IEEE Open Journal of
Intelligent Transportation
Systems

Received 17 May 2022; revised 22 August 2022; accepted 23 September 2022. Date of publication 27 September 2022; date of current version 12 October 2022.

Digital Object Identifier 10.1109/OJITS.2022.3209907

A Mobility Model for Synthetic Travel Demand From Sparse Traces

YUAN LIAO [1], KRISTOFFER EK [2], ERIC WENNERBERG [3], SONIA YEH [1], AND JORGE GIL [4]

[1] Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, 41296 Gothenburg, Sweden

[2] Burt Intelligence AB, 41109 Gothenburg, Sweden

[3] Einride AB, 41129 Gothenburg, Sweden

[4] Department of Architecture and Civil Engineering, Division of Urban Design and Planning, Chalmers University of Technology, 41296 Gothenburg, Sweden

CORRESPONDING AUTHOR: Y. LIAO (e-mail: yuan.liao@chalmers.se)

This work was supported in part by the Swedish Research Council Formas under Project 2016-1326, and in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant 821124.

ABSTRACT Knowing how much people travel is essential for transport planning. Empirical mobility traces collected from call detail records (CDRs), location-based social networks (LBSNs), and social media data have been used widely to study mobility patterns. However, these data suffer from sparsity, an issue that has largely been overlooked. In order to extend the use of these low-cost and accessible data, this study proposes a mobility model that fills the gaps in sparse mobility traces from which one can later synthesise travel demand. The proposed model extends the fundamental mechanisms of exploration and preferential return to synthesise mobility trips. The model is tested on sparse mobility traces from Twitter. We validate our model and find good agreement on origin-destination matrices and trip distance distributions for Sweden, the Netherlands, and São Paulo, Brazil, compared with a benchmark model using a heuristic method, especially for the most frequent trip distance range (1–40 km). Moreover, the learned model parameters are found to be transferable from one region to another. Using the proposed model, reasonable travel demand values can be synthesised from a dataset covering a large enough population of very sparse individual geolocations (around 1.5 geolocations per day covering 100 days on average).

INDEX TERMS Origin-destination estimation, sparse mobility traces, social media data, travel demand, trip distance distribution.

I. INTRODUCTION

Transportation accounts for 24% of global CO_2 emissions annually [1], presenting a major challenge to climate change mitigation. Meeting the challenge will require knowing the details of travel demand: how and how much people travel. Quantifying travel demand often relies on an origin-destination (OD) matrix [2], representing the intensity of flows of people between different zones/regions. Another extensively explored aspect is the trip distance distribution, which characterises how far people travel.

In transport planning and policymaking, different models are used to estimate travel demand either directly at the population level or through the detailed activity chains of agents. They rely on high-quality data collected through traditional methods including road traffic counting, household travel surveys, censuses, and population mobility models. These data collection methods are often costly, have small sample sizes, and are updated infrequently [3].

The increased prevalence of location-aware devices over the last decade has benefited our understanding of human mobility [4], [5]. Common sources include: call detail records (CDRs); GPS-enabled devices; tracking apps on smartphones; location-based social networks (LBSNs), e.g., Foursquare; and social media data, e.g., Twitter. The mobility traces obtained from these sources are promising in quantifying the flows of people between places and how far they travel [6].

Given that geolocations are collected with triggered phone activities or volunteered reports, one salient issue is to what extent the covered traces are incomplete, i.e., the sparsity

The review of this article was arranged by Associate Editor Meng Li.

issue. Data sources like CDRs, LBSNs, and social media data only provide a partial view of the actual mobility trajectories [7]. The incompleteness of the traces limits the accuracy of the estimated travel demand. Nevertheless, these sources are collectively abundant, especially LBSNs and social media data, which are relatively less expensive and available globally. In order to extend the use of these data sources, it is important to have appropriate techniques to fill the gaps in sparse mobility traces.

This study proposes a mobility model that fills the gaps in sparse mobility traces, tested on geolocations collected from social media data. Using the model-processed data, one can subsequently synthesise travel demand on two aspects: the share of trips between spatial zones and the trip distance distribution. The proposed model extends the fundamental mechanisms of exploration and preferential return to synthesise mobility trips [8] for accommodating the individually-sparse but collectively abundant mobility traces. We first calibrate and validate the model with official data on daily travel demand. We then apply the model to represent the travel demand in two countries and one metropolitan region. The model generates good transferability of its parameters from one region to another.

The remainder of this paper is organised as follows. The rest of this section reviews the work related to different data sources used to estimate the two aspects of travel demand: trip distance distribution and flows between spatial zones. It covers the shortcomings of these data sources, specifically related to sparsity, followed by a brief summary of the objectives of the present study. Section II describes the model design, and Section III describes the model experiment. The results are presented in Sections IV, and Section V discusses the findings and identifies future research needs and the conclusions.

A. RELATED WORK

Common models for travel demand estimation include the four-step model [9], activity-based models [10], and agent-based models (ABM) with a synthetic population [11]. These models rely on data collected from traditional travel surveys and censuses. For instance, a study uses the data from an yearly census and a national household travel survey to create a synthetic population and its travel demand [11]. These data sources have careful sampling designed to statistically represent the true population. However, they also have many shortcomings such as being costly to collect and having low sampling rates, short survey duration, under-reporting of trips, and being out-of-date [12]. Travel surveys also fail to capture most of the infrequent long-distance trips [13].

Travel demand estimation has benefited from increasingly available location-aware devices [12] that provide a variety of human mobility records. Using data from GPS-enabled devices, a multi-scale model has been proposed to synthesise mobility traces that yield representative trip distance distribution [5]. Another study has updated origin-destination matrices using aggregated GPS data [14]. The movements of a large population can be captured by CDRs [15], [16], [17]. CDRs have been used to develop a microscopic individual mobility model [8] and reveal fundamental mobility laws such as the distance-frequency scaling law [18]. Wang *et al.* (2018) have explored social disparities of travel distances using 650 million geotagged tweets [19]. Liao *et al.* (2021) have modelled the overall travel demand using geolocations of Twitter data, showing good agreements with the ground truth data [20].

However, data collected from CDRs, LBSNs, and social media are collectively abundant but individually sparse. For example, in Twitter data, the top geotag users generate 1–3 geolocations per active day on average as revealed by the present study. In other words, these data sources capture incomplete mobility trips because they do not record all the locations a user has visited. Due to this sparsity issue, estimating travel demand using CDRs is not very feasible [21]. Similarly, sparse traces from social media data yield sparse origin-destination matrices (ODMs) [22].

In order to address the sparsity issue, studies have developed different techniques to fill the gaps in sparse individual mobility traces. Typical techniques include heuristic methods and mathematical models. Heuristic methods that are widely used in processing sparse traces consist of intuitive rules. For example, a CDR entry can be regarded as a stay that lasts for a certain time period, e.g., one hour [23]; the missing entries between 10 pm and 7 am, when a user is assumed to be at home, are filled with the home location estimated based on the user's historical records [21]. When using sparse traces, the reported geolocations need to be processed to become trips. A widely applied practice is to connect the two consecutive geolocations and filter out connections with a time interval longer than a selected time threshold, e.g., 4 hours [22], [24]. However, these heuristic methods using time-based rules are arbitrary. Moreover, such filtering leads to a massive reduction of available data which does not reflect true mobility patterns.

Beyond the heuristic methods, a variety of mathematical models have been designed to bridge the gaps in the sparse mobility traces to increase their usability in understanding mobility patterns. Chen *et al.* (2019) have developed a technique called Context-enhanced Trajectory Reconstruction that completes individual CDR-based traces using tensor factorisation [7]. The synthesised data deliver a trip distance distribution with a better fit among other key mobility indicators. Their study suggests that filling the gaps in the sparse individual traces results in better representation of travel demand, e.g., the truncated power-law distribution of trip distance distributions. Burkhard *et al.* (2017) have reconstructed regular mobility patterns from users with sparse CDRs using idiosyncratic daily patterns from clustered daily activities [25].

With the exception of these few studies, most studies design methods that directly extract patterns from sparse mobility traces [19], [20], [26]. The generally overlooked bias from data sparsity can affect the observed mobility patterns [7] and limit their usability for travel demand estimation.

B. STUDY OBJECTIVES

Sparse mobility traces collected from CDRs, LBSNs, and social media data have been widely used to study mobility patterns. However, most studies use them directly and ignore the impact of the sparsity issue, or apply simple heuristic methods, both of which lead to results that are potentially biased and inaccurate. In order to extend the use of these data, it is crucial to design appropriate techniques to fill the gaps in sparse mobility traces.

To bridge the gaps in the literature, we propose a mobility model to deal with sparse mobility traces, tested on geolocations of social media data. We calibrate and validate the model with the other established data sources in the form of origin-destination matrices quantifying the daily travel demand in Sweden, the Netherlands, and São Paulo, Brazil. Specifically, we attempt to answer the following research questions:

- Can we develop a model that fills the gaps in sparse mobility data for a more accurate travel demand estimation?
- How well does the model perform compared with heuristic methods?

II. MODEL DESIGN

This section proposes a model that fills the gaps in sparse mobility traces. The model-synthesised data are used to obtain individual trips for synthetic travel demand. We start with a problem statement (Section II-A) defining the sparse input and the synthesised output. In Section II-B, we describe the features extracted from the sparse traces for modelling. Then in Section II-C, we describe how the model components work together to synthesise mobility data.

