CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Estimating Travel Demand from Twitter using an Individual Mobility Model

In Sweden, The Netherlands and São Paulo

Master's thesis in Computer science and engineering

KRISTOFFER EK
ERIC WENNERBERG

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2020

Master's thesis 2020

Estimating Travel Demand from Twitter using an Individual Mobility Model

In Sweden, The Netherlands and São Paulo

KRISTOFFER EK
ERIC WENNERBERG

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2020

Estimating Travel Demand from Twitter
using an Individual Mobility Model

In Sweden, The Netherlands and São Paulo

KRISTOFFER EK
ERIC WENNERBERG

Supervisor: Sonia Yeh, Department of Space, Earth and Environment
Advisor: Yuan Liao, Department of Space, Earth and Environment
Examiner: Carl Seger, Department of Computer Science and Engineering

Master's Thesis 2020
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2020

Estimating Travel Demand from Twitter
using an Individual Mobility Model

KRISTOFFER EK
ERIC WENNERBERG
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

The cost of conducting household travel surveys is increasing, while the response rate is decreasing, pushing researchers to explore new sources of data that can be used to estimate travel demand. Among these new data sources is geotagged tweets from Twitter due to its large quantity of available data and low cost of access. At the same time, using Twitter for travel demand estimation has garnered criticism regarding the biases inherent in Twitter data. This thesis uses geotagged tweets from three regions: Sweden, the Netherlands and São Paulo, to quantify the bias in Twitter data and develop a novel model that estimates travel demand by de-biasing the raw Twitter data. The model integrates two natural dimensions of individual mobility: regularly returning to habitual locations and occasionally exploring new locations. The proposed model addresses the under-representation of habitual places such as home and workplace and corrects the geotagging behavioural bias of overly representing long-distance travel. The model is validated against external data sources in each of the three regions and it is found to result in significant improvements over contemporary methods for using Twitter data for travel demand estimation. The model's parameters are robust across regions studied, and by using the parameters found in this thesis one can expect the same improvements compared to contemporary approaches when applied to other regions.

Keywords: human mobility, travel demand estimation, Twitter, individual mobility model.

Acknowledgements

We would like to thank our supervisors, Sonia Yeh and Yuan Liao, for their tireless support and guidance.

Kristoffer Ek & Eric Wennerberg, Gothenburg, June 2020

Contents

List of Figures

List of Tables

1

Introduction

People visit different places to participate in a variety of activities every day, and the series of these places form a trajectory. Aggregating the whole population' trajectories reveals the flows of the population between places. Travel demand, as quantified by such flows, is vital for making informed policies in transport and other areas such as urban planning, public health, and greenhouse gas (GHG) mitigation. In understanding how people move, empirical data with good quality plays an important role. To date, daily and short-distance travel have been extensively studied by transportation and geographic researchers using traditional household travel surveys. However, the costs of conducting these surveys are increasing, response rate decreasing to an alarmingly low rate, and are typically conducted every 5-10 years, if at all, meaning that it is hard to keep up-to-date [1].

The increased availability of humans' spatiotemporal records via various social media platforms have provided researchers with new data sources for estimating travel demand. Among these sources, Twitter is especially appealing due to its low cost and easy access to a significant volume of mobility traces. Previous studies have used data from Twitter to answer important mobility questions, for example how geographical features and cultural norms affect long-distance travel, how people move within cities, and how people can be clustered based on their mobility patterns.

The main criticism, however, of using Twitter to measure the mobility pattern of the general population pertains to three aspects: (1) Twitter users do not necessarily represent the whole population, (2) The users' incentives for using Twitter, such as showing off being at unusual places, might skew the observed mobility, (3) Twitter data lack of regular, and often sparse, sampling.

So far, there is not a good way to address these criticisms other than to acknowledge that the issues exist, but being able to address them is essential to bringing Twitter into play for travel demand estimation in a more rigorous manner. This thesis examines Criticism (1) and creates individual mobility model to address Criticism (2)-(3), and by doing so, advance the case of using Twitter for mobility studies.

The rest of this chapter reviews the related work on common data sources and models to represent mobility and how the above criticisms are considered in the modelling. Finally, the thesis objectives and ethical considerations conclude the chapter.

1.1 Related work

There are a few data sources that have been used for travel demand estimation, such as household travel surveys, GPS log, Call Detailed Record (CDR), and Twitter as an example of social media. This section reviews the strengths and limitations of Twitter data as compared with the other sources (Section 1.1.1). In Section 1.1.2, we show how mobility is measured by geotagged tweets in the literature and the corresponding problems. Section 1.1.2 reviews a range of mobility models for characterising human mobility at both individual and population level, where we justify the selection of an individual mobility model to build upon in this thesis to model mobility trajectories for travel demand estimation.

1.1.1 Data sources

In the last decade, new technologies and services have offered an alternative source of human mobility data in addition to the traditional household travel survey. Three categories of unconventional data sources have been widely used: GPS-enabled tracking devices (GPS logs), Call Detail Records (CDR), and geotagged social media. These new data sources have different characteristics that impact the type of research questions suitable to explore.

GPS log data contain time-series data of GPS coordinates indicating individuals' whereabouts. They are commonly collected from a limited number of subjects who willingly carry a GPS tracker. Most studies applying GPS log data collect data from a small group of individuals, typically in a range of 20-500 [2]. GPS devices produce positions that have an accuracy of 10 meters [3]. The major advantage of GPS logs is their high temporal resolution (e.g., every 10 seconds [4]). Therefore, they provide a relatively complete and accurate picture of an individual's movements during the observation period, which usually lasts several days to a few months. However, in comparison to other sources, GPS log data is used infrequently by the broader research community due to small sample size, high cost, and privacy concerns.

Mobile phone CDR is collected by cell towers in a specific area and contain information, timestamps and position, about calls and text messages sent in the vicinity of each tower. While more data is collected by cell-providers, researchers typically only have access to records on calls and texts. CDR is collected long-term with a large number of individuals due to the high penetration rate of cell phone users. CDR is the most frequently used data source today to estimate the mobility patterns of the general population. For example, one study used a one-year-long CDR data set with 3 million individuals tracked to model the fundamental patterns of individual mobility, e.g., the long-tailed distribution of trip distance [5]. Because the position attributed to each record is the position of the closest cell tower, the spatial resolution is directly correlated to the density of the cell tower network, where towers are typically spaced 200-300 meters apart in urban areas and up to 30 km in rural areas. The spatial sparsity in rural areas, and the limitation to a single cell phone provider, restrict the usage of CDR to urban cities, and consequently, short distance travel. Furthermore, locations are only recorded when an individual makes a phone

call or sends a text message, leading to sparse and irregular samples. Lastly, while CDR is often used, it is difficult to access, and the data are anonymised for privacy consideration, resulting in various drawbacks depending on the anonymisation techniques.

Despite some drawbacks, in contrast to both CDR and GPS logs, data from Twitter is easy, cheap to access, and scale-free, e.g., adaptive to both regional and global scales [2]. A tweet in which the user has selected to attach its position is called a *geotagged tweet*. Geotagged tweets contain information useful for transportation research: timestamp, position, and text. In general, only a small number of tweets have position attached, varying between different regions of the world, typically 1-3% [6]. The relatively low number of geotagged tweets could be improved by inferring position from the text content, but the accuracy is low [7]. Because the position of geotagged tweets is based on the GPS-enabled device the user sent the tweet from, it shares the same spatial resolution of GPS-log data, typically 10m. Geotagged tweets are collected when an individual sends a tweet, which leads to the same issues, such as sparsity and irregularity, as CDR. Geotagged tweets enabled observation of a large number of individuals' movement over long periods, while not restricted by geographical and administrative boundaries such as cities and countries.

Previous studies have used geotagged tweets to derive summarised mobility patterns around the world. Hawelka et al. (2014) [8] analysed global mobility patterns observed from one billion tweets and found that geographical features and cultural norms influenced the mobility patterns. For example, individuals in isolated countries, such as Australia and New Zealand, exhibited a relatively larger radius of gyration and individuals from Arabic countries travelled significantly less during the period of Ramadan. Lenormand et al. (2014) [9] compared the commuting-mobility, travel between home and work, during weekdays derived from Twitter data and CDR in Barcelona and Madrid to travel surveys and found a high correlation between the three data sources.

To summarise, though geotagged social media as an emerging data source has been used widely to quantify mobility in the last ten years, careful investigations and validation are still needed to further extend its application in estimating travel demand more robustly for the entire population and across different regions. This thesis aims to address these challenges as we describe in the following sections.

1.1.2 Measuring human mobility with geotagged tweets

Challenges with using Twitter data for travel demand estimation pertains to two aspects: user behaviour and temporal sparsity. User behaviour manifests itself in how Twitter users interact with the platform. Tasse et al. (2017) [10] found that users geotag their tweets in order to show off being at cool places and to keep their friends and family updated. This consequently leads to users geotagging tweets at places they rarely visit, and even being reluctant to geotag at routine places. Further, they find that users generally geotag their tweets at places far from home, with only 46.7% of geotagged tweets originating from the user's home city. Due to

geotagged tweets being the result of a conscious decision [10], they are naturally sparse and irregular in time. The issue with temporal sparsity lies in the observed movement from Twitter data. Travel demand is an aggregation of individuals' trips, the connection between two consecutive stays. Due to the temporal sparsity of geotagged tweets, trips are not directly observed in Twitter data [11], but rather what is observed is more precisely named 'displacements', the connection between two, possibly but not necessarily, consecutive stays.

Most previous studies use geotagged tweets without considering, or over-simplifying these challenges. One approach used in literature to translate displacements to trips is to apply a time threshold between 4 and 24 hours [12, 13]. Displacements with a time interval shorter than the threshold are considered trips, and all other displacements are discarded. Regardless of the exact time threshold used, the amount of available data is massively reduced. Lenormand et al. (2014) [9] bypassed the issue of temporal sparsity by estimating the home and workplace of individuals, and derived commuting-mobility, travel between home and work. While this in some sense solves the temporal sparsity, it does so by falsely relying on the second problem that Twitter users are reluctant to post on habitual places. Therefore, careful modelling of the individual mobility trajectories using Twitter data is essential to use geotagged tweets for travel demand estimation.

1.1.3 Modelling travel demand: from individuals to population

Travel demand estimation manifests at the population level, reflecting the flows between regions. There are two types of models, population-based and individual-based. Population-based models, as the name implies, operate on the entire population at once. In contrast, the individual-based models operate on individuals, and its output can be aggregated to the population level.

One of the most applied population-level models in travel demand estimation is the gravity model [14], which states that the trip number between two places can be determined by the production (e.g., population) and attraction (e.g., workplaces) of both places, and their distance from one another. Despite the widespread use of the gravity model, it has notable limitations such as over-simplification and being data-demanding. Another model that recently has gained attention is called Radiation model, which improves the traditional gravity model [15]. The estimated travel demand has fixed zoning once the model is constructed. Its further application is more constrained than the combination of individual-based modelling and aggregation.

Population-based models often require good-quality data at the individual level. Liao et al. (2020) [11] explore the feasibility of using geotagged displacements from Twitter with the gravity model to estimate travel demand where, however, the individual trajectories are not modelled sufficiently to consider behavioural biases. When using Twitter data for travel demand estimation, the individual trajectories require reasonable de-biasing to be better used to estimate the population flows between regions.

Of the individual-based models, Markovian models are among the typical ones that have been used widely. Cárcamo et al. (2017) [16] used CDR and constructed a Markov model capturing the transitions between antennas based on the entire population. Gambs et al. (2012) [17] described an algorithm for next place prediction based on a Markov mobility model of an individual, called *n-MMC*, which achieved up to 95% accuracy. Approaches similar to Markovian models require observed transitions between locations which, however, are not observable in Twitter data due to the sparsity issue. Therefore, these types of models are not suitable for travel demand estimation based on Twitter data.

Lévy-flight and random walk models have been successfully used to model the mobility of animals, and have also been used to model the individual mobility of humans [18]. Although these models exhibit a striking statistical resemblance to human mobility[19], they are based on the assumption that human movement is random. Song et al. (2010), proposes that human mobility is barely random, but follows reproducible scaling laws [5]. In their *Individual Mobility Model*, they focus primarily on two generic mechanisms, exploration and preferential return, both unique to human mobility. According to the model, an individual's next displacement can either be exploring a new location or returning to a previously visited location.

The *Individual Mobility Model* captures the asymptotic/general mobility of individuals, meaning it is a good fit for estimating travel demand after a proper aggregation. Furthermore, the model does not rely on observed trips between places but instead uses the visitation frequency of places to determine the next displacement. The model reflects the input data, and without modification, it would output the behaviour bias observed in Twitter data. However, given the generality of the model, it holds the potential to be adapted to Twitter data, and allow travel demand estimation to be recovered.

