CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Simulating Mobility of Large Population using Mobile Application Data

Master's thesis in Computer Science and Engineering

MATTIAS RYDSTRÖM

DIANA SALIM

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Simulating Mobility of Large Population using Mobile Application Data

MATTIAS RYDSTRÖM
DIANA SALIM

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Simulating Mobility of Large Population using Mobile Application Data

MATTIAS RYDSTRÖM, DIANA SALIM

Supervisor: Yuan Liao, Department of Physical Resource Theory, Space, Earth and Environment
Examiner: Jorge Gil, Architecture and Civil Engineering

Master's Thesis 2023
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 2023

Simulating Mobility of Large Population using Mobile Application Data

MATTIAS RYDSTRÖM, DIANA SALIM
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Understanding people's travel patterns and knowing the frequency of their travel activities provide important insights for effective transport management and infrastructure planning. While traditional travel surveys have historically contributed to this valuable information, they come with shortcomings such as being expensive to collect, easily outdated, and having short time coverage. With technology advancing, new data sources on human mobility, such as Mobile Application Data (MAD), passively log people's geolocations and provide larger and more diverse datasets than traditional travel surveys. However, the resulting information on an individual level is often too sparse to be used in more advanced mobility models. Acknowledging these limitations of MAD, this thesis aims to leverage the strengths of both datasets. By combining the traditional travel survey data with the richer and more extensive dataset from mobile phones, we intend to synthesize comprehensive activity plans for those living in Sweden.

This thesis makes two key contributions. Firstly, it enhances the accuracy of identified home locations by analyzing their temporal visitation patterns and comparing them with survey data. The candidate agents whose patterns align closely with the survey data are selected based on the similarity of their temporal distributions. Secondly, it proposes a simple and transferable generative model for synthesising activity plans, which integrates big geodata and survey data. In this model, for each agent, we identify a corresponding "twin traveler" from the travel diary data. We then enrich the activity sequences of these twins with the extensive location data collected from big geodata sources over several months.

The proposed model identifies home and work as anchor locations and compares the home location with survey data to exclude unreliable ones. It then transforms user data into activity plans and applies a modified Jaccard similarity to find matching twins between datasets. Finally, it creates synthesized activity plans by combining the activity sequences of survey twins with the extensive location data from mobile app users. The resulting 113 488 synthesized activity plans are then validated against the 18 106 survey responses regarding the essential attributes of individual mobility patterns. We employ the Kullback-Leibler divergence to compare the similarities between the two datasets. The validation shows that our model generally agrees with the survey data. These results indicate that, with some future improvements, generative models combining survey and big geodata sources, as MAD in this thesis, are valuable and promising for future mobility studies.

Keywords: Human mobility, mobile application data, synthesize activity plans.

Acknowledgements

We would like to thank and express our gratitude to our supervisor Yuan Liao for her support and valuable inputs during the process of this thesis. It would not be possible to conduct this project without Yuan, and we are very grateful for the generosity of her time and expertise.

Mattias Rydström & Diana Salim, Gothenburg, 2023-12-12

Contents

List of Figures

List of Tables

1

Introduction

People participate in a wide range of activities every day, leading them to visit different locations. These sequential visits form trajectories, and by aggregating the trajectories of an entire population, valuable insights can be gained on people's mobility patterns, which is crucial in decision-making for different domains, including traffic forecasting, city planning (e.g., placing hospitals and schools), infrastructure, and tourism [1], [2]. This requires data that tracks individuals' locations daily over an extended time. Traditionally, this has been done through surveys. However, collecting this information can be expensive and time-consuming, and the resulting data can quickly become outdated [3].

In recent years, a new type of geographical data (geodata) has been introduced through anonymized Mobile Application Data (MAD). This data source is diverse, cost-effective, and readily available. It logs phone users' geographical locations (geolocations) with their consent as they interact with various mobile applications. This enables large-scale observations over longer time periods for millions of individuals at the second-to-minute level and is not restricted by geographical or administrative boundaries, as traditional surveys often are. Consequently, they are attracting more focus for the analysis of mobility patterns. However, MAD is not without its limitations. The data from MAD suffers from sparsity issues in recorded locations and population biases. For instance, capturing the geolocation from a MAD user is only possible when the user is actively using the applications. There is no guarantee that users will utilize mobile applications at every new location they visit. Therefore, data collection from MAD is inconsistent over time and across different visits, making it challenging to trace individuals' complete mobility patterns. Furthermore, there is a possibility of an overrepresentation of smartphone users among younger individuals compared to older age groups or non-smartphone users. Additionally, it is essential to consider that biases towards specific locations can also exist in MAD. For example, individuals may be more likely to use their phones while waiting for the bus at the bus stop than when visiting a restaurant.