A. PROBLEM STATEMENT

Part of the mobility traces of a given individual are observed via CDRs or social media platforms over a certain duration, expressed as: $\mathbf{Trac} = \{(X, Y)_p \mid 1 \leq p \leq N\}$ where X and Y are the decimal degree of latitude and longitude respectively, and p is the chronological order index of the observed **visits** to a variety of **locations** ranging from 1 to the total number of visits by the individual, i.e., N. **Locations** are distinguished by their recorded coordinates (X, Y), however, their spatial resolution varies depending on logging noise, cell tower coverage of CDRs, or different social media platforms. Therefore, in practice, some preprocessing is needed to cluster the raw location coordinates so that their spatial resolution is more consistent. After preprocessing the raw data, we refer to a location as a unique pair of GPS coordinates.

However, the sparse traces **Trac** are incomplete, they do not include all the locations visited by an individual, and are biased by the associated activity, be it tweeting or making a phone call, depending on how frequently, at what time and where the specific activity is typically performed. In order to fill the gaps, the proposed model takes **Trac** as input and synthesises them into a more representative set of mobility data, $\mathbf{Trac}'$, for travel demand estimation.

As model output, the synthesised mobility traces $\mathbf{Trac}' = \{(X, Y)_{day,m} \mid 1 \leq day \leq D, 1 \leq m \leq M_{day}\}$ represent visits of an individual that happen in a series of simulation days (day) where m is the chronological order index of a visit to a location (X, Y) in a simulation day. A simulation day is a working unit of how the model generates synthesised data, as specified in Section II-C.3. The total number of simulation days (D) is determined when the aggregate output of the model-synthesised data stabilises (see Appendix C). The number of visits per simulation day, M_{day}, is empirically determined by looking at how many displacements are usually made by the population or a specific individual. In the experiment of this study, we use the Swedish National Travel Survey (2011–2016) [27] to get the distribution of the number of visits per day across all survey participants. For each simulation day of each individual, M_{day} is randomly drawn from that distribution (detailed in Appendix C).

B. FEATURE EXTRACTION

For a given individual, a number of features can be extracted from the model input **Trac**. These features are later used for synthesising mobility data.

The set $\mathbf{S}$ is defined as a collection of all the distinct locations having different values of (X, Y). The number of distinct locations in $\mathbf{S}$ is indicated by n. The frequency rate of them being visited is expressed as $f_j, j = 1, 2, \ldots, n$. Among these locations, the home location s_h is identified as the most-visited location between 7 pm and 8 am on weekdays and the whole day on weekends [19], [20], [28].

The jump size θ_p connecting two consecutive observed locations, s_p and s_{p+1}, is defined as the Haversine distance between them. The bearing α_p, referring to the direction from s_p to s_{p+1}, is an angle measured clockwise from the north direction. The set of the jump size and the bearing of all the pairs of consecutive locations in **Trac** is expressed as $\mathbf{J} = \{(\theta, \alpha)_p \mid 1 \leq p \leq N - 1\}$.

C. SYNTHESISING MOBILITY DATA

Given **Trac**, the model sets the individual at home (s_h) to start the simulation day. As shown in Figure 1, the model generates the next location given current location s_p with two options: 1) to return to a previously visited location $s_j \in \mathbf{S}, j \neq p$ with a probability of Prob(return) or 2) to explore a new location with a probability of Prob(explore) where we have $\text{Prob(explore)} + \text{Prob(return)} = 1$. According to the individual mobility model [8], the probability of exploring a new location is expressed as:

$$\text{Prob(explore)} = \rho n^{-\gamma} \tag{1}$$

where the greater the n, the smaller the probability of exploring a new location and ρ and γ control how much n affects

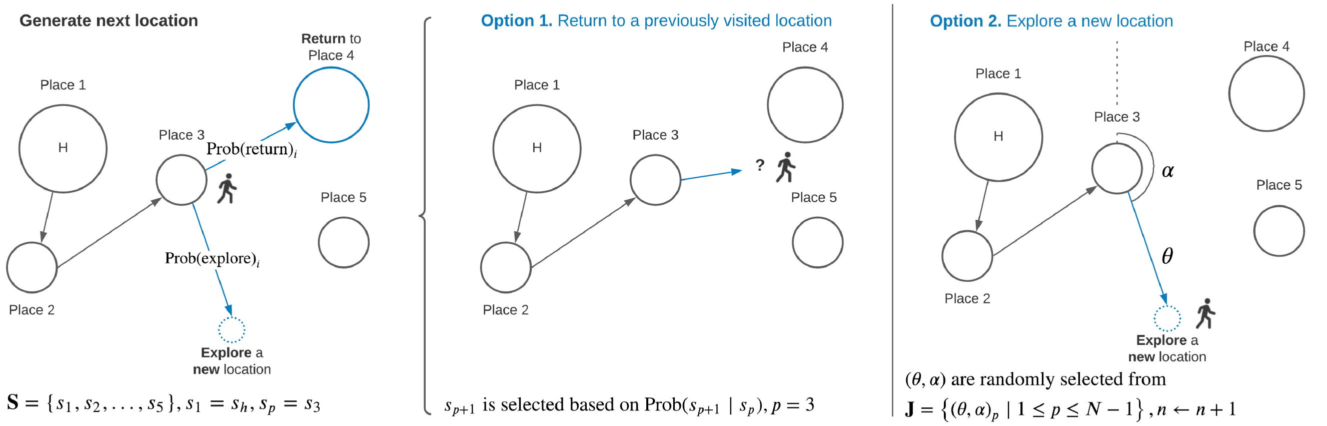

Generate next location
Return to Place 4
Place 1
H
Place 3
Prob(return)$_i$
Place 5
Prob(explore)$_i$
Place 2
Explore a new location
$\mathbf{S} = \{s_1, s_2, \ldots, s_5\}, s_1 = s_h, s_p = s_3$
Option 1. Return to a previously visited location
Place 4
Place 1
H
Place 3
?
Place 5
Place 2
s_{p+1} is selected based on Prob$(s_{p+1} \mid s_p), p = 3$
Option 2. Explore a new location
Place 4
Place 1
H
Place 3
α
Place 5
θ
Place 2
Explore a new location
(θ, α) are randomly selected from
$\mathbf{J} = \{(\theta, \alpha)_p \mid 1 \leq p \leq N-1\}, n \leftarrow n+1$

FIGURE 1. Model framework for generating synthetic mobility data. An example of an individual with 5 unique observed locations. The individual is at present located at Place 3 ($s_p = s_3$) and has the option of either returning to a previously visited location (Option 1) or exploring a new location (Option 2).

the probability. Given the same n and γ, the greater the ρ, the higher the probability of exploring. Given the same ρ, the greater the γ, the more rapidly declining Prob(explore) as n increases.

1) RETURN TO AN OLD PLACE

If a return is generated, the model moves this individual to a previously visited location in $\mathbf{S}$. This location is selected from all the candidate locations in $\mathbf{S}$ that have unequal probabilities. The probability of a candidate place s_{p+1} considering the current location s_p is determined by two factors, **visitation frequency** $P(s_{p+1}|s_p)$ and **impedance to the candidate places** $I(s_{p+1}|s_p)$.

Visitation frequency: The sparse traces are often collected passively, i.e., being opportunistic due to their association with certain activities. They have biased visitation frequency of observed places. For social media data, habitual places such as home and work are much less reported relative to uncommon places [29]. For CDRs, the sparse traces are biased toward locations of phone activities [30]. However, one may expect the rank order of places, based on their visitation frequency from sparse traces, to be preserved, if not the absolute frequency [21]. Therefore, we define visitation frequency $P(s_{p+1}|s_p)$ as:

$$P\big(s_{p+1} \mid s_p\big) = \frac{k_{s_{p+1}}^{-\zeta}}{\sum_{s_{p+1} \in \mathbf{S}, s_{p+1} \neq s_p} k_{s_{p+1}}^{-\zeta}} \tag{2}$$

where $k_{s_{p+1}}$ represents the rank order of location s_{p+1}, which is the kth most visited location whose visitation frequency follows Zipf's law $k_{s_{p+1}}^{-\zeta}$ where $\zeta \approx 1.2 \pm 0.1$ [8].

Impedance to the candidate places: The other factor affecting the selection of returning to an old place is the distance (travel impedance) from the current location to the candidate place. Naturally, people are more likely to visit nearby locations over distant ones [18]. Besides, the incorporation of the travel impedance factor helps to further correct the biases of rank order of locations in the sparse data, to avoid their frequency dominating the visitation probability of different candidate places. We define this impedance term $I(s_{p+1} \mid s_p)$ as:

$$I\big(s_{p+1} \mid s_p\big) = \frac{\exp\big(-\beta \cdot \theta\big(X_p, Y_p, X_{p+1}, Y_{p+1}\big)\big)}{\sum_{s_{p+1} \in \mathbf{S}, s_{p+1} \neq s_p} \exp\big(-\beta \cdot \theta\big(X_p, Y_p, X_{p+1}, Y_{p+1}\big)\big)} \tag{3}$$

where $\theta(X_p, Y_p, X_{p+1}, Y_{p+1})$ is the distance between a candidate place s_{p+1} and the current location s_p. To keep the model generic and boundary-free, we use Harvesine distance. And the parameter β controls the degree to which a given individual is constrained by distance. The higher the β, the more likely the individual is to visit places nearby.