1.2 Thesis objectives

To address the limitations of traditional travel surveys, Twitter data has gained increased interests among emerging data sources in estimating mobility. However, some of the well-known criticisms have continued to be ignored in practice. This thesis examines the biases of geotagged tweets and proposes an individual mobility model for travel demand estimation with the attempt to address some of the biases. The model is applied to several regions globally and validated against other "official" data sources in the field of transport. Specifically, this thesis has four objectives:

- Examine the representativeness of geotagged Twitter users and potential biases.
- Propose an individual mobility model to address the sparsity issue and behavioural biases of using geotagged tweets.
- Calibrate and validate the proposed model against travel surveys and the traffic model output to obtain optimal model parameters. Investigate the

performance of the proposed model compared with the common practice of utilising geotagged tweets.

- Compare the sensitivity of model parameters and discuss the ability to generalising it to the other global regions.

1.3 Disposition of this thesis

The remainder of this thesis consists of five chapters; Methods, Validation, Results, Discussion, and Conclusion. Methods describe the data and how they are processed and the proposed model. Validation describes the data sets that are validated against, and also how the validation is carried out. Results show the effects of applying the validation to the model. Discussion deliberates the main findings, limitations, and future work. At last, Conclusion summarises this thesis.

1.4 Ethical considerations

The available data will contain users' Twitter handles, a unique identifier for the user, meaning that the derived travel behaviour can be tied to a specific user. To combat this, the data will be anonymised, i.e. the Twitter handles will be replaced with pseudonymised strings, to remove the direct connection between the data and the underlying Twitter user. To prevent "reverse engineering", we do not publish individual trajectories and their locations. Furthermore, users have opted in to share their location, meaning that some form of consent is already given. The users have however not consented for their data to be part of this study, but this is, in part, mitigated by the fact that the data is publicly available on their profiles, meaning anyone could potentially retrieve it.

2

Methods

This chapter describes the methods applied in this thesis. Section 2.1 describes the dataset in terms of collection and preprocessing. Section 2.2 describes the features constructed from the dataset. Section 2.3 introduces the model of individual mobility and how to deal with the biases found in Twitter data.

2.1 Data collection and preprocessing

The Twitter data used in this thesis have been collected and extracted by a previous study [20]. The data set consists of 23 regions, both countries and cities, and was collected in two stages. In the first stage, tweets during six months (20 December 2015 - 20 June 2016) were collected using a geographical bounding box containing the region. From these tweets, the users who geotag their tweets most frequently was identified. In the second stage, the 3200 most recent tweets of the identified users were retrieved. Not all users have a total of 3200 tweets, and some tweet more frequently than others, such that the total number of tweets and timespan varies from user to user. In total, there are between 30 and 65 million tweets for each region.

To provide more detailed validation, this thesis focuses on the following regions: Sweden, São Paulo and the Netherlands. Furthermore, only geotagged tweets are of interest, and all regular tweets are removed from the data set. Table 2.1 shows the number of tweets and geotagged tweets in each region.

The geotagged tweets that user i have sent are: $(X, Y, t)_{i,p}, p = 1, 2, ..., N_i$, where X

	Before processing			After processing	
Region	**Tweets**	**Individuals**	**Geotweets**	**Individuals**	**Geotweets**
Sweden	31 591 697	7 773	2 943 731	3 961	1 248 158
São Paulo	65 089 103	22 853	8 059 448	10 943	3 513 796
The Netherlands	31 997 687	12 638	4 418 891	5 375	1 479 674

Table 2.1: Summary of data for each region, before and after processing.

is the decimal degree of latitude, Y is the decimal degree of longitude, and t is the local time converted from the original UTC timestamp of the pth geotagged tweet using the location (X, Y). N_i is the total number of geotagged tweets sent by this user. For each geotagged tweet, we also calculate the day of the week and hour of the day based on t. The sequence of the user's geotagged tweets is, therefore:

$$\mathbf{G}_i = \{X, Y, t, w, h\}_{i,p}, p = 1, 2, ..., N_i \tag{2.1}$$

Cross-platform posting

Twitter supports cross-platform posts, meaning that a user can link their account to other platforms and posts shared on one platform will be shared on the other. Cross-posting is problematic for geotagged tweets, as platforms have different spatial accuracy in their reported locations. For example, when a user shares an image on Instagram and tags it with a place, the reported location will be in the centre of that place. So if an individual sets their location on Instagram as Sao Paulo, the geotagged tweet will be reported in the middle of Sao Paulo, as opposed to an exact location. To deal with this artefact, all geotagged tweets are grouped based on their exact latitudinal and longitudinal coordinates, and groups with more than 0.1% of the tweets in the region are removed. This could potentially remove some regular geotagged tweets (not cross-posted), but the likelihood of several geotagged tweets having exact coordinates is minimal.

Clustering locations into places

The spatial resolution of GPS devices is typically within 10 meters [3], meaning that geotagged tweets from the same place can have slightly different coordinates, X and Y. To deal with this problem, each individual's observed locations are grouped to *places* using DBSCAN [21], a clustering algorithm that groups points tightly packed based on density. The algorithm is parameterised with distance threshold, ϵ, and the minimum number of points per cluster, n_{min}. Parameter ϵ controls the geographic size of the clusters but has been found to not be very sensitive for travel demand estimation as the overall patterns emerge despite different values [3]. We use $\epsilon = 100m$ as it prevents a large number of small places being identified, while still separating different places. We use $n_{min} = 1$ in order to not ignore places that have only been visited once.

Let n_i be the number of distinct places for user i, X_j and Y_j be the spatial centroid of the of place j, obtained from DBSCAN. Furthermore, let K denote the number of geotagged tweets at this place. The set of the user's distinct places is, therefore:

$$\mathbf{S}_i = \{X, Y, K\}_{i,j}, j = 1, 2, ..., n_i \tag{2.2}$$

Estimating home place

It is possible that during the first stage of data collection, some top geotag users were only visiting the region, but live somewhere else. As we focus on the residents in

each region, this artefact needs to be removed. Therefore, each user's home location, $s_h \in \mathbf{S}_i$, is detected using the assumption that they live at the most visited location during weekends and 7 pm-8 am on weekdays [22]. Due to Twitter users' reluctance to geotag their tweets at habitual places, including their home, combined with the fact that not everyone has the same circadian rhythm (some working night shifts for example) this method is not perfect. Although the method is not perfect, it is the most common approach for identifying home locations for this type of data. Finally, individuals with an estimated home located outside of the region are removed from the data set.

Considering the long study period, up to 9 years for some users, it is probable that some have moved during the study period, resulting in multiple home locations detected from their timelines. The home location is an essential part of a person's mobility pattern; thus, moving will most likely results in a significant change in their mobility pattern. Hence, as a measure to reduce the complexity of the analysis, we only consider the period where the user lives at his/her latest home location.

Bot accounts and insufficient geotagged tweets

Another artefact of the data set is bot accounts, which, for example, only tweet about job postings and weather updates. Some of these accounts also geotag their tweets, often at the same place. These are identified by selecting users with only one distinct place and subsequently removed.

After the processing to deal with the artefacts in the original data set, we further remove users based on the amount of available data. This is done to ensure mobility patterns can be identified. Users with less than 20 tweets are removed. Table 2.1 presents the data sets, before and after processing.

2.2 Feature construction

For the trajectory of each Twitter user, not all distinct places are visited with the same frequency; some are visited more frequently than others. Thus, a place j has a visitation frequency of being visited, $f_{i,j}$, that is calculated based on the number of geotagged tweets individual i has sent at the place j.

$$f_{i,j} = \frac{K_{i,j}}{\sum_{j=1}^{n_i} K_{i,j}} \tag{2.3}$$

Jump size, $\theta_{p,p+1}$, refers to the distance between two consecutive geotagged tweets p and $p+1$, and is defined as

$$\theta_{p,p+1} = \text{haversine}(X_p, Y_p, X_{p+1}, Y_{p+1}) \tag{2.4}$$

where $\text{haversine}(X_p, Y_p, X_{p+1}, Y_{p+1})$ is the Haversine distance (distance along the curved surface of the earth) between two coordinates.

Bearing, $\alpha_{p,p+1}$, refers to the direction of the straight line one travels between two consecutive geotagged tweets p and $p+1$ and it is defined as

$$\Delta Y_{p,p+1} = Y_{p+1} - Y_p \tag{2.5}$$
$$y_{p,p+1} = \sin(\Delta Y_{p,p+1}) * \cos(X_{p+1}) \tag{2.6}$$
$$x_{p,p+1} = \cos(X_p) * \sin(X_{p+1}) - \sin(X_p) * \cos(X_{p+1}) * \cos(\Delta Y_{p,p+1}) \tag{2.7}$$
$$\alpha_{p,p+1} = \arctan2(y_{p,p+1}, x_{p,p+1}) \tag{2.8}$$

Feature	Description
$\mathbf{S}_i$	Set of distinct places visited by the individual
$s_h \in \mathbf{S}_i$	Estimated home place of the individual
$f_{i,j}$	Visitation frequency of place j, $j = 1, 2, ..., n_i$
$\text{Prob}(\theta)_i$	Jump size distribution
$\text{Prob}(\alpha)_i$	Bearing distribution

Table 2.2: Summary of mobility features constructed for individual i.

2.3 Individual mobility model

As mentioned earlier in Section 1.2, the goal of this thesis is to demonstrate a novel method to estimate travel demand using Twitter data while addressing biases such as under-representation of short-distance trips. In this section, we describe the *Individual Mobility Model*, first proposed by Song et al. (2010) [5], without considering any potential bias in the data source, and our adaptions. We start this section with an overview of the framework in which the model operates, followed by a detailed description of how the steps in the model are carried out.

2.3.1 Framework

The raw trajectory of geotagged tweets constitutes a biased observation of the actual mobility trajectory for a Twitter user. The model developed aims to construct the mobility pattern of *an average week* for individuals. To achieve this, we create a timeline, $\mathbf{L}_i$, for each individual i. The timeline consists of trajectories of multiple days, $\mathbf{L}_i = (\mathbf{T}_{i,d})$, $d = 1, 2, ..., D$. Each daily trajectory, $\mathbf{T}_{i,d}$, consists of multiple visits $\mathbf{T}_{i,d} = (v_{i,d,m})$, $m = 1, 2, ..., M_d$. This is depicted in Figure 2.1.

The number of daily trajectories, D, is set at twenty weeks ($D = 140$). It is selected to be large enough such that we can find a regular pattern for an average week. We found that increasing D, i.e. increasing the number of modelling days, does not change our results. The number of visits generated per day, M_d, is drawn from a normal distribution estimated from travel survey from Sweden[23], $\text{N}(3.14, 1.8)$, and

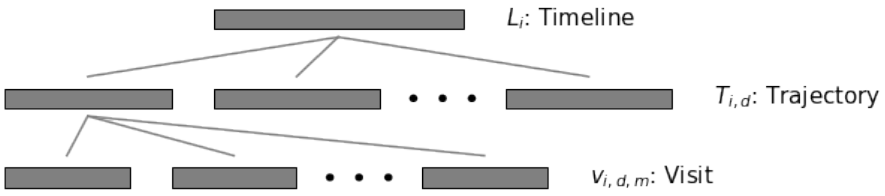

Figure 2.1: Hierarchy of timeline for an individual.

the same distribution is used for all regions. Due to drawing M_d from a distribution, each day has a different number of visits per day.

Each visit $v_{i,d,m}$ consists of latitude and longitude, X and Y, expressed in decimal degrees.

$$v_{i,d,m} = (X, Y) \tag{2.9}$$

For the first visit of each daily trajectory, $v_{i,d,1}$, it is assumed that the individual is located at their estimated home, s_h. This assumption reflects humans' tendency to return home at the end of every day.

To generate the remaining visits for the daily trajectory, it is assumed that the individual can perform one of two choices: *exploration* or *preferential return*. When exploring, the model generates a visit to a place j not observed in the individual's distinct places, $j \notin \mathbf{S}_i$. On the other hand, when returning, the model generates a visit to place j observed in the individual's places, $j \in \mathbf{S}_i$.

The probability of each choice is dependant on the number of distinct places, n_i, derived from the individual's geotagged tweets. The more distinct places, the smaller the probability of exploration is. How much the number of distinct locations influences the probabilities is controlled via two parameters: $0 < \rho \leq 1$ and $0 \leq \gamma$.

$$\text{Prob(explore)}_i = \rho n_i^{-\gamma} \tag{2.10}$$
$$\text{Prob(return)}_i = 1 - \text{Prob(explore)}_i \tag{2.11}$$

The parameters ρ and γ are not specific for an individual but is shared across the population. Figure 2.2 shows their influence on the exploration probability.

From an individuals timeline, a set of trips is constructed by considering each pair of consecutive visits to be the origin and destination of a trip. For the remainder of this chapter, we describe the process of modelling individual mobility in detail with two possible options: exploration and preferential return. The chapter is concluded with an example showing how the model works.

