procedia Computer Science **83**, 934–939 (2016)
20. Pappalardo, L., Simini, F.: Data-driven generation of spatio-temporal routines in human mobility. Data Mining and Knowledge Discovery **32**(3), 787–829 (2018)
21. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2019)
22. Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. Journal of The Royal Society Interface **10**(84), 20130246 (2013)
23. Schneider, M., Gehrke, L., Christen, P., Rahm, E.: D-TOUR: Detour-based point of interest detection in privacy-sensitive trajectories. In: INFORMATIK 2022. pp. 219–230. Gesellschaft für Informatik, Bonn (2022)
24. Song, C., Koren, T., Wang, P., Barabási, A.L.: Modelling the scaling properties of human mobility. Nature Physics **6**(10), 818–823 (2010)
25. Strobel, L., Pruckner, M.: OMOD: An open-source tool for creating disaggregated mobility demand based on OpenStreetMap. Computers, Environment and Urban Systems **106**, 102029 (2023)
26. Yang, D., Zhang, D., Zheng, V.W., Yu, Z.: Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems **45**(1), 129–142 (2015)
27. Yuan, N.J., Zhang, F., Lian, D., Zheng, K., Yu, S., Xie, X.: We know how you live: Exploring the spectrum of urban lifestyles. In: Proceedings of the First ACM Conference on Online Social Networks. pp. 3–14 (2013)
28. Zhang, Z., Zhou, L., Zhao, X., Wang, G., Su, Y., Metzger, M., Zheng, H., Zhao, B.Y.: On the validity of geosocial mobility traces. In: Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks. pp. 1–7 (2013)
29. Zheng, Q., Hong, X., Liu, J.: An Agenda Based Mobility Model. In: 39th Annual Simulation Symposium (ANSS'06). pp. 188–195 (2006)
30. Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X.: You are where you go: Inferring demographic attributes from location check-ins. WSDM'15 pp. 295–304 (2015)