Combining 2 and 3, the selection of a return location is associated with the distances from the current location to the candidate places as well as the historical visitation frequency indicating the importance levels of these candidate places:

$$\text{Prob}\big(s_{p+1} \mid s_p\big) = \frac{P\big(_{p+1} \mid s_p\big) I\big(s_{p+1} \mid s_p\big)}{\sum_{s_{p+1} \in \mathbf{S}, s_{p+1} \neq s_p} P\big(s_{p+1} \mid s_p\big) I\big(s_{p+1} \mid s_p\big)} \tag{4}$$

2) EXPLORE A NEW PLACE

If exploring a new location, the model moves individual i to an unobserved location s_{p+1} ($s_{p+1} \notin \mathbf{S}$). The new location is determined by the current location s_p and the jump size θ and bearing α randomly selected from $\mathbf{J}$ as illustrated in Figure 1-Option 2:

$$X_{p+1}, Y_{p+1} = \text{shift}\big(X_p, Y_p, \theta, \alpha\big) \tag{5}$$

where the function shift computes the coordinates of the new location by moving the jump size of θ along the clockwise direction of the bearing angle α (the north as zero degrees).

Every time a new place is selected, the total number of distinct places visited n is updated, $n \leftarrow n + 1$.

3) GENERATE SIMULATION DAYS

For a simulation day with M_{day} visits, the individual departs from s_h to visit a series of locations, where the last one is

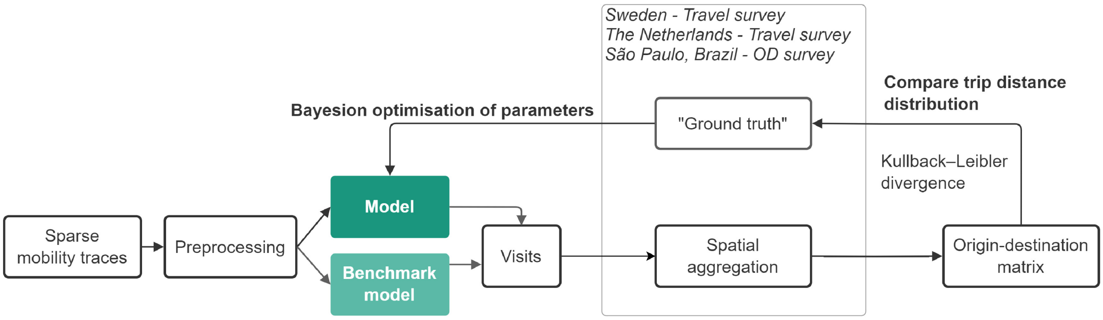

Sweden - Travel survey
The Netherlands - Travel survey
São Paulo, Brazil - OD survey
Compare trip distance distribution
Bayesion optimisation of parameters
"Ground truth"
Kullback–Leibler divergence
Model
Sparse mobility traces
Preprocessing
Benchmark model
Visits
Spatial aggregation
Origin-destination matrix

FIGURE 2. Experiment of the proposed model.

Algorithm 1: Synthesising Mobility Data Using Sparse Traces From an Individual

```
Data: $\rho, \gamma, \zeta, \beta, D, M$, Trac
Result: Trac$'$
$n$, S, $s_h$, J ← FeatureExtraction[Trac];
Trac$'$ ← [ ];
while $day < D$ do
    append $s_h$ to Trac$'$;
    $s_p \leftarrow s_h$;
    while $m < M_{day} - 2$ do
        Prob(explore) ← $\rho n^{-\gamma}$;
        t ← generateRandomNumber[0,1];
        if $t \leq$ Prob(explore) then
            $\theta, \alpha$ ← selectJumpBearing[J];
            $s_{p+1}$ ← shift($s_p, \theta, \alpha$);
            $n \leftarrow n + 1$;
        else
            Prob$(s_{p+1} \mid s_p)$ ←
            ReturnProbability[$s_{p+1}, s_p, \zeta, \beta$];
            $s_{p+1}$ ← selectPlace[Prob$(s_{p+1} \mid s_p)$];
        end
        append $s_{p+1}$ to Trac$'$;
        $s_p \leftarrow s_{p+1}$;
        $m \leftarrow m + 1$;
    end
    append $s_h$ to Trac$'$;
    $day \leftarrow day + 1$;
end
```

also s_h. For $M_{day} - 1$ visits, each location is created by either returning to an old place (Section II-C.1) or exploring a new place (Section II-C.2). As illustrated in Algorithm 1, after the specified simulation days (D) are finished, the mobility data of the individual (**Trac**$'$) are synthesised by using the sparse input **Trac**.

III. MODEL EXPERIMENT

Considering the ground-truth data availability and the potential impact of geographical scales on the model performance, we select Sweden, the Netherlands, and São Paulo to do the model experiment. These three regions have similar population sizes but distinct areas; São Paulo is a metropolitan area whereas Sweden and the Netherlands are two countries of different sizes (detailed in Table 1). Specifically, we use a benchmark model (Section III-A) with geotagged tweets as an example of sparse mobility traces (detailed in Appendix B).

TABLE 1. Descriptions of model parameters and assumptions.

	Description	Value
ρ, γ	Two parameters that control exploring probability	0.01–0.99
β	The parameter that controls returning impedance	0.01–0.99
ζ	The parameter of Zipf's Law ($f \sim k^{-\zeta}$)	1.2
D	No. of simulation days	260
M_{day}	No. of visits to locations per simulation day (day)	$M_{day} \sim \text{Prob}(\mathbf{F})$

As illustrated in Figure 2, we first construct models for the three study areas and calibrate the models against the official travel survey data as the ground truth to find the optimal parameters. The aim of the experiment (Section III-B) is to see how the model performs in representing the travel demand, as quantified by the aggregated population flows between spatial zones, when validated against official data sources. The model performance is evaluated by comparing the ODM and its trip distance distribution with the ground truth in contrast with the benchmark model.

A. BENCHMARK MODEL

In assessing the performance of the model's synthetic travel demand estimation, we create a benchmark model using a common heuristic method of generating an origin-destination matrix (ODM) based on sparse mobility traces. The benchmark model converts the displacements of two consecutive geolocations generated by the same individual with a time interval below 24 hours into trips [22], [24], [31]. The origin-destination pairs of these converted trips go through spatial aggregation for all the covered individuals to formulate the benchmark ODM to be compared with the ground truth together with the proposed model. The performance gain between the proposed model and the benchmark model quantifies to what extent the proposed model corrects the biases in sparse traces, thus contributing to an improved travel demand estimation at the aggregate level.

B. MODEL EXPERIMENT

The preprocessed sparse geolocations, as described in Appendix B, are ordered chronologically and divided into two equal-length parts, one part for calibration and the other for validation. With an initial parameter setting, the model takes in sparse traces for each individual (**Trac**) to generate visits (**Trac**′). All the individuals' visits are further aggregated on the spatial zones consistent with the ground-truth data to calculate the ODM. The calculated ODM is compared with the ground truth in terms of the trip distance distribution using the Kullback-Leibler (KL) divergence measure [20], [32], [33]. A small KL divergence value indicates that the two distributions are similar. The optimal model parameters are those that yield the smallest KL divergence with Bayesian optimisation. The model with optimal parameters is applied to the validation dataset, and the performance (KL divergence from the ground truth) is compared to that for the calibration dataset.

1) MODEL SETTINGS

The initial model setup is illustrated in Algorithm 1 (Data). Except for the input of sparse traces (**Trac**), the model has a few parameters that need to be set in order for it to synthesise mobility data. The meanings and values of these parameters are displayed in Table 1. Prob(**F**) is the probability of a set of values of No. of visits to locations per day, **F**, which is empirically derived from the Swedish National Travel Survey (2011–2016) [27]. See the detailed distribution in Figure 1. D is determined based on the exploration of the relationship between the model's performance, KL divergence, with a varying value of the number of simulation days (detailed in Figure 1). For three of these parameters, ρ, γ, and β, Bayesian optimisation on model outputs against the ground-truth data is used to specify the values within the intervals in Table 1. This is introduced in the rest of this section.

2) GROUND-TRUTH DATA

We use the travel survey data covering detailed trip information, such as the origin, destination and distance for individual trips, from three selected regions as shown in Figure 3. Given that some validation data only report weekday travel, for the sake of consistency, we focus on weekday trips.

Sweden: The Swedish National Travel Survey collects one-day travel diaries for 2011 to 2016 [27]. The survey includes 171,553 trips from 38,258 participants with 2,189 record days [20]. This dataset contains the origins and destinations of trips as well as trip distance. The spatial resolution is the DeSO zone defined as 5,984 demographic statistics areas by Statistics Sweden.

The Netherlands: The dataset of daily mobility OViN (Onderzoek Verplaatsingen in Nederland) [34] is a survey conducted in 2017 with 37,016 respondents at the national level. All trips originate and end in postal code areas, grouped by their first four digits. In total, there are 4,066 zones.

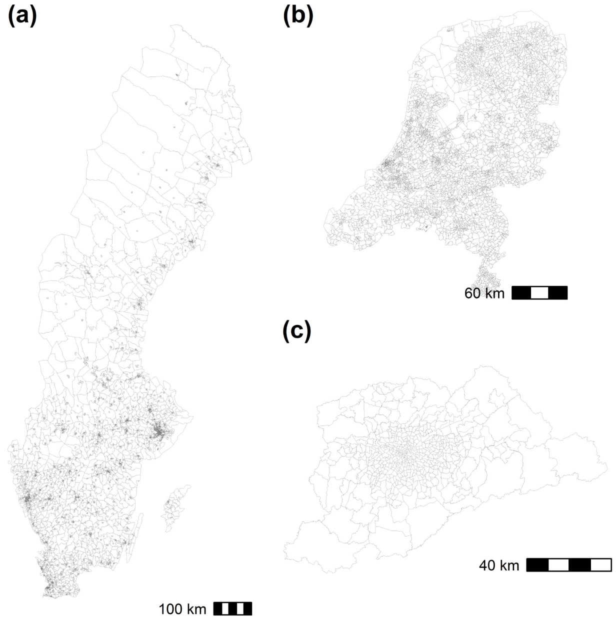


FIGURE 3. Spatial zones. (a) Sweden. (b) The Netherlands. (c) São Paulo, Brazil.