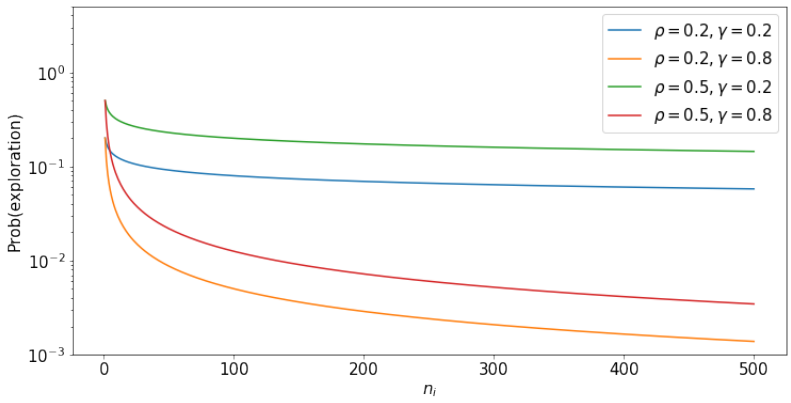

Figure 2.2: Influence of parameters ρ and γ on the exploration probability. n_i is the number of distinct places visited by an individual.

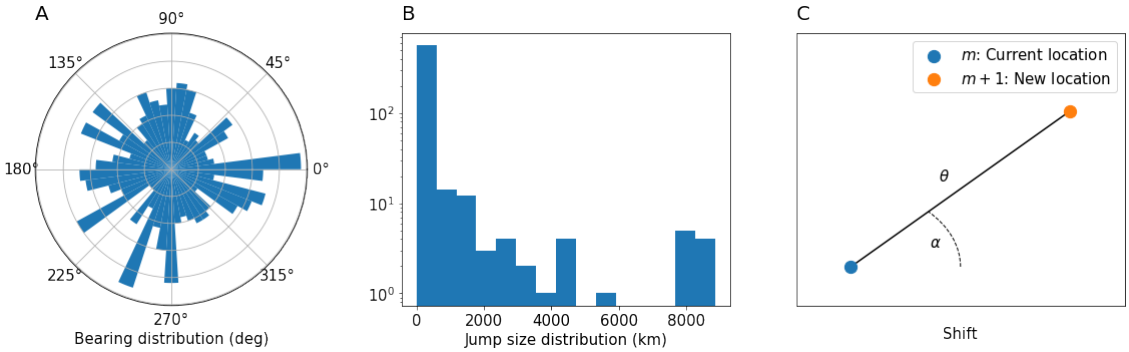

Figure 2.3: **A** Bearing distribution for one individual i. **B** Jump size distribution for one individual i. **C** Visual explanation of the shift function, where θ is drawn from the jump size distribution and α is drawn from the bearing distribution.

2.3.2 Exploration

For day d, let m denote the current place. When exploring, the individual i makes a visit to an unobserved location, $m+1 \notin \mathbf{S}_i$. The new location's coordinates, X_{m+1} and Y_{m+1}, is generated based on the individual's jump size distribution $(\text{Prob}(\theta)_i)$, bearing distribution $(\text{Prob}(\alpha)_i)$, and current location, (X_m, Y_m), as depicted in Figure 2.3.

$$\theta \leftarrow \text{Prob}(\theta)_i \tag{2.12}$$

$$\alpha \leftarrow \text{Prob}(\alpha)_i \tag{2.13}$$

$$X_{m+1}, Y_{m+1} = \text{shift}(X_m, Y_m, \theta, \alpha) \tag{2.14}$$

$$v_{i,d,m+1} = (X_{m+1}, Y_{m+1}) \tag{2.15}$$

2.3.3 Returning

For day d, let m denote the current place. When returning, the individual moves to one of their previously visited places $m+1 \in \mathbf{S}_i$. The selection of place $m+1$ among

all places in $\mathbf{S}_i$ depends on two factors: the visitation frequency of the candidate place $m+1$, and the travel distance from the current place, m, to the candidate place.

Visitation frequency

Previous research on human mobility suggests that the visitation frequency of humans is uneven, such that the frequency z of the kth most visited location follows Zipf's law[5] with parameter $\zeta \approx 1.2 \pm 0.1$.

$$z_k \sim k^{-\zeta} \tag{2.16}$$

Because Twitter users are reluctant to geotag tweets at habitual places, we assume that the observed visitation frequency $f_{i,j}$ is skewed. Habitual places, such as home (s_h) and work, have an observed visitation frequency that is lower than the actual visitation frequency of the place. On the other hand, infrequent place, such as one-time visits to bars, have an observed visitation frequency that is higher than the actual visitation frequency. We assume that, while the observed visitation frequency is skewed, the order of places based on visitation frequency is correct. Therefore we use the below equation to re-scale the visitation frequency of geotagged places where $s \in \mathbf{S}_i$ and $\text{rank}(s)$ denotes the relative order of places by visitation frequency.

$$P(s) = \frac{z_{\text{rank}(s)}}{\sum_{s' \in S_i} z_{\text{rank}(s')}} \tag{2.17}$$

Figure 2.4 shows the effect of re-scaling the visitation frequency as described in Equation 2.17 for one individual. The re-scaling results in habitual places being visited more frequently than they have been observed in the geotagged tweets.

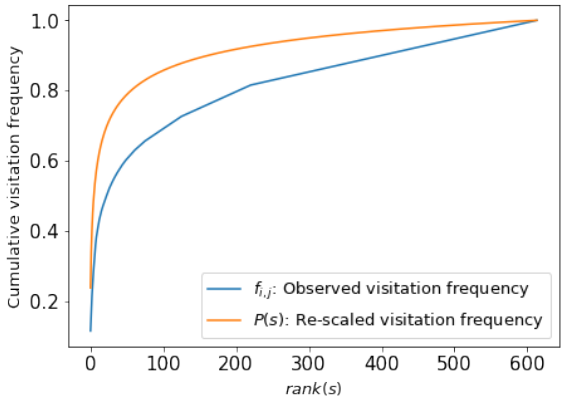

Figure 2.4: Comparison of cumulative distributions of observed visitation frequency, f_j, and re-scaled visitation frequency, $P(s)$, for one individual.

Impedance to the candidate places

The other factor determining the selection of next place, $m+1$, is its distance from the last visit's location, X_m and Y_m. Including distance in the selection helps de-bias the trajectories generated, accounting for the fact that Twitter users are more likely to geotag tweets far from home. The intuition is that we want to slightly increase the probability of visiting places closer to where the individual is currently located. To achieve this, we use an approach similar to that of the Gravity model, by modelling the probability of travelling to a place to be inversely proportional to the distance to it. In other words, the longer the distance one needs to travel from a place to another, the more unlikely the visit happens. Therefore, the impedance between a candidate place s and previous place m is expressed as $\exp(-\beta * \text{haversine}(X_m, Y_m, X_s, Y_s))$ and normalised with the below equation.

$$I(s) = \frac{\exp(-\beta * \text{haversine}(X_m, Y_m, X_s, Y_s))}{\sum_{j=1}^{n_j} \exp(-\beta * \text{haversine}(X_m, Y_m, X_j, Y_j))} \tag{2.18}$$

The strength of preference for short-distance travel, $I(s)$, is controlled by parameter β. Figure 2.5 shows how different values for β yield a stronger och weaker preference.

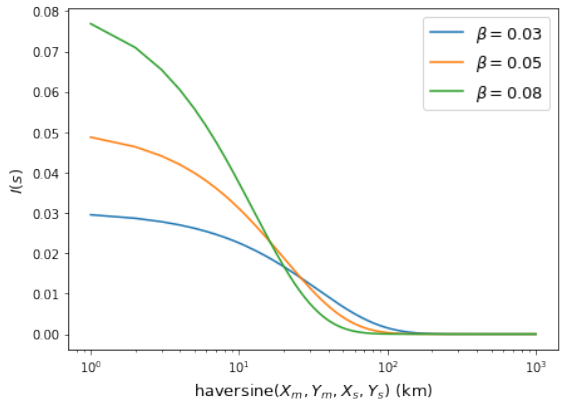

Figure 2.5: Preference for short-distance travel for varying values of β.

Combining visitation frequency and impedance

Both factors, $P(s)$ and $I(s)$, are in the range $[0, 1]$. And they are combined with multiplication and re-normalised to sum to 1. The place the individual moves to for the next visit, $v_{i,d,m+1} = (X_{m+1}, Y_{m+1})$, is drawn from the resulting distribution of the candidate places' probability.

$$\text{Prob}(s) = \frac{P(s) \cdot I(s)}{\sum_{s' \in S_i} P(s') \cdot I(s')} \tag{2.19}$$

$$v_{i,d,m+1} \leftarrow \text{Prob}(s) \tag{2.20}$$

2.3.4 Example

This section show an example of how the model works, by simulating the choices made during one daily trajectory, $\mathbf{T}_{i,d}$. Three different visits are simulated and illustrated in Figure 2.6. The individual in the example have visited three distinct places, and their observed visitation frequency is depicted in the left-most part of the figure, indicated by their size.

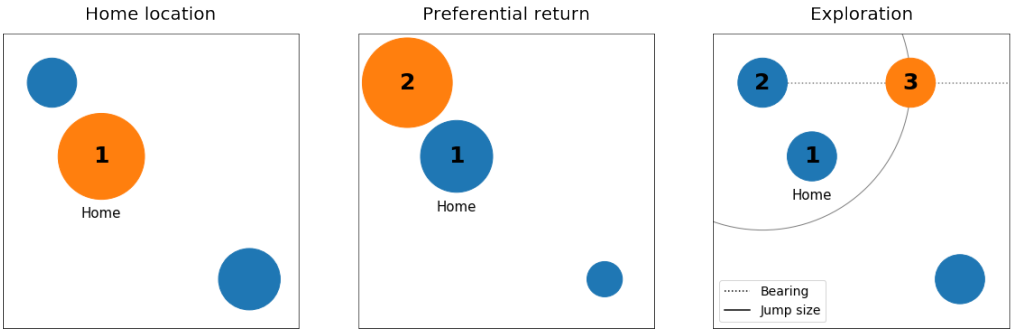

Figure 2.6: Example of the model choices when simulating three visits of a daily trajectory.

The first visit of the daily trajectory is to the individual's home location, s_h. In the example, the second visit is assumed to be preferential return. The figure shows the combination of visitation frequency and impedance, $P(s) * I(s)$, indicated by the size of the circles. Note how the place in the bottom right has lower probability to returned to, due to the distance to the current location. Instead, the top-left place is returned to, marked with 2 in the figure, because of it's proximity to the current location. For the final visit in the example, it is assumed to be exploration, and an unobserved place will be visited based on the bearing and jump size distributions, Prob(θ) and Prob(α). The sampled values from these two distributions are depicted in the figure. The individual will move to where they intersect, marked with 3 in the figure. This process is repeated until the daily trajectory is completed, and then repeated for the remaining daily trajectories of the individual's timeline.

3

Validation

To validate the proposed model, we compare the model outputs with established data sources including travel survey and traffic model output. When validating individual mobility patterns, it is uncommon to have two data sources which include the same group of individuals. Hence, validation is conducted at the population level. Aggregating individual trajectories of mobility yields a picture of how the population flows between regions. One of the most common ways to validate travel demand is to construct and compare Origin-Destination (OD) matrices from different data sources. An OD matrix represents the volume of trips between any two zones in a study area. Using OD matrices, it is possible to compare and analyse two independent travel demand estimations of the same study area.

In this chapter, Section 3.1 presents the external data sources used for validation in each region: Sweden, the Netherlands, and São Paulo. Section 3.2 and Section 3.3 describe the methods used in population representation and trip distance validation, respectively.

3.1 External data sources

This section details four different data sources; three of them are survey-based travel demand estimations, and one is a population distribution estimation.

3.1.1 EU-wide population grid

The GEOSTAT initiative was taken jointly by Eurostat and the National Statistical Institutes to establish a data and production infrastructure for geospatial statistics. The GEOSTAT 2011 dataset [24] represent the main characteristics of the 2011 population and housing census in a 1 km^2 grid system for the entire European Union. This information allows us to identify potential population biases in terms of the spatial distribution of Twitter users' detected home location as compared with the general population in Sweden and the Netherlands.