MAD has been commonly used in traditional population models, such as in the generation of origin-destination (OD) matrices, to describe the mobility flow of individuals within a specific area of interest. However, there are less explored ways of using such data to simulate human mobility, one of which is Agent-based modeling (ABM) [4], [5]. This approach, unlike traditional models, allows for exploration of different scenarios based on policy decisions, and is capable of modeling situations

that go beyond observed data or real-world experiments. Agents are created based on information about how people typically move; with these, it is possible to simulate the flow of people in an area at different scales. To create these models, detailed and realistic activity plans are required. This conflicts with the sparsity of the MAD, where each individual's stays at locations are not entirely recorded.

In this thesis, to extend the use of valuable MAD in travel demand modeling, we design a generative model that applies traditional survey data to complement the MAD. The benefit of utilizing survey data is that it provides complete one-day activity-travel records for each participant, hence, no missing trips in the travel routes or gaps in the data. This completeness is essential for synthesizing activity plans. By combining these two datasets, the proposed generative model leverages both "small" and "big" datasets regarding their completeness and large-scale population and geographical coverage, respectively. The proposed model generates 113 488 activity plans based on agents living in Sweden, covering a typical weekday with 316 881 trips. The model captures genuine travel patterns compared to the travel survey data yet underestimates the number of activity patterns with work locations.

1.1 Background

Human mobility refers to how people travel in terms of place and time. The continued urbanization and expanded human movement have increased the complexity of predicting human mobility and led to challenges in comprehending and addressing these dynamics [6], [7]. Traditional population models, such as the OD-matrix and gravity model [8], have played a significant role, where how a population travels between geographical zones is the primary focus. Research within this field often relies on survey data, a source that comes with both advantages and disadvantages. While survey data provides a comprehensive source of information, including both travel and individuals' socioeconomic attributes, it can be costly to produce and quickly outdated. An alternative data source that has flourished in recent years, due to the increased use of smart devices, is large-scale geolocation data from mobile applications. This data is beneficial due to its flexibility, lower cost, and diversity, offering possibilities to enhance the exploration of human movement.

Another approach to studying human mobility emerged in the latter part of the 1970s when activity-based models gained attention for their ability to provide valuable insights into individuals' travel behavior and capture complex travel interactions [9]. An activity plan for a typical day includes information about when, where, and for how long an individual is engaged in various activities, as well as how the person travels between these activities. Thus, detailed data is essential for developing realistic activity plans, limiting the number of available data sources. Despite the potential of big geolocation data contributing valuable insights into individuals' travel behavior, its sparsity makes it challenging to apply in these models.

1.2 Thesis objectives

MAD consists of a large quantity of records and a variety of locations, yet it lacks the completeness required to extract activity plans due to its inability to record all locations a user has visited. To address this gap, we propose a generative model that combines MAD with the Swedish National Travel Survey [10] dataset. By leveraging the strengths of both data sources, the diversity of MAD will merge with the complete travel records of the survey data, generating a large quantity of varied and complex activity plans. The objectives of this thesis are as follows:

- Infer home and work locations and other points of interest for each MAD user.
- Match MAD users with survey participants to synthesize complete one-day activity plans from incomplete observations of the MAD.
- Evaluate the synthesized activity plans created by the generative model from various individual mobility aspects.

The inferred home locations will be validated to ensure the reliability of the estimated homes. Existing literature lacks a comprehensive validation framework for such inferred home locations, indicating an area for improvement in geolocation data analysis. Moreover, we introduce a generative model that extends the use of MAD into more advanced transport models, such as ABM. This model is transferable and easily applicable to different data sources, not limited to a specific geographical area.

1.3 Ethical considerations

The GDPR-complied data used in this thesis are generated from mobile applications where the device carriers have consented to share their locations. The data is anonymized; each device is assigned a user ID and is not connected to personal information such as a name or phone number. Nevertheless, the travel behavior and visited locations of each device can be identified, posing a risk of revealing the user's identity. Studies have shown that one to three of the most frequent visits are often enough to identify a person [11]. Hence, the violation of privacy becomes a significant concern. Taking this into account, no individual trajectories or locations will be published or used to tie back to any particular individual.