São Paulo, Brazil: The OD survey [35] carried out in 2017 interviewed 32,000 households (100,000 people) for their recorded weekday. There are 517 spatial zones, of which 342 zones correspond to the municipality of São Paulo, and the rest cover the neighbouring municipalities. This dataset does not have detailed trip distances. The trip distances of the OD pairs are calculated based on the Haversine distance between the centroids of the corresponding origin and destination zones.

3) BAYESIAN OPTIMISATION

In the optimisation process, we aim to find the optimal values of the undetermined parameters listed in Table 1 so that the calibrated model approximates the ground truth as closely as possible. Bayesian optimisation is a global optimisation that does not specify any forms of functions; it finds the optimal parameters given the objective function by taking advantage of the full information provided by the history of the optimisation [36].

In this study, the objective function KL divergence is defined below:

$$D_{KL}(P\|Q) = \sum_{d \in d_{group}} P(d) \log \frac{P(d)}{Q(d)} \tag{6}$$

where d_{group} is a set of quantile-based distance groups (100 quantiles) based on the spatial zones of the study area and $P(d)$ is the frequency rate of trips that fall in a given distance group $d \in d_{group}$ based on the ground-truth data:

$$P(d) = F(\text{ground truth}) \tag{7}$$

TABLE 2. Optimal model parameters for the three regions in comparison with the benchmark.

Region	Parameter			KL divergence			
				Calibration data		Validation data	
	ρ	γ	β	Model	Benchmark	Model	Benchmark
Sweden	0.98	0.24	0.01	0.007	0.091	0.011	0.017
The Netherlands	0.80	0.17	0.17	0.004	0.012	0.004	0.018
São Paulo, Brazil	0.98	0.18	0.16	0.003	0.074	0.003	0.140

while $Q(d)$ is the frequency rate of trips in a given distance group $d \in d_{group}$ based on the model output i.e., its synthesised mobility data from all the individuals $i = 1, 2, \ldots, I$:

$$Q(d) = F\big(\rho, \gamma, \beta, \big[\mathbf{Trac}'_1, \mathbf{Trac}'_2, \ldots, \mathbf{Trac}'_I\big]\big) \tag{8}$$

where ρ, γ, and β are the target parameters whose optimal values are selected to maximise $-D_{KL}$ (minimise D_{KL}).

We use a constrained global optimisation package in Python that is built upon Bayesian inference and Gaussian process [37]. The technique is chosen over other alternatives, e.g., a grid search, due to the high computation cost of calculating the objective function starting with sparse traces. Moreover, this technique allows a balance between exploration and exploitation in searching for the optimal parameters [37].

IV. RESULTS

In this section, we first present the model calibration and validation results with the optimal parameters (Section IV-A) and then test the model's performance in representing travel demand (Section IV-B), and the impact of trip distance and length of sparse traces on the model's performance (Section IV-C). In the last of this section, we discuss model parameter transferability (Section IV-D).

A. CALIBRATED MODELS FOR SWEDEN, THE NETHERLANDS, AND SÃO PAULO, BRAZIL

In model calibration, the Bayesian optimisation searches over the parameters' value space to find the optimal set of ρ, γ, and β for the three case study regions. The results are presented in Figure 4. In the search through the parameter space, the KL divergence varies similarly for the three geographical regions.

Table 2 summarises the optimal model parameters and corresponding model performance in terms of KL divergence for the calibration and validation datasets. The performance difference between the calibration and validation datasets is small for the Netherlands and São Paulo; it is slightly greater for Sweden. Compared with the benchmark, the proposed model approximates the ground truth better: KL divergence decreases from the benchmark to the proposed model 67% – 96% for the calibration data and 35% – 98% for the validation data.

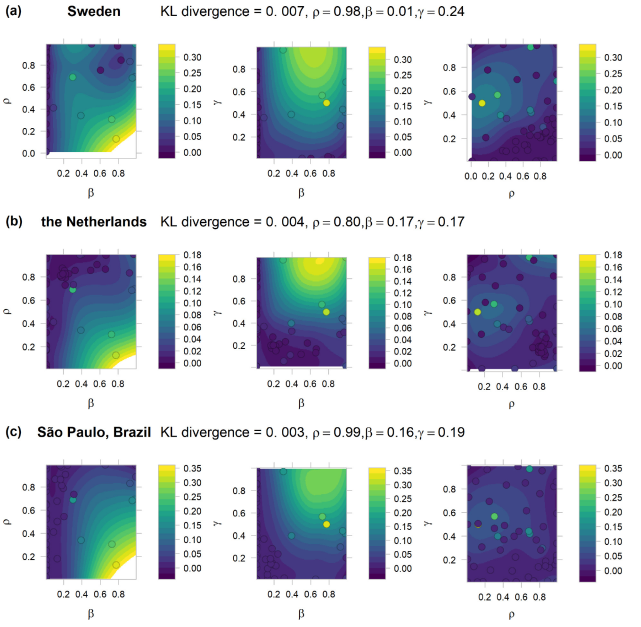


FIGURE 4. Parameter search results. (a) Sweden. (b) The Netherlands. (c) São Paulo, Brazil. A circle represents one combination of parameters with its colour indicating the KL divergence. The surface is interpolated from the circles. The cooler the colour (deep blue), the smaller the KL divergence.

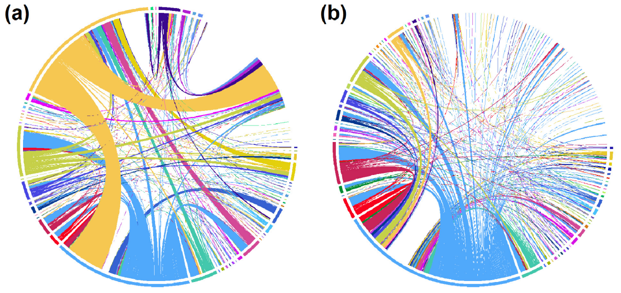


FIGURE 5. Individual mobility ODM from a selected individual living in São Paulo, Brazil based on the (a) benchmark model, and (b) proposed model. Each arc indicates a spatial zone. The more arcs, the more spatial zones covered by visits.

Figure 5 shows an example of generated individual ODMs using the benchmark model vs. the proposed model based on sparse geolocations of an individual covering 315 days. In Figure 5(a), the sparse geolocations are directly used by the benchmark to produce the individual ODM, resulting in 64 spatial zones between which the trips are created. The proposed model, on the other hand, fills the gaps in the sparse data resulting in more diverse synthetic trips covering 123 spatial zones (Figure 5(b)). For both ODMs from the benchmark and the proposed model, the blue arc represents the home location. For daily travel, many trips are originated from or attracted to the home. We see the proposed model better reflects such a pattern compared with the benchmark.

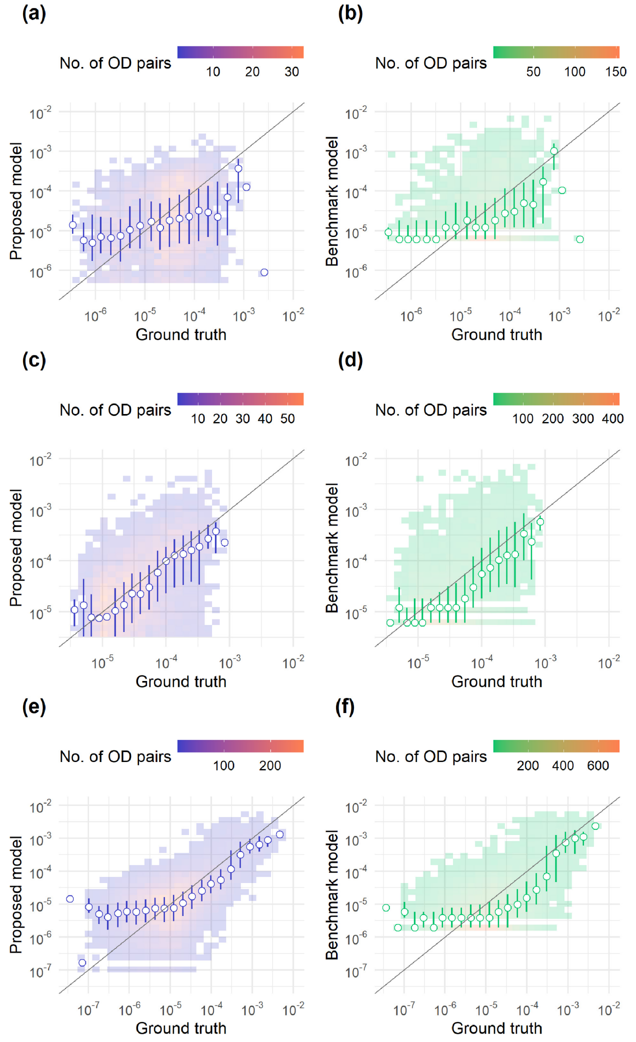


FIGURE 6. Trip frequency rate between zones (calibration results) using the proposed model (y axis) in contrast with ground truth (x axis). The gray diagonal line corresponds to a perfect agreement between the ground-truth data and the model/benchmark output. Heat maps of point counts show the distributions of No. of OD pairs. Circles are median values for each bin and lines are the 0.25-0.75 quantiles. Left columns are the proposed model and right are the benchmark model, for (a-b) Sweden, (c-d) the Netherlands, and (e-f) São Paulo, Brazil.