3.1.2 Sweden

Sampers [25] is a tool owned and managed by the Swedish Transport Administration (Trafikverket), that estimates the historical and future traffic volumes based on

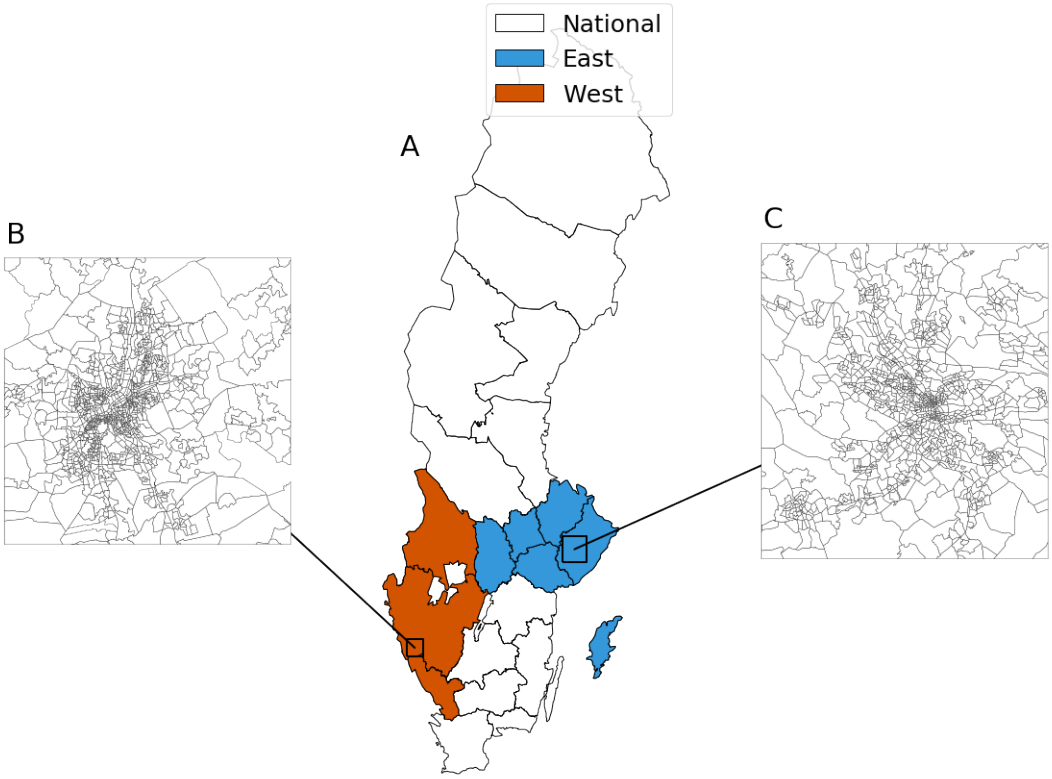

Figure 3.1: (A) A geographical overview of Sweden's national and regional boundaries. (B) Snapshot of zones in West area zoomed in on Gothenburg. (C) Snapshot of zones in East area zoomed in on Stockholm.

studies of travel demand. From the Sampers model, three OD matrices have been retrieved for Sweden from 2014; "National", "East", and "West" (see Figure 3.1). The cell value of the OD matrices represents the estimated number of trips between the origin and destination zone. Although the data is, at time of this thesis, six years old, it is the most representative ground truth for Sweden. The OD matrices represent the domestic travel demand during an average weekday. Each area is segmented into zones, and the segmentation depends on the area. The national model considers trips longer than 100 km done by residents in all of Sweden and is segmented into 682 zones. The models in East and West consider all trips within Sweden, done by residents in the respective study area and is segmented such that spatial resolution decreases further away from the largest city in the area, Stockholm and Gothenburg respectively. The East and West study area each contains approximately 3000 zones.

3.1.3 The Netherlands

OViN (Onderzoek Verplaatsingen in Nederland) is a recent dataset on daily mobility of the Dutch population. The dataset consists of a basic survey at a national level and possible follow-up surveys. The research is a continuous daily study of the travel behaviour of Dutch people. Respondents are asked to keep track of where

Figure 3.2: Geographical overview of OViN zones in the Netherlands

they go for that particular day of the year, for what purpose, with what means of transport and how long it takes to get there. Based on this research, information can be obtained about all daily trips by Dutch people on Dutch territory.

All trips in the OViN data set originates and ends in postal code areas, grouped by their first four digits. In other words, the geographical partitioning into zones is defined by the location's first four postal code digits. In the Netherlands, this results in 4066 zones that trips can occur in between. A geographical overview of these zones can be seen in Figure 3.2.

3.1.4 São Paulo

The city of São Paulo has collected information on trips from citizens in São Paulo. The study, carried out in 2017, interviewed 32,000 households distributed in 517 research zones. In total, approximately 100,000 people were interviewed. Trips were collected from each respondent over 24 hours, and only weekdays were considered.

Of the 517 research zones, 342 represents the municipality of São Paulo, and 175 represents the neighbouring municipalities. Their geographical distribution is shown in Figure 3.3.

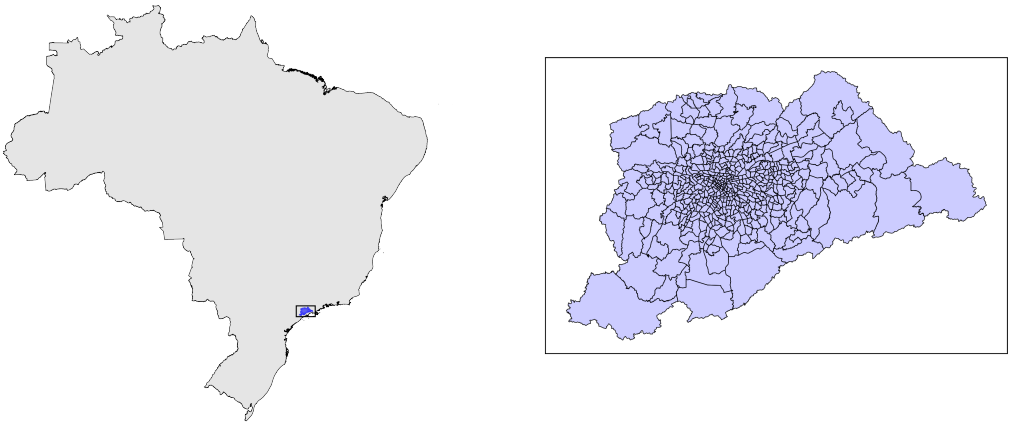

Figure 3.3: A geographical overview of the São Paulo Metropolitan region in Brazil (left) and its distribution of research zones (right).

3.2 Population representation of top geotag Twitter users

The representativeness of top geotag Twitter users for the whole population is crucial for our study and the validity using Twitter data for mobility estimation. The density distribution at the zone-level shows the discrepancy between Twitter users' derived home locations and the census number of residents in the corresponding zones. A similar process has been done at the county level [20], compared to which this project will move one step forward to quantify the population bias at a much finer geographical resolution.

For Sweden and the Netherlands, the GEOSTAT 2011 data source is used as the ground truth. The spatial resolution of GEOSTAT, 1 km^2 grids, is too detailed for our analysis. Hence, the grids are grouped into counties and municipalities before comparison. The derived distribution from Twitter will then be compared to the GEOSTAT distribution for each of the two resolutions (counties and municipalities).

For São Paulo, the population distribution is included in the data source, i.e. there is an estimated number of residents in each of the 517 research zones included in the study. Thus, the comparison of population distribution will be conducted at the zone level for the São Paulo region.

3.3 Mobility representation of the proposed model

While trip distance (d, km) does not encompass the direction of flow between places, it is an essential metric of mobility whose distribution reveals the validity of using Twitter to estimate travel demand [11]. Therefore, it is selected to test the proposed model as calibrated and validated against the external data sources.

Two baselines

In order to quantify the improvement of the model, we compare our model to two other models (hereafter called baselines) from the literature[20, 13, 12, 26]. The two baselines differ in how they construct trips from displacements observed in the raw geotagged tweets. The first baseline, henceforth called just *baseline*, considers every displacement to be a trip, i.e. every pair of consecutive geotagged tweets form the origin and destination of a trip. The second baseline, henceforth called *baseline-24*, considers displacements with a duration shorter than 24 hours to be trips, i.e. every pair of consecutive geotagged tweets posted within 24 hours of each other becomes an origin and a destination of a trip.

Comparison procedure

For each set of trips (*baseline*, *baseline-24*, *model*), an OD matrix is created. First, the coordinates of origin and destination for each trip is geographically joined to the zones of the external data source. Secondly, the number of trips between each origin-destination pair is calculated and normalised so that the sum of all cells in the OD matrix add up to 1.

Based on the distance between zones, 100 distance quantiles are calculated ($Q = 100$), such that each quantile contains the same number of origin-destination pairs. For each quantile, $0 < q \leq Q$, the share of trips from the external data source, t_q, and the model, t'_q is calculated. The similarity of the model compared to the external data source is quantified by Mean Squared Error, MSE (see eqs. (3.1) and (3.2)).

$$\mathrm{SE}_q = (t_q - t'_q)^2 \tag{3.1}$$

$$\mathrm{MSE} = \frac{\sum_{q=1}^{Q} \mathrm{SE}_q}{Q} \tag{3.2}$$

Model calibration

Because the model is parameterized (ρ, γ, and β), the question of which parameters are optimal, or close to optimal, should be addressed. Furthermore, it is possible that the optimal parameters in one region is not optimal for another region, for example due to geographical differences. Hence, the difference in optimal parameters for different regions should also be explored.

To find the optimal model parameters within a region a two-phased grid search is conducted. In the first phase a wide range of parameters are tested, as seen Eq. 3.3. After evaluating all parameters, the ones that achieves the lowest MSE of trip distance distribution, in comparison to external data source, are selected.

$$\begin{aligned} \rho &\in [0.3, 0.6, 0.9] \\ \gamma &\in [0.2, 0.5, 0.8] \\ \beta &\in [0.01, 0.04, 0.07] \end{aligned} \tag{3.3}$$

In the second phase of the grid search, a narrow set of parameters is selected for evaluation centred around the best parameters found in the first phase. For example, if $\rho = 0.6, \gamma = 0.5, \beta = 0.04$ was selected in the first phase, the parameters to be evaluated in second phase are depicted in Eq. 3.4. After evaluating all parameters in the second phase, the ones that achieve the lowest MSE are considered the optimal parameters for that region.

$$\begin{aligned} \rho &\in [0.5, 0.6, 0.7] \\ \gamma &\in [0.45, 0.5, 0.55] \\ \beta &\in [0.03, 0.04, 0.05] \end{aligned} \tag{3.4}$$

When the two phased grid search is completed for the three regions, the optimal parameters of the regions, and their respective MSE, are compared. In the best case, the optimal parameters in each region would be the same, and thus indicate that the model is robust across these regions.

4

Results

In this chapter, the results are presented; Section 4.1 shows the representativeness of the estimated home locations of observed Twitter geotag users compared with the external data sources described in the previous section. Section 4.2 presents the parameters of the model and the process used for setting them, and Section 4.2 shows the performance of the model in terms of the trip distance distribution.

4.1 Population representation

This section presents the spatial distribution of the estimated home locations of geotag Twitter users and compare their spatial distributions with census/survey data sources in Sweden, the Netherlands, and São Paulo.

Sweden

Figure 4.1 shows the representativeness of the estimated home locations of geotag Twitter users compared with the census population in Sweden. The top geotag Twitter users in Sweden are overly representing the residents in Stockholm county, where the capital of Sweden is located, but under-representing the residents in Västra Götaland county, where the second-largest city Gothenburg is located. At the municipality level, Figure 4.1.B shows that Twitter users are overly representing the residents who live in urban centres as well as a few more rural municipalities, such as Åre and Rättvik.The municipality of Sweden's capital, Stockholm, is over-represented by a factor of 2.2. Comparing Figure 4.1.A and Figure 4.1.B, shows that even if the population level is close to census at the county level, individual municipalities within the county can still be over- or under-represented. This is due to the smaller zones at municipality level, which sheds light on how the population is distributed in each county.

Figure 4.2 compares the estimated home locations of Twitter users and census population. For both county and municipality level, less populated areas are more likely to be under-represented by top Twitter users.