1.4 Disposition of this thesis

The remaining thesis comprises five chapters: Related work, Methodology, Results, Discussion, and Conclusion. The Related work describes the current literature on human mobility, and different models applied in this research area are outlined. The Methodology describes the data used in this thesis, the pre-processing steps, and the proposed generative model for synthesizing activity plans. The Results chapter

presents the outcomes of the model and the evaluation performance. The discussion assesses the limitations, analyzes the main findings, and outlines areas for future work. Lastly, the Conclusion provides a summary of this thesis and its findings.

2

Related Work

This chapter explores innovative approaches to utilizing geodata for comprehending mobility and addressing biases in large-scale geospatial data. Following this, we examine the distinctions between activity-based and population-based approaches in travel demand modeling, noting their unique capabilities. The focus then shifts to a more detailed exploration of activity-based models, highlighting their advantages and justifying their selection for this thesis. Subsequently, we outline methods for generating synthetic activity plans using extensive geodata. To conclude, we identify research gaps in the current literature, laying the groundwork for the generative model proposed in this thesis.

2.1 Understanding mobility using big geolocation data

Using big data to explore the movement of a population is a growing field partly due to the increasing usage of smart devices that generate large amounts of geolocation data. In urban studies, the data are used to explore how people interact with each other and move around in space and time. Studies [1], [2] use mobile phone call data to create a dynamic OD-matrix representing the population travel demand, i.e., the number of trips generated between regions in a study area. Peoples' homes, workplaces, and visited locations, and how they travel between areas of interest are classified. These studies mainly show two benefits compared to the traditional quantification of population travel demand. Firstly, using big geolocation data improves the cost efficiency and speed of data gathering compared to traditional surveys. Moreover, these emerging data sources on human mobility add a temporal aspect missing from the classic population demand derived from mobility models, e.g., gravity models. Secondly, these data enable the exploration of how the population uses amenities in a city with high spatiotemporal resolution. For instance, one study focuses on the correlation between catchment area and the size of parks in Tokyo [12], providing city planners with suggestions on where new parks or transportation infrastructure should be constructed.

Another field that has recently seen a considerable upturn in studies is the change in mobility and illness transmission in connection with the COVID-19 pandemic. One study uses big mobility data to explore the relationship between the spread of the virus and mobility in an area [13]. Other studies focus on the impact of

restrictions related to the spread of the virus [14], as well as their effects on mobility and economics in regions, taking into account factors such as income levels and population inequality [15]. What they all have in common is the utilization of large-scale mobility data to track changes in the movement of populations, considering both temporal and geographical aspects.

2.1.1 Data biases in big geodata

One must deal with selection bias and sparsity issues when working with big data on human mobility. Selection bias refers to the fact that the data used is distributed unevenly among the population [16]–[18]. Since the data is collected from people interacting with the internet throughout their day, access to smart devices is required for a person to be registered in the data. In Sweden (Table 2.1), mobile phone usage is over 90% in most age groups. However, among the elderly population, this percentage significantly decreases, and within the oldest recorded age group (75-85), it drops to 46% [19]. This disparity introduces a bias in the data towards younger individuals, and it inadequately represents the mobility patterns of elder people. Another factor contributing to bias in the data is the variance in the amount of information generated by each device. A recent study analyzing mobility data from Iraq, Congo, and Sierra Leone [16] discovered that the wealthiest 20% of the population accounted for 50% of the collected data points. This disparity in phone usage primarily stems from differences in phone usage and internet access costs.

Age group	16-24	25-34	35-44	45-54	55-64	65-74	75-85
Total mobile phone users	91%	98%	95%	94%	87%	75%	46%

Table 2.1: The distribution of mobile phone users in Sweden 2020 [19].

The impact of these biases depends on how the data is utilized. A study by Garber et al. [17] investigates whether these biases significantly affect the results and identifies scenarios where the bias may be less problematic. For instance, the study examines changes in mobility among individuals with varying economic standards during the COVID-19 pandemic. It acknowledges that a potential bias exists, with the lower-income group being less represented in the dataset due to reduced smart device usage. However, since this bias remains consistent over time, the study suggests that it does not significantly affect the observed changes in movement from before the pandemic to a few months afterward within these groups.