B. POPULATION FLOWS: ODMS AND DISTANCE DISTRIBUTION

We quantify the population flows between the spatial zones in the study areas by aggregating the results of all the individuals from the proposed and benchmark models, and compare with the trips in the ground-truth data. Compared with the ground-truth data, four trip frequency rate values are calculated from the proposed model vs. the benchmark model, using the calibration data vs. the validation data. These four model-based frequency rates are each compared with the one from the ground-truth data.

As illustrated in Figure 6, if the model performs the same as the ground truth, all points will fall on the diagonal line. We see that the model generally performs better for OD pairs of higher frequency rate than for those of lower frequency rate. We also observe that the performance varies between the three regions. Compared with the benchmark model results, the proposed model generates more representative trips that generally approximate the ground truth better.

TABLE 3. Model performance regarding the similarity between ODMs. For all correlation tests, $p < 0.001$.

Region	Kendall's tau		SSI	
	Model	Benchmark	Model	Benchmark
Sweden	0.19	0.24	0.31	0.31
The Netherlands	0.44	0.37	0.43	0.39
São Paulo	0.45	0.33	0.54	0.45

Besides the visualisation in Figure 6, we use two indicators, Kendall's tau and the Sørensen–Dice similarity index (SSI) [38], to further compare the performance of the proposed and benchmark models. Kendall's tau quantifies the correlation of the trip frequency rate of all the spatial zones between the ground truth and the model vs. the benchmark outputs. The SSI takes values between 0, when there is no similarity, and 1, when the model output and the ground-truth data are identical. Taking the average of validation and calibration, their results are shown in Table 3.

The similarity scores (KL divergence) for the trip distance distribution of the ground truth and model outputs against the benchmark are included in the CDF plots of Figure 7. The proposed model approximates the ground truth better than the benchmark model, i.e., the blue curves are closer to the orange curve than the green curves. Moreover, the benchmark model tends to underestimate the trip distance. For example, trips below 10 km account for 75 – 90% of total trips in all three regions according to the benchmark models. However, the ground-truth data and the model outputs suggest the shares of 75%, 80%, and 55% approximately, which largely depend on the regions. The overall similarity results are consistent with the results of ODMs. In all three regions, the model applied to the calibration dataset approximates the ground-truth data slightly better than the one applied to the validation dataset.

For ODMs and distance distributions, the proposed model generally performs better than the benchmark model. There is one exception for Sweden: the similarity of ODMs between the model output and the ground-truth data is the same or worse than the benchmark. But its KL divergence indicates better performance than the benchmark. In summary, there is a consistent regional difference for both ODMs and distance distribution: the proposed model performs the best in São Paulo, followed by the Netherlands, and Sweden.

C. IMPACT OF TRIP DISTANCE AND LENGTH OF SPARSE TRACES

How the model approximates the ground truth of trip frequency rate depends on trip distance and region (Figure 8). The model output is very close to the ground truth data for the most frequent trip distance range (1–10 km). For the rest of the trip distance ranges, the model slightly underestimates the trip frequency for distances between 10–30 km and overestimates above 30 km up to 100–300 km in the two countries (Figure 8a-d). When the trip distance increases

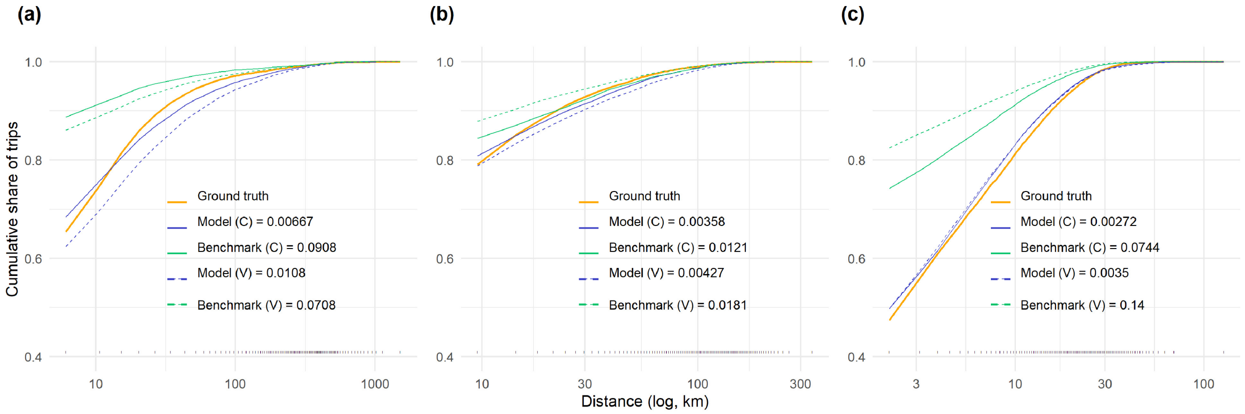


FIGURE 7. CDF (cumulative distribution function) plots to compare the trip distance distributions of the ground-truth data (orange), the proposed model (blue), and the benchmark model (green): (a) Sweden. (b) The Netherlands. (c) São Paulo, Brazil. C stands for calibration data and V validation data. Values following the symbol "=" are the model's corresponding KL divergence as compared with the ground truth.

above 100–300 km, the trip frequency in the ground-truth data starts to fluctuate, and its value difference between the model output rises. For São Paulo (Figure 8e-f), the model approximates the ground truth well as opposed to the benchmark that greatly overestimates short-distance trips below 3 km. The model output is similar to the ground-truth data for the rest of the distance ranges up to 40 km. However, the model overestimates the occurrence of long-distance trips above 40 km within São Paulo.

The similarity between the model output and the ground truth of trip distance distribution depends on data length (Figure 9). We consider two types of data length: the total number of geolocations and the maximum number of geolocations used for each individual. The more geolocations we have in our model, the better its output resembles the ground truth (Figure 9a). For all the regions, we see a continuous increase in performance (declining KL divergence) and such trend even holds after we include all the individuals' data, especially for Sweden, the largest among the study areas, whose performance is far from saturation, unlike São Paulo. However, the model performance is not sensitive to increasing the maximum number of geolocations of each individual (Figure 9b). It seems a maximum of 200 geolocations per individual, even a large number of individuals have much less than 200 (median value about 140 per individual), suffices for generating similar trip distance distribution to the ground truth. Figure 9 suggests that a dataset covering a large enough population with a relatively small number of individual geolocations can be enough for the model to generate sensible travel demand.

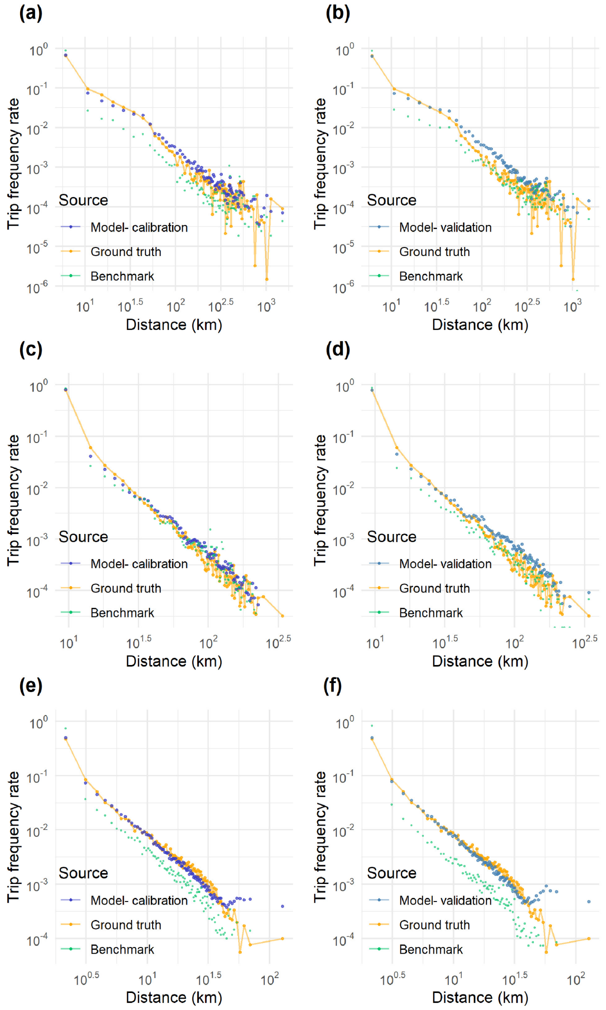


FIGURE 8. Trip frequency rate between zones (y axis) as a function of trip distance (x axis). Left columns are the calibration results and right are the validation results, for (a-b) Sweden, (c-d) the Netherlands, and (e-f) São Paulo, Brazil.