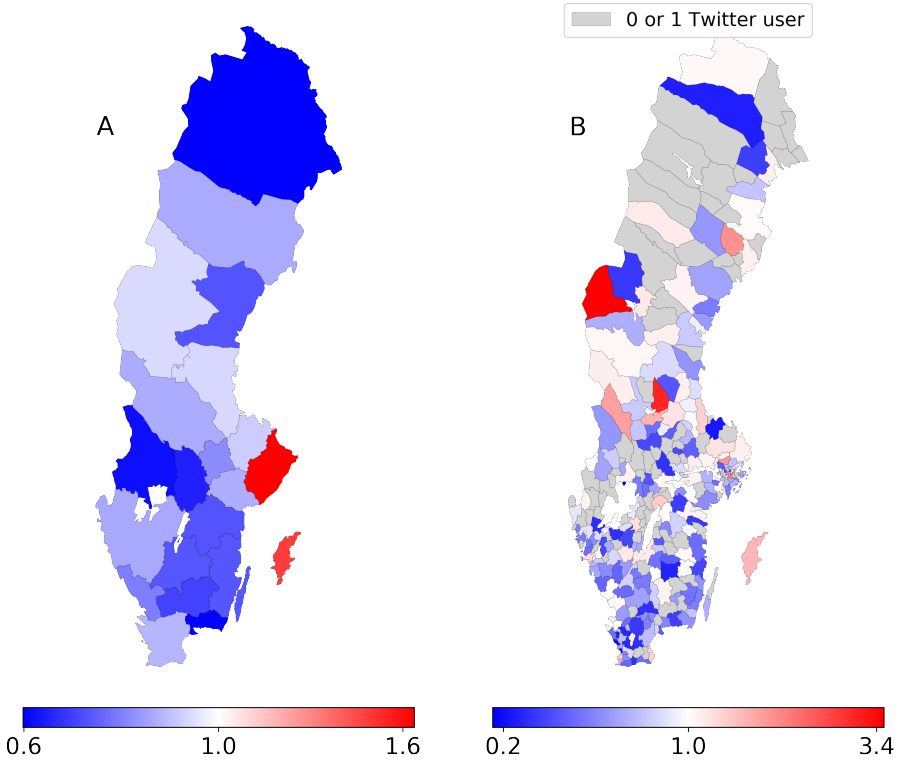

Figure 4.1: Spatial distribution of estimated home locations of Twitter users compared to census data in Sweden. The numbers on the colour bar represent the Twitter-derived population percentage divided by the percentage derived from GEOSTAT. 1 represents an equal ratio of residents between the Twitter users and census data, in the specific zone. **A**: Comparison at the county level. **B**: Comparison at the municipality level.

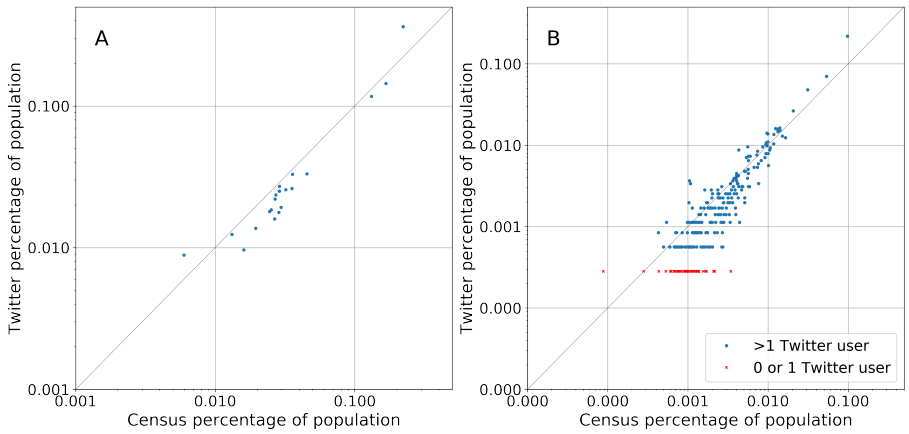

Figure 4.2: Comparison of estimated home locations of Twitter users with census data in Sweden. The diagonal line represents a perfect correlation. Each data point represents the share of population in a zone calculated from census (x axis) and top geotag Twitter users (y axis). **A**: County level. **B**: Municipality level.

The Netherlands

Figure 4.3 shows the population representation of estimated home locations of geotag Twitter users relative to the derived GEOSTAT distribution. The population distribution at the county level is similar to that of GEOSTAT. The county of North Holland is slightly over-represented, by a factor of 1.7. Figure 4.3.B shows the distribution at the municipality level. At this resolution, it is evident that the two sparsely populated islands Vlieland and Terschelling are over-represented by the Twitter users. More urban municipalities, such as Amsterdam and Utrecht are also over-represented; Amsterdam by a factor of 3.35 and Utrecht by a factor of 1.69.

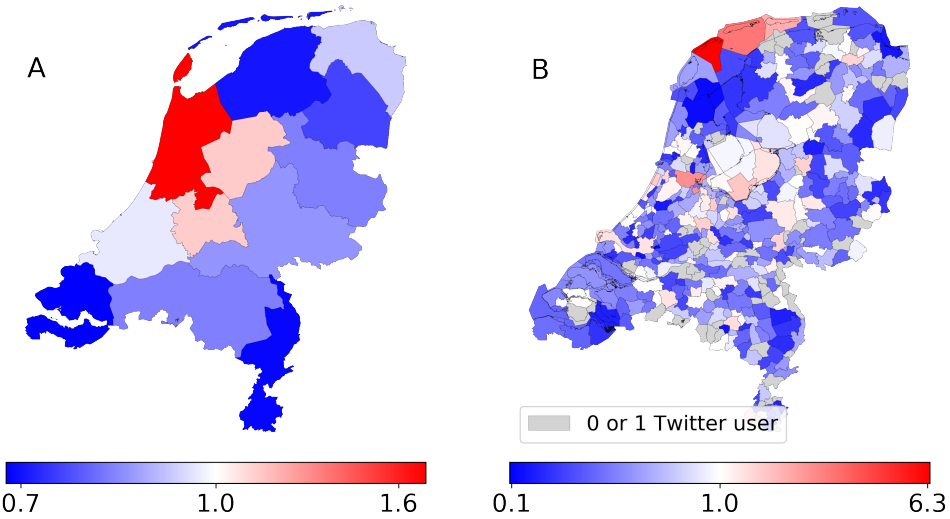

Figure 4.3: Spatial distribution of estimated home locations of Twitter users compared to census data in the Netherlands. The numbers on the colour bar represent the Twitter-derived population percentage divided by the percentage derived from GEOSTAT. 1 represents an equal ratio of residents between the Twitter users and census data, in the specific zone. **A**: Comparison at the county level. **B**: Comparison at the municipality level.

Figure 4.4 shows the relationship between the two sources at both municipality and county level. Similar to the findings in Sweden (Figure 4.2), geotagged tweets in the Netherlands tend to be more prevalent among the residents in urban and populated areas.

São Paulo

Figure 4.5 illustrates the results of population representation for the 517 research zones. It indicates that rural zones, located far from the inner city of São Paulo, are under-represented. Moreover, the Twitter distribution of inner São Paulo resembles that of the travel survey, with one exception. That is the research zone of Barra Funda, which has a travel survey estimate of 324 out of the 20 821 671 residents in São Paulo. For twitter, 22 out of 10 686 users are estimated to have a home location in Barra Funda. This results in an over-representation by a factor of 132.

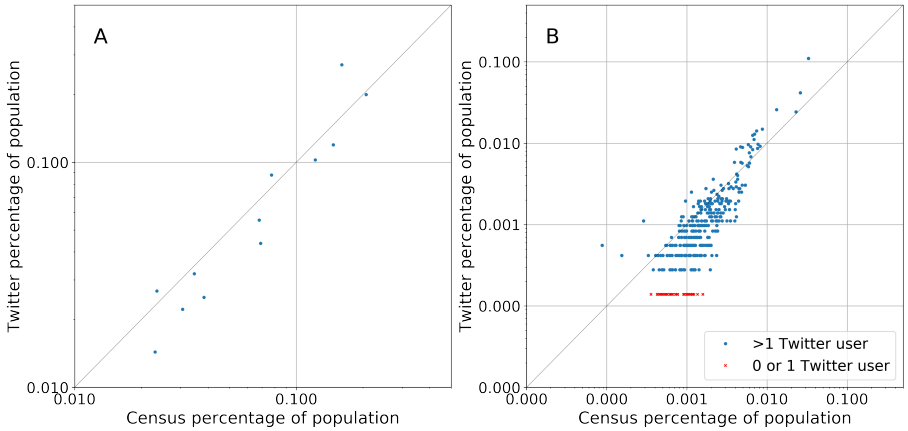

Figure 4.4: Comparison of estimated home locations of Twitter users with census data in the Netherlands. The diagonal line represents a perfect correlation. Each data point represents the share of population in a zone calculated from census (x axis) and top geotag Twitter users (y axis). **A**: County level. **B**: Municipality level.

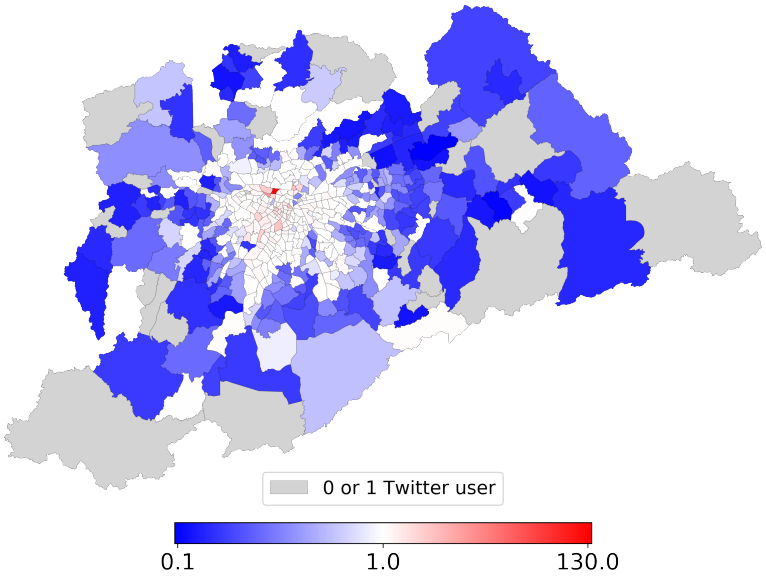

Figure 4.5: Spatial distribution of estimated home locations of Twitter users compared to census data in São Paulo. The numbers on the colour bar represent the Twitter-derived population percentage divided by the percentage derived from GEOSTAT. 1 represents an equal ratio of residents between the Twitter users and census data, in the specific study zone.

Resembling what is observed in Sweden and the Netherlands, top geotag Twitter users in São Paulo tend to overly represent the residents in densely populated areas. However, as shown Figure 4.6 the discrepancy between Twitter users and census population is more salient than Sweden and the Netherlands, i.e., the top geotag Twitter users in São Paulo display a lower population representation.

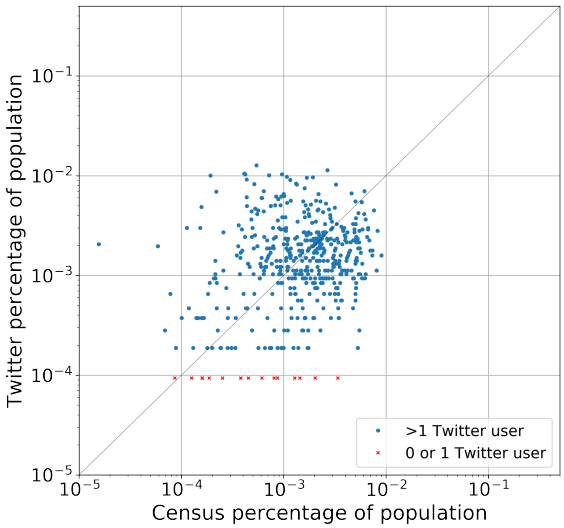

Figure 4.6: Comparison of estimated home locations of Twitter users with census data in São Paulo. The diagonal line represents a perfect correlation. Each data point represents the share of population in a zone calculated from census (x axis) and top geotag Twitter users (y axis).

4.2 Individual mobility model: parameters and validation

Model parameters

This section presents the results obtained in each region with different model configurations.

Figure 4.7 shows the influence of different model configurations for Sweden, the Netherlands and São Paulo. For Sweden, which consists of three areas, the optimisation is based on the sum of the three MSE values. The optimal values for exploration rate parameters γ and ρ are different for Sweden compared to the other two regions. In Sweden, the exploration rate parameters are optimal when $\gamma \in [0.75, 0.8]$ and $\rho \in [0.3, 0.4]$, while in São Paulo and the Netherlands they are optimal when $\gamma \in [0.45, 0.5]$ and $\rho \in [0.6, 0.7]$. The results indicate that a lower probability for exploration is preferential in Sweden, while a slightly higher exploration rate in the Netherlands and São Paulo.

Parameter β, controlling preference for short-distance travel, is subsequently analysed when exploration parameters is fixed in their respective range, depicted in the right column of figure 4.7. For all regions, there is one value of β that achieves a better score, regardless of the values of exploration rate parameters. In Sweden, that value is $\beta = 0.03$, in the Netherlands $\beta = 0.04$, and São Paulo $\beta = 0.05$. This indicates that the smaller the region in the study, the larger the optimal value of β is. Furthermore, the smaller the region under study, the less influence the exact value of β has.