Inherent sparsity within big geolocation data pertains to sources such as MAD, Call Detail Records (CDRs), and location-based social networks. While these data sources are collectively abundant, the information they provide tends to be sparse individually. These sources offer geolocations from device carriers but provide a limited view of actual trajectories. For example, tracking mobility locations from CDRs is feasible only if a call has been made. Therefore, it is not guaranteed that a call has occurred in all areas the user has visited [20]–[22]. In other words, big

geolocation data is irregularly distributed in time and captures a partial view of users' mobility [21].

2.2 Activity-based vs. traditional population models

Activity-based approaches, introduced in the late 1970s, have received extensive interest [9]. The approach forecasts travel behavior and obtains new insights into how individuals allocate their time. In addition to this, activity-based models are also adopted to capture complex interactions in activity [23]. The modeling requires careful and extensive data preparation to frame the entire flow of activities and travel patterns. Such scrutiny will help identify possible inconsistencies in the data that could have been missed in a traditional population model, e.g., a trip-based approach. Commonly, surveys are carried out to collect the activities pursued by different individuals during a day or multiple days, i.e., activity plans. This provides detailed information about how, when, and where people tend to travel. Although survey data is widely used, it is time-consuming and costly to update [24]. Due to this, the number of survey participants is often small compared to the entire population, and the survey typically fetches one-day travel dairies or rarely a few days [25].

In contrast to activity-based models, traditional population models are interested in the mobility of the population in a study area. An example of this approach is an OD-matrix, which measures and aggregates the number of trips between zones to comprehend the travel demand and intensity in a study area [2], [6], [20]. Other varieties of populational models are the four-step model, the gravity model, and the radiation model [8]. However, traditional models are considered less advanced and valuable than activity-based models. This is because activity-based models have unique abilities such as (1) capturing the entire activity pattern of individuals, including the duration and intention of each activity, which gives a detailed insight into the travel behavior, and (2) simulating travel behaviors of each individual based on their characteristics [23], [26].

2.3 Activity plans from big geolocation data

In activity-based models, activity plans refer to patterns and activities performed by individuals or a population during a given time frame. The intention is to analyze mobility through geolocation data to identify typical routines and patterns across different locations. By studying and simulating activity plans, a deeper understanding of human mobility can be gained and used in, for example, traffic forecasting, transport planning, and understanding urban land use dynamics [27].

The choice of dataset is crucial when aiming to develop activity plans. Traditionally, research in this field has relied on travel surveys as the primary resource for understanding human mobility behavior. However, as outlined in Section 2.2, data

from travel surveys come with certain limitations. In comparison to traditional surveys, big geolocation data appears as a promising alternative as it offers large population and geographical coverage with high spatial and temporal resolution and is continuously updated, all while remaining cost-effective.

Synthesizing activity plans from big geolocation data is not a single straightforward task. Most previous studies have used different generative models to conduct this. However, creating reliable and useful activity plans using big geolocation data remains challenging due to data biases such as sparsity [20]. The challenges are twofold: generating reasonable routing activities, i.e., home and work (Section 2.3.1), and overcoming the sparsity of big geolocation data at an individual level (Section 2.3.2).

2.3.1 Routine activities at home and work

One crucial step in generating activity plans with big data is inferring each device's home and workplace locations. Due to human beings' circadian rhythm, temporal rules have been widely used to infer individuals' home areas. For example, Pappalardo et al. [28] studied several algorithms to detect home locations by testing different time slots. The data covered 65 individuals working at a company in Chile, and their actual home location was provided for validation purposes. The highest accuracy on home detection was found to be between 7 p.m. and 7 a.m. It is important to mention that as all participants worked during ordinary labor hours, this analysis is biased toward individuals with similar mobility routines. In contrast to this, another study [29] infers the home location by identifying the period of inactivity when one is most likely sleeping; hence, the first call after waking up, or the last call before bed, was used to detect the home location.

Similar temporal rules can be applied to estimate work location. One paper by Tongsinoot et al. [29] identifies the hours with the highest daily phone activity. The location where this occurs during the most distinct days is classified as the work location. However, the risk of using this method is that it excludes the employees who work from home and, hence, cannot be detected by this identification method. Further challenges with this approach are the workers who do not have a fixed work location [30] and people who may not use their phones at work. Additionally, workers with two phones, one personal and one for work purposes, might affect the data as the check-ins will be biased towards these locations only.