D. PARAMETER TRANSFERABILITY

Consistent ground-truth data are not always available for those regions where one can collect sparse mobility data. If the good performance of the proposed model largely relies on external data sources to calibrate its parameters, its application is limited. Can we use a set of parameters learned from one region's ground-truth data to another without compromising the performance too much? To answer this question, we test how transferable the calibrated parameters are from one region to another. To do so, for each of the three regions, we run the model to synthesise mobility data

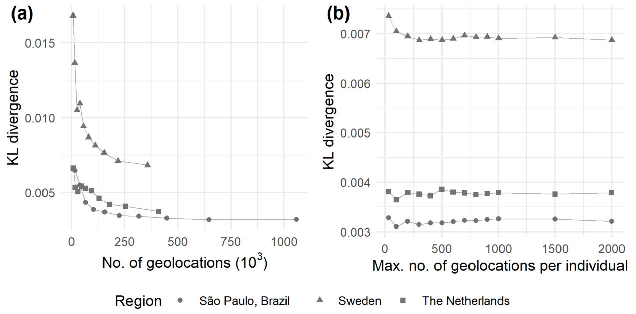
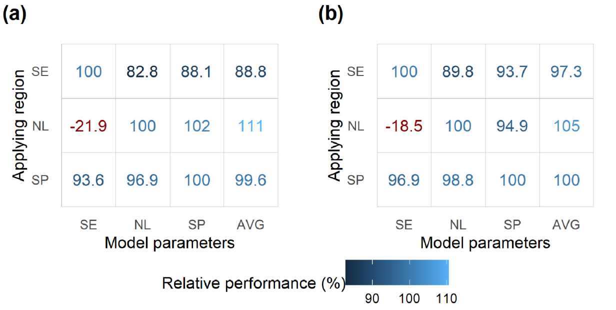


FIGURE 9. Model performance as a function of data length (against the calibration dataset). (a) Total number of sparse geolocations, from individuals in the order of the ones with least geolocations to the ones with most. (b) Maximum number of geolocations per individual. Parameters are optimal for each region. The smaller the KL divergence, the better the model performs.



(a)

Applying region	SE	NL	SP	AVG
SE	100	82.8	88.1	88.8
NL	-21.9	100	102	111
SP	93.6	96.9	100	99.6

(b)

Applying region	SE	NL	SP	AVG
SE	100	89.8	93.7	97.3
NL	-18.5	100	94.9	105
SP	96.9	98.8	100	100

FIGURE 10. Comparison of the relative performance of each region using its calibrated model parameters (diagonal scores), against model parameters calibrated for the other regions. SE = Sweden, NL = The Netherlands, SP = São Paulo, and AVG = average model parameters. (a) Calibration dataset. (b) Validation dataset.

with the calibrated parameters of the other two regions and with the average value of the parameters of all three regions, and compare the results with their ground truth results. The performance gain is calculated as the relative decrease of KL divergence of the model as compared with the benchmark in %. A negative value of the performance gain indicates that the proposed model performs worse than the benchmark model. The relative performance, i.e., how good the model parameters of one region are to another, is quantified by the ratio of the performance gain (applying region / ego region). For the region results using its own model parameters, this relative performance is 100%.

Figure 10 shows the results of the test of how transferable the calibrated model parameters are from one region to another. Except for the use of Sweden's parameters on the Netherlands, we observe only a small variation in relative performance. This indicates that the model performance is not very sensitive to the change of the parameters' values given a certain level of knowledge. And it is promising for reaching a good performance when using the calibrated parameters in other regions with similar sparse data. It is worth noting that, in some cases, we have a relative performance above 100%, which means that some other regions' model parameters are better than the ones found for one region. This is due to the fact that the Bayesian optimisation approximates the optimal parameters.

V. DISCUSSION AND CONCLUSION

This study proposes a model that fills the gaps in sparse mobility traces. The synthesised mobility data can be used for quantifying travel demand in terms of population flows and trip distance distributions. The proposed model extends the fundamental mechanisms of exploration and preferential return to synthesise mobility [8], and is tested on sparse individual traces found in geolocated social media data.

The proposed model generally performs better than the benchmark (heuristic) model in terms of quantifying population flows and trip distance distribution. Compared with the other methods addressing sparsity issues, the proposed model has a few advantages. First, instead of trajectory reconstruction which risks the invasion of privacy, our model estimates travel demand based on collective travel patterns. Second, it is based on fundamental mechanisms of human mobility, expressed in a simpler form than in previous studies [7]. Third, based on real-world data that are very sparse (around 1.5 geolocations per day covering 100 days on average), the proposed model shows good performance. This level of sparsity is higher than previous studies using CDRs [7], [25], [39].

A. MODEL DESIGN FOR SPARSE TRACES

Sparse mobility traces are often collected passively, only when the phone users are engaged in certain phone activities: making a call, messaging, tweeting with a geotag, or using location-aware applications. Hence, these geolocations are incomplete and sparse observations of the individuals' mobility. In a previous study of sparse geolocations from Twitter, we found that the long-term observation of individual geolocations captures both routine mobility and occasional exploration to new places [40], despite the proportion of regular locations to uncommon places deviating from the users' actual mobility [29]. Therefore, we follow the assumption that the rank order of places, based on their visitation frequency from sparse traces, are preserved [21].

According to the literature, we make two designs in the model accounting for the sparsity issue. First, we use the visitation frequency obtained from the Zipf's law when designing the probability function for returning to an old place (Section II-C.1), instead of the a visitation frequency directly calculated from the sparse input. In doing so, we attempt to exclude the bias of overly representing uncommon places in the sparse geolocations. Second, we create a two-dimensional collection of jump size (trip distance) and bearing for exploring a new place, instead of replicating the biased displacements in the sparse traces (Section II-C.2). This distribution is shaped by the individual's returning and exploring behaviour observed in the sparse input, and the visits to new places are constrained by where the individual lives and stays most of the time. The second design is similar to a study that introduces the heterogeneity of visiting directions [18] to the individual mobility model [8]. The difference is that we consider this directional preference at the individual level. In contrast, they consider how a large group of people influence each other, i.e., people tend to visit places that are frequently visited by others based on their empirical findings [18]. The integrated heterogeneity of visiting directions provides more spatial details. By these two designs, the proposed model synthesises the sparse traces into more representative mobility data.

The proposed model protects personal data and privacy by 1) clustering raw geolocations for identifying the home regions (see Appendix B) and 2) not reconstructing individual mobility trajectories that could potentially reveal the precise movement of each data contributor; instead, it creates *synthetic* mobility data from sparse inputs. The objective of the proposed model is to fill the gaps in sparse traces so that the synthesised mobility data are more representative of average daily visits and total distance travelled for further aggregation. Apart from constructing ODMs, we can also develop activity-based models driven by the model-synthesised data for simulating individuals' daily activities. Using these synthetic data from easy-to-access geolocation big data, we can provide more timely and realistic trip data than traditional data-driven approaches [41].

B. MODEL PERFORMANCE

We use the available travel survey data to calibrate the customisable parameters of the proposed model for Sweden, the Netherlands, and São Paulo (Section III-B). It is worth noting that the model is designed in such a way that if there are "ground-truth" trajectories, the model can be calibrated against these data. In reality, it is difficult to access high quality ground-truth data, which often are either non-existence or outdated. Therefore, in this study, we calibrate the model against population-level data for three selected regions. The difference between the results using the calibration and the validation sample is small (Table 2).

There are regional differences between the model outputs for the three regions. Overall, the model for São Paulo performs better than the ones for Sweden and the Netherlands. One reason for this relates to how the individuals' home locations are distributed across the study area. Previous studies have suggested that most active Twitter users live in urban areas [40], [42] and that using sparse geolocations of Twitter data for simulating travel demand is more suitable for urban residents than for the population as a whole. Another reason is that São Paulo has the smallest area but the greatest number of individuals and geolocations in the sparse traces. Given the impact of data length on the model performance (Figure 9), abundant data may contribute to its best performance among the three study areas. The same reasoning may explain the less ideal model performance in Sweden, where its performance may be further improved by covering a larger population (Figure 9a). The other reason may be due to the effect of the modifiable areal unit problem (MAUP), a phenomenon where spatial results vary depending on how the study area is divided into smaller analysis units [43], [44], [45]. We could not use a consistent gridding system to compare the model performance due to the predefined region-specific spatial zones of the ground-truth data. Therefore, the origins and destinations of trips are aggregated to different spatial zones for the three regions. With more precise ground-truth data, the model can be further investigated using a uniform gridding system to exclude the MAUP effect.

Based on the results of the parameter search, we observe that there is a large parameter space where the model performance is quite robust to a moderate range of values for the three parameters (Figure 4). Our results suggest that the parameters calibrated for one region are transferable to another (Figure 10), except for using Sweden's parameters on the Netherlands. The exception may be due to the distinct geographical scales of these two countries and the MAUP issue. We need more in-depth analysis and a broader model test in different regions to understand the reasons better. In general, the proposed model with the parameters' average values has the potential to be applied to the other regions in the absence of ground-truth data.

C. LIMITATIONS AND FUTURE WORK

The proposed model for filling the gaps in sparse mobility traces has some limitations. (1) The proposed model synthesise mobility data by filling in the data gaps of sparse individual traces. However, due to the lack of matching individuals, our validation data represent the aggregated pictures of population flows and trip distances from daily trips. More steps can be taken to address the inherent inconsistency between the proposed individual-based model and the calibration to the population data. One future direction is to test the performance of the proposed model using high-resolution GPS data: with a more complete set of mobility trajectories, we can simulate a variety of sparsity levels by downsampling the observed locations and evaluate the impact of sparsity on the model's performance. (2) The model can be extended in future studies by integrating spatial context and temporal dimensions [7] to account for the types of activities based on the semantic context of historical trips. These improvements can make the synthesised mobility data more useful in transport planning. (3) The model simulates daily trips that always return to home therefore, an important aspect of mobility, overnight trips, is yet to be integrated for future improvements. (4) We use geolocations from Twitter as an example of sparse traces. However, Twitter has recently changed its policy, making the geolocations less precise [46]. Despite using the data collected before this significant change, future work will need to test the feasibility of the proposed model by using more sources of sparse traces such as mobile application data from more regions.

Code is available at https://github.com/TheYuanLiao/individual_mobility_model.

APPENDIX A NOTATIONS

The main symbols used in this manuscript and their definitions are briefly summarised in Table 4.

TABLE 4. Notation table. Main symbols and their definitions.