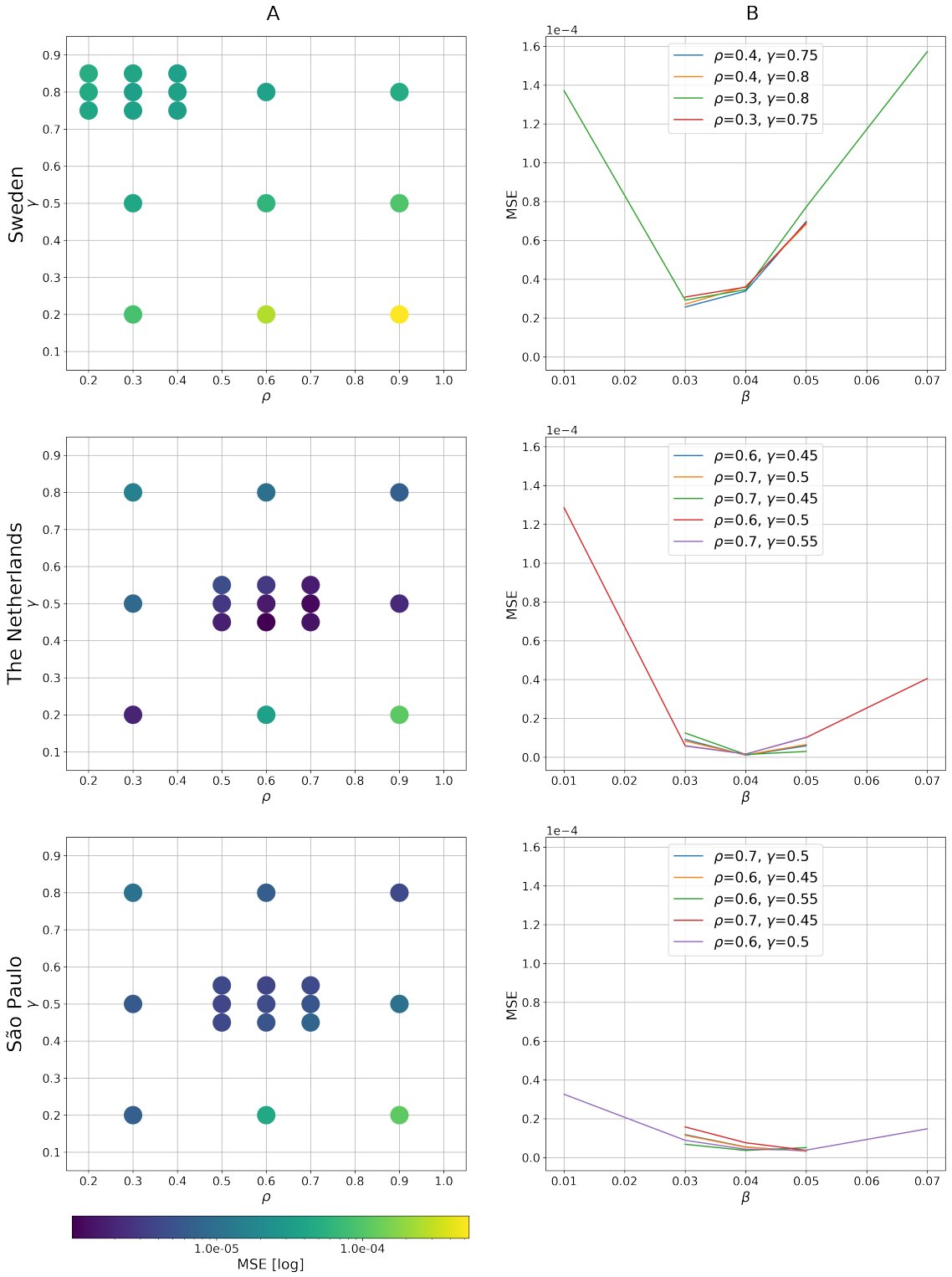

Figure 4.7: Parameter topology in Sweden, the Netherlands and São Paulo. **A** Influence of exploration parameters γ and ρ on MSE - β is fixed at 0.04. **B** Influence of β on MSE - γ and ρ is fixed at different values. One pair of ρ and γ was included in both the first and the second phase of grid search, thus, having results for additional β values.

Table 4.1 shows the best model configuration for each region. The complete table of results for different model configurations across the three regions can be found in Appendix B.

Region	γ	ρ	β
Sweden	0.75	0.4	0.03
The Netherlands	0.45	0.6	0.04
São Paulo	0.45	0.6	0.05

Table 4.1: Optimal set of parameters for the model in each region.

Validation

This section presents the results of validation for each region using the method described in Section 3.3. In this section, the model's results are presented using the best set of parameters found in the previous section.

Sweden

The trip distance distribution for the region of Sweden has three areas: "National", "East" and "West". Note that the maximum straight-line distance for Sweden is around 1500 km.

National

Sampers-National model considers trips with a minimum distance of 100 km by residents from all over Sweden. Figure 4.8 shows the trip distance distributions for the National region from the external source (Sampers), the two baselines, and the model. Both the *baseline* and the *baseline-24* under-represents medium distance trips, 100 km to 250 km, and over-represents long-distance trips up to about 600 km. Although *baseline-24* yields an improvement of the *baseline* distribution, it still deviates from Sampers. Our model closely follows the cumulative distribution of Sampers up until 200 km. For the distance above 200 km, it slightly deviates from Sampers for the distance of 200 - 500 km.

East

For the East area, the Sampers output considers trips in all of Sweden, made by the residents who live in the East area. The trip distance distributions for the area are illustrated in Figure 4.9. The first distance quantile, which contains trips up to 3.2 km, is under-represented by both the *baseline* and the model. The *baseline* continues to under-represent trips up to 100 km and then heavily over-represents long-distance trips. *baseline-24* resembles the *baseline* except for having a lot more short-distance trips, with 76% of the trips in the first distance quantile. The model generates slightly more trips in the range of 3.2 - 7 km than Sampers. For distance above 7 km, the model approximates the cumulative distribution derived from Sampers.

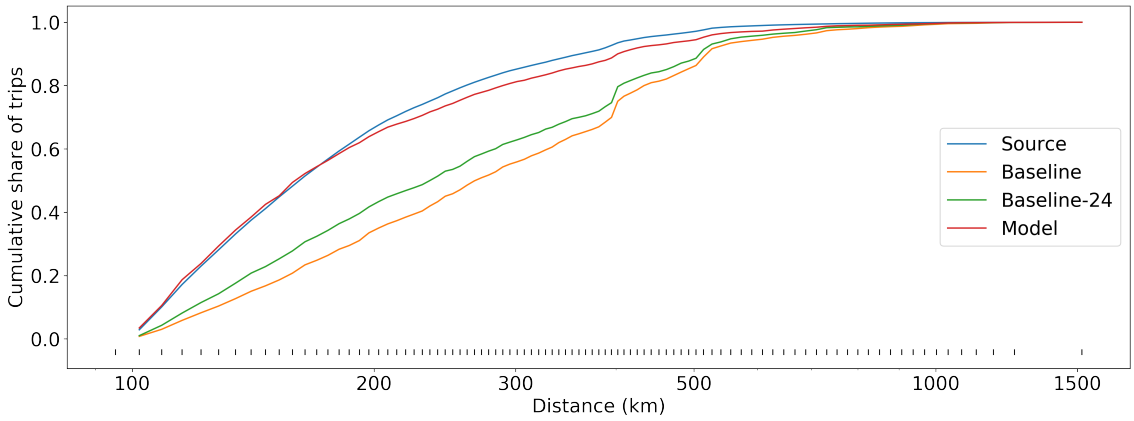

Figure 4.8: Trip distance distributions for the National area, Sweden (Source=Sampers-National). Cumulative percentage of trips in each distance quantile. The black vertical lines indicate the upper and lower boundaries for the distance quantiles. The same below for all figures on distance distributions.

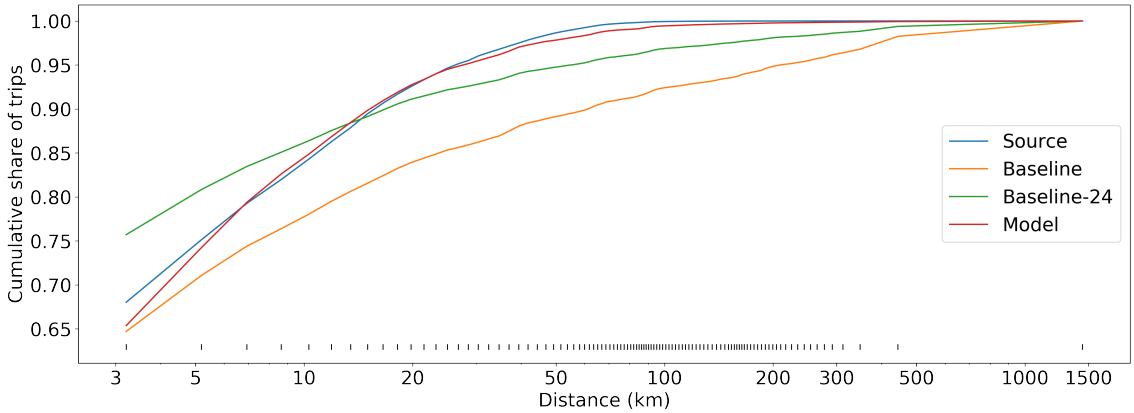

Figure 4.9: Trip distance distributions for the East area (Source=Sampers-East).

West

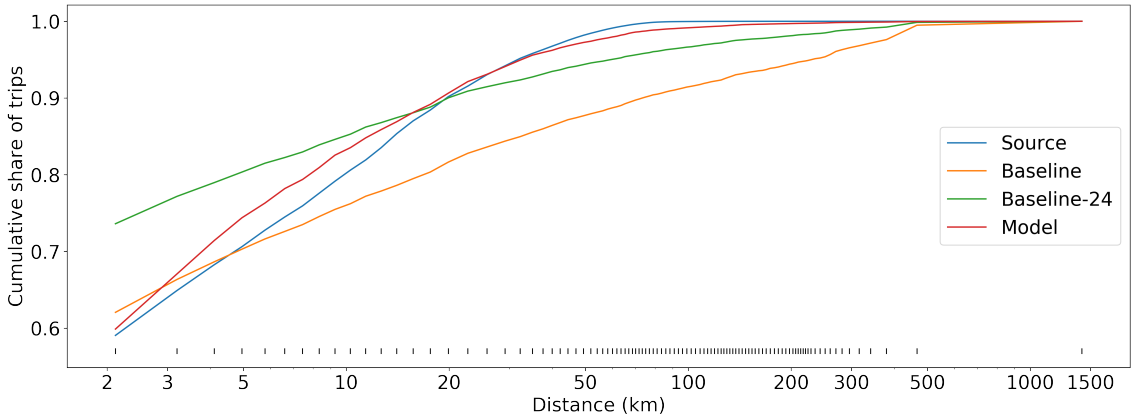

Figure 4.10: Trip distance distributions for the West area (Source=Sampers-West).

For the West area, the Sampers output considers trips in all of Sweden, made by the

residents who live in the West area, and the trip distance distributions are illustrated in Figure 4.10. Compared to Sampers, the two baselines overly represent the first quantile, 0 - 2.1 km, but then under-represents trips with a distance less than 100 km. The time threshold in *baseline-24* generates more trips in the first quantile than the unbounded *baseline*. Moreover, the baseline also over-represents long-distance trips. In contrast, the model slightly over-represents trips up to 5 km and then follows the same cumulative distribution as Sampers.

In summary, the model improves over the two baselines across all three areas, especially on the national level. On the national level *baseline-24* is slightly better than *baseline* while being worse on both East and West. According to figures 4.10 and 4.9, this is due to the much larger share of short-distance trips (<10 km). The MSE between the Sampers' output and the baselines' as well as the model are summarised in Table 4.2.

	MSE (10^{-5}) between Source (Sampers)		
Region	Baseline	Baseline-24	Model
National	14.9	9.49	**0.79**
East	2.06	7.13	**0.61**
West	2.51	23.2	**1.17**

Table 4.2: Summary of MSE in Sweden, comparing baselines to the best model configuration.

The Netherlands

The longest distance one can travel within the Netherlands is around 300 km, shorter than Sweden due to its different geometry of the territory. Figure 4.11 shows the cumulative trip distance distribution, comparing the source (OViN), *baseline*, *baseline-24* and model.

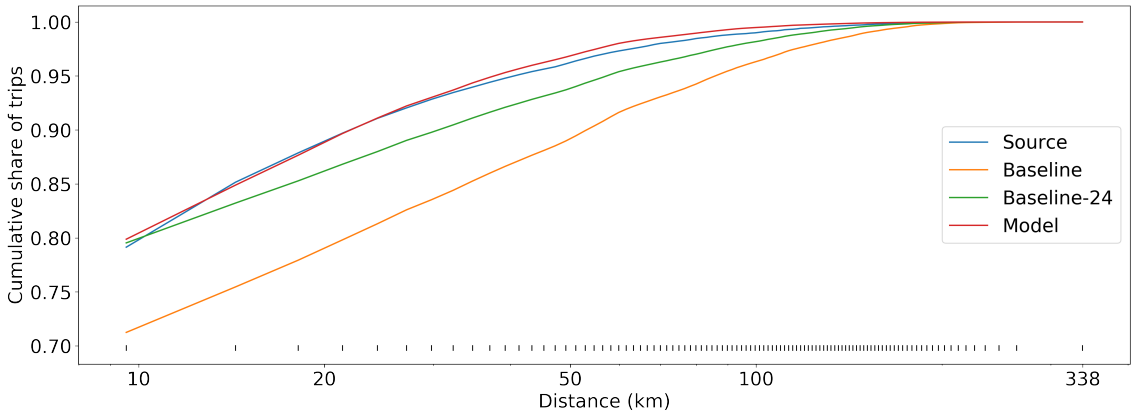

Figure 4.11: Trip distance distributions for the Netherlands (Source=OViN).

The *baseline* significantly under-represents travel shorter than 10km, being off by

10%. The cumulative trips distance distribution for the *baseline* continues to be lower than expected until around 50km, and trips longer than 50km are slightly over-represented. The *baseline-24*, however, has a significantly better share of short distance trips but still follow the under and over-representation pattern as the *baseline*. In comparison, the model corrects the short distance under-representation of Twitter data, and achieves a good fit with the source data for all the distance ranges.

The MSE between the trips in the external source (OViN) and the baselines' as well as the model are summarised in Table 4.3. The *baseline-24* outperforms the *baseline*, due to it's higher share of short-distance trips (<10 km). The model further improves on both of the baselines, because of a better share of medium-distance trips, which are very similar to OViN.

	MSE (10^{-5}) between Source (OViN)		
Region	Baseline	Baseline-24	Model
The Netherlands	6.72	0.63	**0.10**

Table 4.3: Summary of MSE in the Netherlands, comparing baseline to the best model configuration.

São Paulo

São Paulo is different from Sweden and the Netherlands because it is a city where the longest distance one can travel is significantly shorter, slightly more than 100km. Figure 4.12 shows the cumulative trip distance distribution, comparing the source (travel survey), *baseline*, *baseline-24* and the model. The *baseline* overly represents travel in the shortest distance interval, 0 - 2.15 km, while slightly under represents travel on distances up to 20 km. Here, the *baseline-24* actually performs worse than the *baseline*, by having a 20 percentage point increase in the shortest distance interval as compared with the source data. The model, however, corrects the short distance over-representation, and all together achieves a good fit with the source data.

The MSE between the trips in the travel survey and the baselines' as well as the model are summarised in Table 4.4. The *baseline* model over-represents the share of short-distance trips (<3 km), and the *baseline-24* further increase the share of short-distance trips leading to a worse MSE. The model, however, achieves a good fit for the share of short-distance trips, which leads to the best MSE of the three models.

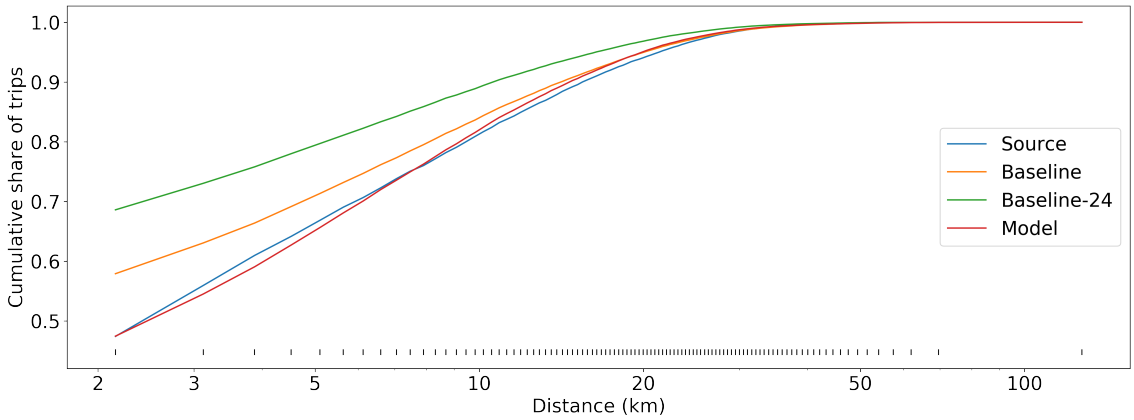

Figure 4.12: Trip distance distributions for São Paulo (Source=OD Survey 2017 in São Paulo).

	MSE (10^{-5}) between Source (travel survey)		
Region	Baseline	Baseline-24	Model
São Paulo	12.6	47.7	**0.33**

Table 4.4: Summary of MSE in São Paulo, comparing baseline to the best model configuration.

5

Discussion

This thesis estimates travel demand using geotagged tweets. We develop an individual mobility model that accounts for and corrects the observed mobility behaviour biases found in geotagged tweets so that geotagged tweets can improve the estimates of travel demand. The model is calibrated and validated in two countries, Sweden and the Netherlands, and one city, São Paulo.

5.1 Top geotag Twitter users vs general population

In agreement with previous studies, we find that the top geotag users on Twitter overly represent residents in urban areas (see figs. 4.1, 4.3 and 4.5). Stockholm county in Sweden and North Holland in the Netherlands both indicate an over-representation by a factor of, approximately, 1.7. In addition to previous work, we also find that the urban area over-representation can be attributed to the most central areas. The centre of Stockholm city in Sweden is overly represented by a factor of 2.3, and this number is 3.4 for Amsterdam in the Netherlands, and 130 for Barra Funda in São Paulo. Quantifying the population biases sheds light on the need for further de-biasing. For instance, the discrepancy between the spatial distribution of top geotag Twitter users and the general population could be used to attach a weight to each top geotag Twitter user when modelling the population flows between places.

An issue with the high-resolution comparison of population density arises due to the small sample size of Twitter users in each region. This leads to the summary statistics in small zones being very sensitive to scale, and them being disproportionately over- or under-represented in comparison to the true population. This indicates the importance of properly choosing study zones when using geotagged tweets.

5.2 Mobility measured by geotagged tweets

The *baseline* trips are created by connecting every two consecutive geotagged tweets by the same user without any time threshold, similar to other methods in the literature [20, 26]. We found that the trips tend to overly represent long-distance travel

(see figs. 4.8 to 4.11). For Sweden and the Netherlands, the distance at which the *baseline* starts to overly represent is approximately 50 km. The *baseline*'s tendency to over-represent long distance trips further confirms the behaviour bias described by Tasse et al (2017) [10]. However, it should also be noted that the external data sources used in validation might not perfectly represent the true mobility for the region. Traditional household travel surveys have, for example, been shown to under-report long-distance trips [27, 28].

We also show that, if the considered region is geographically small, a short distance over-representation, less than 5 km, emerges in the trip distance distribution. For example, in São Paulo, where the maximum distance is 100 km, trips less below 3 km are over-represented by 10% (see Figure 4.12). Hence, regardless of the study area's scale, trips derived directly from geotagged tweets will not yield a representative trip distance distribution. Furthermore, the share of short distance trips is affected by the spatial aggregation used, and the spatial resolution of the source. Geotagged tweets have a higher spatial resolution, leading to many short-distance trips which is more realistic than travel surveys because surveys sometimes only have tips with minimum distance of 1 km.

The *baseline-24* uses a time threshold of 24 hours, as commonly found in literature [12, 13]. The threshold appears to primarily remove long-distance trips, which in turn yields a more representative share of long-distance trips compared to the *baseline*. We also find that the threshold produces a considerably greater share of short-distance trips, i.e. 0 - 10 km, than the baseline. In all regions except for the Netherlands, this impacts the MSE negatively, as the ground truth has a lower share of short-distance trips. Overall, we find that the *baseline-24* produce trip distance distributions worse than the *baseline*. Furthermore, the common practice [12, 13] of adding a time threshold to convert displacements into trips reduces the amount of available data, and as the accuracy of the estimations is actually decreased, there are not many benefits of this practice.

5.3 Individual mobility model

Based on the mobility biases found in Twitter data, we propose an *Individual Mobility Model* that accounts for the observed behavioural biases of geotagged tweets. The proposed model integrates two natural dimensions of individual mobility: regularly returning to habitual locations and occasionally exploring new locations. To address the caveat of under-representation of habitual places such as home and workplace, the model re-scales the observed visitation frequency of various locations by combining their order and Zipf's law [5]. In addition, the model combines visitation frequency and distance when selecting a location to return to. By doing so, the tendency for long-distance travel observed in Twitter data, is corrected. As a result, the modelled mobility prefers making visits to places closer to where the individual is currently located more than the baselines.

The model consistently improves upon the baseline models across all regions (see Table 5.1), especially on short-distance trips (< 50 km). Despite the good results,

the model have some limitations that should be explored further; the dependence on estimated home location, and assumed independence of jump size and bearing distributions. First, the model is dependant on the estimated home location from the tweets, as the first visit every day is assumed to be at home. Consequently, inherent in the the model is the uncertainties of the estimated home locations. Theoretically, the model only requires the first visit of the user's timeline to be at a known preferred location, and the remaining visits could be generated without starting at the home location every day. This would drastically decrease the dependence on estimated home locations, but other effects are still unknown.

Secondly, the jump size and bearing distributions are independent in the proposed model which is a simplification of the reality. For example, it is not likely that a person living in Sweden makes a 6000 km north (crossing the north pole), while taking a flight to New York of the same distance is much more likely. This could be addressed by sampling the jump size conditionally based on the bearing in future work. The consequences of this assumption do not emerge in the results, however, because the trip distance distribution is one aspect of mobility, and does not take bearing into account.

	MSE (10^{-5})			Parameters		
Region	Baseline	Baseline-24	Model	γ	ρ	β
Sweden: National	14.9	9.49	**0.79**	0.75	0.4	0.03
Sweden: East	2.06	7.13	**0.61**	0.75	0.4	0.03
Sweden: West	2.51	23.2	**1.17**	0.75	0.4	0.03
The Netherlands	6.72	0.63	**0.10**	0.45	0.6	0.04
São Paulo	12.6	47.7	**0.33**	0.45	0.6	0.05

Table 5.1: MSE and parameters of the best performing model in each region compared to the baselines.

5.4 Model sensitivity to different parameters

There are three parameters in the proposed model: ρ, γ and β We calibrate the model parameters for three of the regions in this thesis: Sweden, the Netherlands and São Paulo.

The β parameter, controlling preference for short-distance travel, is the parameter that influences the MSE the most. Furthermore it appears to be correlated with the maximum length travel within the region. That is, the longer the maximum travel, the lower β is found. Despite this, it is found that the shorter the maximum travel in the region, the less influence β has on the model. This is because the lower probability of long-distance trips, controlled by β, becomes irrelevant when only considering short-distance trips. The optimal value of β in each region is summarised in Table 5.1. Parameters controlling exploration rate, ρ and γ, have

less influence on the MSE in the studied regions than β. However, across all regions the optimal parameters suggest that small amounts of exploration is beneficial for the MSE presumably due to biases towards unusual locations in geotagged tweets as compared with the actual mobility.

Although the optimal model parameters found for each region is slightly different, our grid search results (see Figure 4.7) suggests that the optimal values are in the same range for all regions: β around 0.04, ρ around 0.5 and γ between 0.45 and 0.75. We expect that applying the model with these parameters to a new region would yield better results than using consecutive tweets directly, with or without a time threshold.

5.5 Future work

This thesis carefully examines the biases of geotagged tweets in two aspects, population representation and mobility representation as measured by geotagged tweets. Despite addressing the behavioural biases by proposing the individual mobility model, the population bias of the top geotag Twitter users are not integrated into the modelling. In order to de-bias on both aspects for travel demand estimation, one future direction can be adding varying weights to those top geotag Twitter users as compared with the general population when aggregating their trajectories to create an origin-destination matrix.

The proposed individual mobility model performs well as compared with the other established data sources in the three regions that we studied. However, the evaluation method in this thesis is limited to the trip distance distribution which constitutes one part of the travel demand. The next step is to further validate the model on where visits are generated to take spatial orientation into consideration.

Another future work is to generalise the proposed model into global regions to further test its feasibility and conduct cross-regional analysis on the individual mobility to make full use of social media data as an emerging data source in mobility study.

6

Conclusion

Traditional household travel surveys is a commonly used method of estimating travel demand. However, the cost of conducting these travel surveys is increasing, while the response rate is decreasing. This has led researchers to explore new sources of data that can be used to estimate travel demand. Among these new data sources is social media data such as geotagged tweets from Twitter, which is promising due to it's large quantity of available data and low cost of access. At the same time, using Twitter for travel demand estimation has garnered criticism regarding the biases inherent in Twitter data.