Notation	Definition
(X, Y)	Decimal coordinates of a location
$\mathbf{Trac}$	Sparse input of visits to locations for a given individual
$\mathbf{Trac}'$	Synthesised visits to locations for a given individual
N	No. of observed locations in the sparse input
D	No. of simulation days
M_{day}	No. of visits to locations per simulation day (day)
$\mathbf{S}$	Set of all the distinct locations in $\mathbf{Trac}$
n	No. of distinct locations in $\mathbf{S}$
s_h	Individual's home location
θ_p	Euclidean distance between s_p and s_{p+1}
α_p	Bearing angle between s_p and s_{p+1}
$\mathbf{J}$	Set of $(\theta, \alpha)_p$ in $\mathbf{Trac}$
s_p	Individual's current location
s_{p+1}	Individual's next location
γ, ρ	Two parameters that control exploration
$\text{Prob}(s_{p+1} \mid s_p)$	Probability of returning to s_{p+1} given s_p
k_s	Rank order of location s by its visiting frequency
ζ	The parameter of Zipf's Law
β	The parameter that controls returning impedance
d	Trip distance (Euclidean), $d \in d_{group}$
$P\left(d\right)$	Frequency rate of d
$p\left(d\right)$	Empirical probability density of d
$\hat{p}\left(d\right)$	Theoretical probability density of d fitted to a model

APPENDIX B DATA DESCRIPTION OF APPLIED SPARSE TRACES

Geotagged tweets are a typical source of sparse mobility traces. Twitter users can choose to geotag tweets, in which case the social media data include geolocation information. One can collect tweets from the Twitter User Timeline API to get a maximum of 3,200 tweets from a Twitter user's history, where a (small) portion of these are geotagged. We purchased data from a Twitter subsidiary, Gnip, to get a complete archive of geotagged tweets from a six-month period (20 Dec 2015 – 20 Jun 2016), generated within the study areas: Sweden, the Netherlands, and São Paulo, Brazil. Using this Gnip dataset, we identified the top geotag users, i.e., those who generated at least 30 geotagged tweets during the data collection period. For the model experiment, we collected their user timelines to get their historical geotagged tweets.

Before these geotagged tweets can be used, we carefully preprocess them to reduce artefacts [20]. We remove: 1) Users who only geotag tweets of a single place, on suspicion of bot accounts, e.g., for job posting or weather updates. 2) Tweets for which the Twitter user posts a place's location, e.g., the centre of a country, instead of the tweet's precise GPS coordinates. 3) Those top geotag Twitter users who nevertheless have insufficiently many (< 20) geotagged tweets. 4) Tweets from before an apparent move to a study region. To protect privacy, we further cluster raw geolocations using DBSCAN so that the identified home location refers to an area [47] instead of a precise point on the map. The distance threshold for merging is set as 0.1 km. The minimum number of location for a region is set as 1.

The geotagged tweets after the above preprocessing are summarised in Table 5. The sparsity of the data is observed in all three regions, given that the number of geolocations per day ranges from 1.4 to 3.2, with all having fewer than two locations, which is far lower than the typical number of visits per day such as 3.1 for Sweden. This makes it challenging to directly use these sparse traces to adequately model travel demand [20].

TABLE 5. Statistics of the sparse traces from Twitter covering the time span of 2010 – 2019.

Region	São Paulo, Brazil	The Netherlands	Sweden
Population (million)	12.2	17.3	10.2
GDP/capita (kUSD/yr, nominal)	27.1	53	54.6
No. of users	10,943	5,375	3,961
No. of geolocations	3,513,796	1,479,674	1,248,158
No. of days covered /user[a]	96	100	111
No. of geolocations/day /user[a]	1.5	1.4	1.4
Area (10^3 km^2)	1.5	42	450

[a] Median value of all Twitter users.

Source: GDP – The World Bank (2020), Population – UNEP (2020).

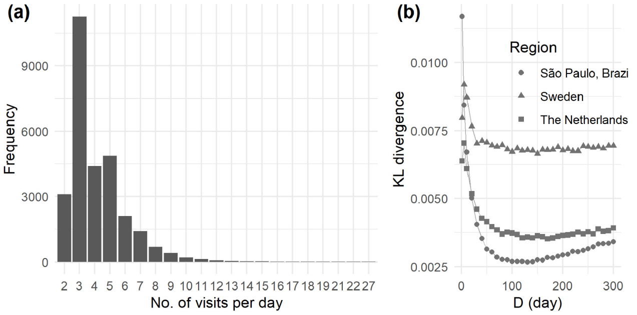


FIGURE 11. Value settings of the model parameters M_{day} and D. (a) M_{day}. Empirical distribution of the number of visits per day derived from the Swedish National Travel Survey [27]. (b) D. Relationship between the model performance (KL divergence compared with ground truth) and the value of parameter D. The other parameters are optimal for each region. The smaller the KL divergence, the better the model performs.

APPENDIX C DETERMINATION OF MODEL PARAMETERS M_{DAY} AND D

The model parameter M_{day} determines how many visits to generate for each simulation day, which can be empirically informed. In this study, we use the Swedish National Travel Survey to get the distribution (Figure 11a) that the model draws M_{day} from. The model parameter D decides how many simulation days to generate data, which can be determined by experiments. In this study, we compare the model-synthesised ODMs with the ground truth (quantified by KL divergence) using varying values of D. Due to the stochasticity of the model output, we need more than one simulation day to achieve stable results of individual mobility trajectories. Figure 11b suggests a stabilised KL divergence after 260 days for all three regions. To balance the model performance and computation efficiency, we set D to 260.

APPENDIX D POPULATION FLOWS USING VALIDATION DATASET

Figure 12 shows the trip frequency rate between zones from the validation results, comparing the model output with the benchmark output. The overall trend is consistent with the results from the calibration dataset as shown in Figure 6.

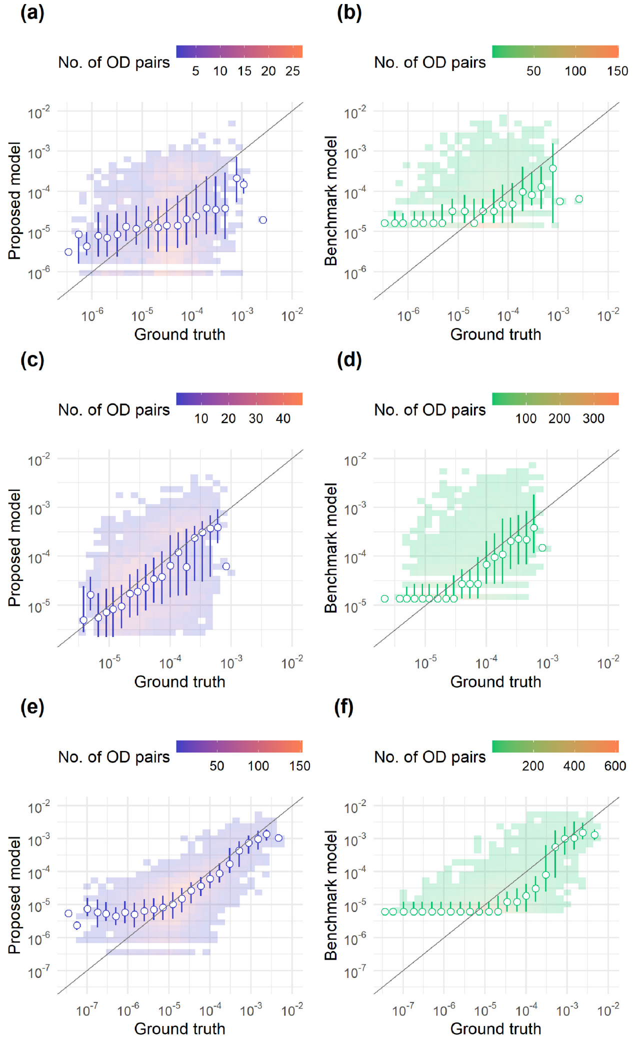


FIGURE 12. Modelling the trip frequency rate between zones (validation results). The gray diagonal line corresponds to a perfect agreement between the ground-truth data and the model/benchmark output. Heat maps of point counts show the distributions of No. of OD pairs. Circles are median values for each bin and lines are the 0.25-0.75 quantiles. (a-b) Sweden, model and benchmark. (c-d) The Netherlands, model and benchmark. (e-f) São Paulo, Brazil, model and benchmark.

REFERENCES

[1] (IEA, Paris, France). "Tracking transport 2020." 2020. [Online]. Available: https://www.iea.org/reports/tracking-transport-2020

[2] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Apr. 2011.

[3] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González, "The TimeGeo modeling framework for urban mobility without travel surveys," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 37, pp. E5370–E5378, 2016.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[5] L. Alessandretti, U. Aslak, and S. Lehmann, "The scales of human mobility," *Nature*, vol. 587, no. 7834, pp. 402–407, 2020.

[6] H. Barbosa *et al.*, "Human mobility: Models and applications," *Phys. Rep.*, vol. 734, pp. 1–74, Mar. 2018.

[7] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete trajectory reconstruction from sparse mobile phone data," *EPJ Data Sci.*, vol. 8, no. 1, p. 30, 2019.

[8] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nat. Phys.*, vol. 6, no. 10, p. 818, 2010.

[9] M. G. McNally, "The four step model," Inst. Transp. Stud., Univ. California at Irvine, Irvine, CA, USA, Rep. UCI-ITS-AS-WP-07-2, 2000.

[10] K. W. Axhausen and T. Gärling, "Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems," *Transp. Rev.*, vol. 12, no. 4, pp. 323–341, 1992.

[11] S. Hörl and M. Balac, "Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data," *Transp. Res. C Emerg. Technol.*, vol. 130, Sep. 2021, Art. no. 103291.

[12] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources," *Transp. Res. C Emerg. Technol.*, vol. 58, pp. 162–177, Sep. 2015.

[13] G. Mattioli and M. Adeel, "Long-distance travel," in *International Encyclopedia of Transportation*, R. Vickerman, Ed. Oxford, U.K.: Elsevier, 2021, pp. 272–277. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081026717106955

[14] Q. Ge and D. Fukuda, "Updating origin—Destination matrices with aggregated data of GPS traces," *Transp. Res. C Emerg. Technol.*, vol. 69, pp. 291–312, Aug. 2016.

[15] Y. Yue, T. Lan, A. G. Yeh, and Q.-Q. Li, "Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies," *Travel Behav. Soc.*, vol. 1, no. 2, pp. 69–78, 2014.

[16] K. H. Grantz *et al.*, "The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology," *Nat. Commun.*, vol. 11, no. 1, pp. 1–8, 2020.

[17] Y. Liao, "Understanding human mobility with emerging data sources: Validation, spatiotemporal patterns, and transport modal disparity," Dept. Space Earth Environ., Chalmers Univ. Technol., Gothenburg, Sweden, Rep. TR-2020, 2020.

[18] M. Schläpfer *et al.*, "The universal visitation law of human mobility," *Nature*, vol. 593, no. 7860, pp. 522–527, 2021.

[19] Q. Wang, N. E. Phillips, M. L. Small, and R. J. Sampson, "Urban mobility and neighborhood isolation in America's 50 largest cities," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 30, pp. 7735–7740, 2018.

[20] Y. Liao, S. Yeh, and J. Gil, "Feasibility of estimating travel demand using geolocations of social media data," *Transportation*, vol. 49, no. 1, pp. 137–161, 2022.

[21] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Enriching sparse mobility information in call detail records," *Comput. Commun.*, vol. 122, pp. 44–58, Jun. 2018.

[22] J. H. Lee, A. Davis, E. McBride, and K. G. Goulias, "Statewide comparison of origin-destination matrices between california travel model and Twitter," in *Mobility Patterns, Big Data and Transport Analytics*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 201–228.

[23] H.-H. Jo, M. Karsai, J. Karikoski, and K. Kaski, "Spatiotemporal correlations of handset-based service usages," *EPJ Data Sci.*, vol. 1, no. 1, p. 10, 2012.

[24] A. Kheiri, F. Karimipour, and M. Forghani, "Intra-urban movement flow estimation using location based social networking data," in *Proc. Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 40, 2015, pp. 781–785.

[25] O. Burkhard, R. Ahas, E. Saluveer, and R. Weibel, "Extracting regular mobility patterns from sparse CDR data without a priori assumptions," *J. Location Services*, vol. 11, no. 2, pp. 78–97, 2017.

[26] C. Anda, A. Erath, and P. J. Fourie, "Transport modelling in the age of big data," *Int. J. Urban Sci.*, vol. 21, no. s1, pp. 19–42, 2017.

[27] Official Statistics of Sweden. "Swedish National Travel Survey (RVU Sweden) 2011—2016." 2016. [Online]. Available: https://www.trafa.se/en/travel-survey/travel-survey/

[28] J. Osorio-Arjona and J. C. García-Palomares, "Social media and urban mobility: Using Twitter to calculate home-work travel matrices," *Cities*, vol. 89, pp. 268–280, Jun. 2019.

[29] D. Tasse, Z. Liu, A. Sciuto, and J. I. Hong, "State of the Geotags: Motivations and recent changes," in *Proc. ICWSM*, 2017, pp. 250–259.

[30] A. Rodriguez-Carrion, C. Garcia-Rubio, and C. Campo, "Detecting and reducing biases in cellular-based mobility data sets," *Entropy*, vol. 20, no. 10, p. 736, 2018.

[31] S. Gao, J.-A. Yang, B. Yan, Y. Hu, K. Janowicz, and G. McKenzie, "Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area," in *Proc. 8th Int. Conf. Geograph. Inf. Sci. (GIScience)*, 2014, pp. 2–9.

[32] J.-X. Wang, J. Huang, L. Duan, and H. Xiao, "Prediction of reynolds stresses in high-mach-number turbulent boundary layers using physics-informed machine learning," *Theor. Comput. Fluid Dyn.*, vol. 33, no. 1, pp. 1–19, 2019.

[33] K. Smolak, W. Rohm, K. Knop, and K. Siła-Nowicka, "Population mobility modelling for mobility data simulation," *Comput. Environ. Urban Syst.*, vol. 84, Nov. 2020, Art. no. 101526.

[34] Statistics Netherlands. "Onderzoek Verplaatsingen in Nederland (OViN) 2017." 2018. [Online]. Available: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingen-in-nederland--ovin--

[35] S. P. G. Do Estado. "Research Origin and Destination 2017." 2017. [Online]. Available: http://www.metro.sp.gov.br/pesquisa-od/

[36] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[37] F. Nogueira. "Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python." 2014. [Online]. Available: https://github.com/fmfn/BayesianOptimization

[38] A. P. Masucci, J. Serras, A. Johansson, and M. Batty, "Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 88, no. 2, 2013, Art. no. 022812.

[39] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Netw.*, vol. 64, pp. 296–307, Apr. 2014.

[40] Y. Liao, S. Yeh, and G. S. Jeuken, "From individual to collective behaviours: Exploring population heterogeneity of human mobility based on social media data," *EPJ Data Sci.*, vol. 8, no. 1, p. 34, 2019.

[41] J. Drchal, M. Čertickỳ, and M. Jakob, "Data-driven activity scheduler for agent-based mobility models," *Transp. Res. C Emerg. Technol.*, vol. 98, pp. 370–390, Jan. 2019.

[42] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users," in *Proc. 5th Int. AAAI Conf. Soc. Media*, 2011, pp. 554–557. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234

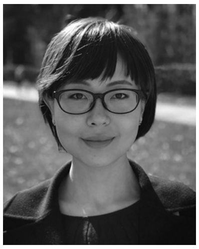
[43] S. Openshaw, *The Modifiable Areal Unit Problem*, vol. 38. Norwich, U.K.: Geo Books, 1983.

[44] A. S. Fotheringham and D. W. Wong, "The modifiable areal unit problem in multivariate statistical analysis," *Environ. Plan. A*, vol. 23, no. 7, pp. 1025–1044, 1991.

[45] D. W. Wong, "The modifiable areal unit problem (MAUP)," in *WorldMinds: Geographical Perspectives on 100 Problems*. Dordrecht, The Netherlands: Springer, 2004, pp. 571–575.

[46] Y. Hu and R.-Q. Wang, "Understanding the removal of precise geotagging in tweets," *Nat. Human Behav.*, vol. 4, no. 12, pp. 1219–1221, 2020.

[47] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Kdd*, vol. 96, 1996, pp. 226–231.



YUAN LIAO received the B.Sc. and M.Sc. degrees in automotive engineering from Tsinghua University, China, in 2013 and 2016, respectively, and the Ph.D. degree in energy and environment from the Chalmers University of Technology, Sweden in 2021, where she is currently a Postdoctoral Researcher working on big data, agent-based modeling, and synthetic population for future urban mobility, such as electromobility infrastructure planning. Her research vision is the data-driven understanding of human mobility to empower the sustainable transformation of global cities. The vision highlights using big and continuous data from newly emerged sources, such as twitter, mobile phones, and other online platforms. Her active research interests include mobility data science, urban big data, GIS, and sustainable transport.

KRISTOFFER EK received the M.Sc. degree in software engineering from the Chalmers University of Technology Sweden in 2020. He is currently a Software Engineer with Burt Intelligence, Sweden, working with big data and automated insights for the advertising industry, helping leading media brands make data-driven decisions.

ERIC WENNERBERG received the M.Sc. degree in software engineering from the Chalmers University of Technology Sweden in 2020. He is currently a Software Engineer with Einride, Sweden, working with vehicle routing and optimization of heavy electric trucks.

SONIA YEH is a Professor of Transport and Energy Systems with the Department of Space, Earth and Environment, Chalmers University of Technology Sweden and the Vice Director of Area of Advance Energy with Chalmers. She is an Adjunct Professor with the Department of Engineering and Public Policy, Carnegie Mellon University. Her expertise is in energy economics and energy system modeling, alternative transportation fuels, sustainability standards, technological change, and consumer behavior and mobility. Throughout her work, she has advised and worked broadly with U.S. state and international advisers, policymakers, a wide range of stakeholder groups and academic researchers in developing climate policies toward reducing the environmental impacts and GHG emissions from transport. She served as the Fulbright Distinguished Chair Professor of Alternative Energy Technology from 2016 to 2017, and received the Håkan Frisinger Award by Volvo Research and Educational Foundations in 2019. She has been a Senior Editor for *Energy Policy* since 2018.

JORGE GIL received the M.Sc. degree in virtual environments from UCL, Bartlett, in 2000, and the Ph.D. degree in urbanism from TU Delft, in 2016. He is an Associate Professor of Urban Analytics and Informatics with the Department of Architecture and Civil Engineering, Chalmers University of Technology, Sweden. His primary research focus is on the conceptualisation and development of integrated urban models for digital urban planning research and practice, in the context of Smart Cities, City Information Modelling, and Urban Digital Twins. These are applied in the domains of sustainable mobility of people and goods, social inclusion, liveability, energy transition, and circular economy. He develops GIS and geo-database solutions with a user centred approach, including spatial analysis, network analysis, machine learning and visualisation methods. The research leverages the potential of open data and open source adopting an open science approach.