We quantify and confirm the results of previous research regarding the biases of Twitter users. (1) The users of Twitter are overly represented in urban areas, and even more so in the absolute centre of these areas. (2) Twitter users geotag their tweets during trips far from home leading to an over-representation of long distance trips if used directly as a proxy of human mobility.

Our main contribution to the field, corresponding to the revealed biases, is to develop a novel model to generate individual mobility trajectories for travel demand estimation using geotagged tweets. It takes behavioural biases into consideration. The proposed model produces individual based series of visits based on the observed geotagged activities. The proposed model integrates two natural dimensions of individual mobility: regularly returning to habitual locations and occasionally exploring new locations that have not been visited before. The proposed model addresses the under-representation of habitual places such as home and workplace and corrects the geotagging behavioural bias of being less constrained by distance. Validation on three different regions suggests that the model is able to capture the essential travel demand in multiple regions of distinct geographical properties as compared with the other established data sources. Finally, the results suggest that the model's parameters are robust across regions studied, and by using the parameters found in this thesis one can expect similar improvements compared to contemporary approaches across other regions.

Given that geotagged tweets as an emerging data source have been used widely to characterise travel demand, it is imperative to address the known biases, as we have demonstrated here. Future work includes examining the performance of the model by using more validation metrics than trip distance distribution, integrating the population de-biasing in the model and test the model in other regions.

Bibliography

[1] Yang Yue, Tian Lan, Anthony Yeh, and Qing-Quan Li. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1:69–78, 05 2014.

[2] Yuan Liao. Understanding human mobility with emerging data sources: Validation, spatiotemporal patterns, and transport modal disparity. 2020.

[3] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from twitter. *PLOS ONE*, 10(7):1–16, 07 2015.

[4] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

[5] Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6, 10 2010.

[6] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv:1306.5204 [physics]*, June 2013. arXiv: 1306.5204.

[7] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '13, pages 605–613, Chicago, Illinois, USA, August 2013. Association for Computing Machinery.

[8] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, May 2014.

[9] Maxime Lenormand, Miguel Picornell, Oliva Garcia Cantu Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frias-Martinez, and Jose Javier Ramasco. Cross-checking different sources of mobility information. *PLoS ONE*, 9, 04 2014.

[10] Dan Tasse, Zichen Liu, Alex Sciuto, and Jason I Hong. State of the geotags: Motivations and recent changes. In *ICWSM*, pages 250–259, 2017.

[11] Yuan Liao, Sonia Yeh, and Jorge Gil. Feasibility of estimating travel demand using social media data. *Transportation*, 2020.

[12] Jae Hyun Lee, Adam Davis, Elizabeth McBride, and Konstadinos G Goulias. Statewide comparison of origin-destination matrices between california travel model and twitter. In *Mobility Patterns, Big Data and Transport Analytics*, pages 201–228. Elsevier, 2019.

[13] A. Kheiri, F. Karimipour, and M. Forghani. Intra-Urban Movement Flow Estimation Using Location Based Social Networking Data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 15:781–785, December 2015.

[14] Fan Yang, Peter J. Jin, Yang Cheng, Jian Zhang, and Bin Ran. Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *International Journal of Sustainable Transportation*, 9(8):551–564, November 2015.

[15] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[16] Juan Gonzalo Cárcamo, Roderick Grahm Vogel, Adam M. Terwilliger, Jonathan P. Leidig, and Greg Wolffe. Generative models for synthetic populations. In *Proceedings of the Summer Simulation Multi-Conference*, SummerSim '17, San Diego, CA, USA, 2017. Society for Computer Simulation International.

[17] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. Next place prediction using mobility markov chains. 04 2012.

[18] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4):463–483, October 2012.

[19] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking*, 19(3):630–643, 2011.

[20] Yuan Liao, Sonia Yeh, and Gustavo S Jeuken. From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. *EPJ Data Science*, 8(1):34, 2019.

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[22] Christian M. Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[23] Official Statistics of Sweden. *Swedish National Travel survey (RVU Sweden) 2011—2016*, 2016.

[24] Eurostat. Population grids, 2018. Data retrieved from Statistics Explained.

[25] Trafikverket. Sampers, Nov 2019.

[26] Song Gao, Jiue-An Yang, Bo Yan, Yingjie Hu, Krzysztof Janowicz, and Grant McKenzie. Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area. 09 2014.

[27] Maxim Janzen. Population synthesis for long-distance travel demand simulations. In *6th symposium of the European association for research in transportation (hEART 2017)*. ETH Zurich, Institute for Transport Planning and Systems, 2017.

[28] Zhenzhen Wang, Sylvia He, and Yee Leung. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 03 2017.

A

Notations

Notation	Definition
i	Individual index
(X, Y)	Decimal coordinates of a geotagged tweet
t	Local time of a geotagged tweet converted
$\mathbf{G}_i$	Sequence of a user's geotagged tweets
p	Index of geotagged tweet in $\mathbf{G}_i$
N_i	Total number of geotagged tweets in $\mathbf{G}_i$
w	Day of week
h	Hour of day
$\mathbf{S}_i$	Set of distinct places visited by individual i
j	Index of distinct place in $\mathbf{S}_i$
n_i	Total number of distinct places in $\mathbf{S}_i$
K_j	Frequency of visiting a place j for individual i
f_j	Frequency rate of place j among total visited places in $\mathbf{S}_i$
$\theta_{p,p+1}$	Distance between two consecutive geotagged tweets
$\alpha_{p,p+1}$	Bearing between two consecutive geotagged tweets
s	A place in $\mathbf{S}_i$
s_h	Identified home place of the individual
$\mathbf{L}_i$	Model output of individual mobility trajectory for individual i
D	Total number of days for $\mathbf{L}_i$
$\mathbf{T}_{i,d}$	A series of visits for individual i at day d
M_d	Total number of visits for $\mathbf{T}_{i,d}$
$v_{i,d,m}$	The mth visit at dth day for individual i
γ, ρ	Two parameters that control exploration in the proposed model
β	The parameter that controls returning in the proposed model
ζ	The parameter of Zipf's Law
k	Rank of visited places by their visiting frequency

Table A.1: Lookup table with the main symbols and relevant notations used in this thesis.

B

Parameter tuning

ρ	γ	β	MSE* (10^{-5})
0.4	0.75	0.03	2.57
0.4	0.8	0.03	2.71
0.3	0.8	0.03	2.92
0.3	0.75	0.03	3.08
0.4	0.85	0.03	3.22
0.4	0.75	0.04	3.39
0.3	0.8	0.04	3.46
0.4	0.85	0.04	3.51
0.3	0.75	0.04	3.59
0.4	0.8	0.04	3.62
0.3	0.85	0.03	3.74
0.6	0.8	0.04	3.75
0.3	0.85	0.04	4.04
0.2	0.75	0.04	4.13
0.3	0.5	0.04	4.23
0.2	0.8	0.04	4.61
0.2	0.75	0.03	4.74
0.9	0.8	0.04	4.82
0.2	0.85	0.04	5.05
0.2	0.8	0.03	5.83
0.6	0.5	0.04	6.23
0.4	0.8	0.05	6.84
0.3	0.75	0.05	6.92
0.9	0.5	0.07	6.92
0.4	0.75	0.05	6.98
0.2	0.85	0.03	7.0
0.3	0.2	0.07	7.09
0.4	0.85	0.05	7.25
0.2	0.75	0.05	7.27
0.2	0.85	0.05	7.45
0.2	0.8	0.05	7.67
0.3	0.8	0.05	7.73
0.3	0.85	0.05	7.83
0.6	0.5	0.07	8.05
0.3	0.2	0.04	8.77
0.9	0.5	0.04	9.66
0.9	0.8	0.07	10.75
0.3	0.5	0.07	12.54
0.6	0.8	0.07	13.03
0.3	0.8	0.01	13.72
0.6	0.8	0.01	15.69
0.3	0.8	0.07	15.73
0.3	0.5	0.01	16.55
0.6	0.2	0.07	16.91
0.9	0.8	0.01	18.57
0.6	0.5	0.01	23.72
0.6	0.2	0.04	25.87
0.3	0.2	0.01	28.8
0.9	0.5	0.01	31.22
0.9	0.2	0.07	42.41
0.6	0.2	0.01	52.99
0.9	0.2	0.04	53.84
0.9	0.2	0.01	81.65

Table B.1: Performance, in terms of MSE*, for the different model configurations in Sweden, sorted by MSE*. MSE* is the sum of the three MSE values received for the areas “National”, “East” and “West”.

ρ	γ	β	MSE (10^{-5})
0.6	0.45	0.04	0.1
0.7	0.5	0.04	0.12
0.7	0.45	0.04	0.14
0.6	0.5	0.04	0.16
0.7	0.55	0.04	0.16
0.6	0.5	0.04	0.17
0.5	0.45	0.04	0.18
0.3	0.2	0.04	0.19
0.9	0.5	0.04	0.21
0.5	0.55	0.03	0.26
0.5	0.5	0.04	0.29
0.6	0.55	0.04	0.3
0.7	0.45	0.05	0.3
0.5	0.5	0.03	0.4
0.6	0.55	0.03	0.4
0.5	0.55	0.04	0.44
0.6	0.2	0.07	0.49
0.5	0.45	0.03	0.57
0.7	0.55	0.03	0.58
0.6	0.45	0.05	0.59
0.6	0.5	0.03	0.59
0.7	0.5	0.05	0.65
0.9	0.8	0.04	0.66
0.3	0.5	0.04	0.82
0.7	0.5	0.03	0.83
0.6	0.45	0.03	0.92
0.6	0.8	0.04	1.0
0.6	0.5	0.05	1.02
0.7	0.55	0.05	1.03
0.5	0.45	0.05	1.04
0.7	0.45	0.03	1.25
0.6	0.55	0.05	1.41
0.5	0.5	0.05	1.42
0.3	0.8	0.04	1.52
0.9	0.5	0.07	1.73
0.5	0.55	0.05	1.88
0.3	0.2	0.07	1.91
0.6	0.2	0.04	3.49
0.6	0.5	0.07	4.06
0.9	0.2	0.07	5.77
0.9	0.8	0.07	6.82
0.3	0.5	0.07	7.37
0.6	0.8	0.07	8.43
0.3	0.8	0.01	9.42
0.3	0.8	0.07	10.09
0.6	0.8	0.01	10.39
0.3	0.5	0.01	10.64
0.9	0.8	0.01	11.0
0.9	0.2	0.04	11.08
0.6	0.5	0.01	12.85
0.3	0.2	0.01	14.76
0.9	0.5	0.01	15.21
0.6	0.2	0.01	21.77
0.9	0.2	0.01	29.77

Table B.2: Performance, in terms of MSE, for the different model configurations in the Netherlands, sorted by MSE.

ρ	γ	β	MSE (10^{-5})
0.7	0.5	0.05	0.33
0.6	0.45	0.05	0.34
0.6	0.55	0.04	0.37
0.7	0.45	0.05	0.38
0.5	0.5	0.04	0.38
0.6	0.5	0.05	0.39
0.7	0.55	0.05	0.39
0.5	0.45	0.04	0.4
0.5	0.55	0.04	0.4
0.9	0.5	0.07	0.41
0.9	0.8	0.04	0.42
0.7	0.55	0.04	0.42
0.6	0.5	0.04	0.43
0.5	0.45	0.05	0.43
0.6	0.55	0.05	0.51
0.6	0.45	0.04	0.53
0.5	0.5	0.05	0.54
0.7	0.5	0.04	0.55
0.5	0.55	0.03	0.55
0.3	0.5	0.04	0.58
0.6	0.8	0.04	0.64
0.3	0.2	0.04	0.66
0.5	0.5	0.03	0.67
0.6	0.55	0.03	0.69
0.5	0.55	0.05	0.71
0.7	0.45	0.04	0.76
0.3	0.2	0.07	0.79
0.5	0.45	0.03	0.82
0.6	0.5	0.03	0.89
0.7	0.55	0.03	0.92
0.9	0.5	0.04	1.04
0.3	0.8	0.04	1.06
0.6	0.45	0.03	1.15
0.7	0.5	0.03	1.18
0.3	0.8	0.01	1.29
0.6	0.2	0.07	1.39
0.6	0.5	0.07	1.48
0.7	0.45	0.03	1.58
0.6	0.8	0.01	1.76
0.3	0.5	0.01	1.86
0.9	0.8	0.01	2.33
0.9	0.8	0.07	2.95
0.6	0.5	0.01	3.26
0.3	0.5	0.07	3.75
0.3	0.2	0.01	4.02
0.6	0.8	0.07	4.09
0.6	0.2	0.04	4.39
0.9	0.5	0.01	5.12
0.3	0.8	0.07	5.67
0.9	0.2	0.07	7.15
0.6	0.2	0.01	9.75
0.9	0.2	0.04	11.32
0.9	0.2	0.01	16.83

Table B.3: Performance, in terms of MSE, for the different model configurations in São Paulo, sorted by MSE.