To summarize, most of the previous studies in the field of mobility data rely on temporal rules to infer home and work locations without proper validation [28]. This lack of validation is due to the challenges in obtaining ground truth data. Big geolocation data is anonymized to protect the users privacy, making it difficult to confirm a users home location. However, home and work are anchor locations and remain fundamental in understanding daily mobility [29]. Thus, along with the prevalence of big geolocation data, one should improve the validity to make these data more reliable for real-world applications.

2.3.2 Synthesizing activity plans from incomplete data

The incompleteness of big geolocation data, such as MAD, stems from the partial information obtained from each user's phone activities. The use of mobile applications can vary across spatial zones and times, leading to gaps in users' overall visited locations. This sparsity limits the usability of the data to accurately estimate travel demand [31].

To address the sparsity issue of big geolocation data for representing activity plans, it is necessary to create synthetic activity plans in combination with other data sources. Various methods exist for synthesizing activity plans; one technique is through a probabilistic approach where the probability of a device being at a specific geolocation, based on historical data, is calculated. A study conducted in 2022 [30] utilizes GPS-based survey data and uses a person's home, work, and school location as a fixed entity. These locations are then used as an anchor for where the estimated locations can be located. The result showed that people with an equally socio-demographic background have similar everyday schedules. However, the authors claim that their probabilistic model cannot capture complex human travel behavior and fails to account for non-typical mobility patterns. For instance, some people live in a rented flat during weekdays and travel back to their place or parents home on weekends. On such occasions, the second home identification is often ignored and might lead to biases in travel demand as well as generating incorrect estimations of mobility behavior.

Another way to generate synthetic data involves the application of deep learning methods such as Convolutional Neural Networks (CNNs) or Conditional Generative Adversarial Networks (CGANs) [32]. These networks produce new data based on each device's spatial and temporal patterns, creating synthesized mobility chains. However, they require rather complete datasets in individuals' mobility [24]. A study conducted in 2022 [33] developed a composite GAN with two models where the first learned to generate socioeconomic patterns of the individuals and the second generated sequential mobility data. While deep learning models appear helpful when dealing with large amounts of data in various applications, it is important to consider the limitations. For example, transferring synthesized mobility data to different regions can be complex regarding data usage and methodological framework. Training mobility data from one country and using the trained model in another may not work seamlessly since the data distribution and user behaviors may vary across regions. This mismatch can lead to reduced model performance and the need for extensive adaptation to make it applicable in different settings.

A third approach is to use a heuristic technique in generative models, which proves beneficial when dealing with complex or open-ended tasks. In contrast to the other mentioned approaches, heuristic methods do not solely rely on traditional mathematical or algorithmic models [34], [35]. The idea behind a heuristic approach is to apply and test a combination of different procedures to reach a feasible solution for the specific task of interest, as in this case, creating a generative model that is both simple and transferable. The various challenges in using MAD for generative models in transportation studies, including the generation of activity plans, have

led to the development of different heuristic approaches that prove to be effective in addressing real-world practices. For example, a study [36] focused on modeling activity-travel rescheduling and introduced a heuristic search model to cope with unexpected events. In this model, individuals responded to unforeseen situations by adjusting their activities and travel plans until no further improvements could be made in their schedules. This approach offers a practical way to adapt to changing circumstances and optimize daily activity plans using a heuristic technique.

2.4 Research gaps

Based on the literature review, we identify the below research gaps in travel demand modeling.

1. Innovative data sources:
    - Mobile application data is an emerging data source of big geolocation data on human mobility. It is ubiquitous, less explored, and not widely employed in the literature related to ABMs. However, it is cost-effective and has long-term observations of individuals. Despite the potential of this data source, it tends to be sparse as it lacks the ability to record all locations a user has visited. Still, with its large population and spatial coverage compared to traditional survey data, making it promising for creating easy-to-update and realistic activity plans.
2. Generative models of activity plans:
    - The existing literature lacks a simple yet transferable model that is more widely applicable for different data sources not limited to a single data source, such as Call Detailed Records or data collected from a specific region.
    - Temporal rules are widely used for estimating home and work locations. However, there is a limited effort to improve their validity due to the lack of ground truth data.

3

Methodology

This section outlines the thesis methodology. Section 3.1 presents the datasets used in this study. In Section 3.2, we estimate home and work locations, connect home locations with statistical geographical regions, and study bias in people's home locations. In Section 3.3, we transform the data and use the Jaccard similarity comparison to compare MAD users with survey participants in terms of their travel patterns. Subsequently, we use the activity sequence of the survey participant in conjunction with the estimated locations of the MAD user to generate activity plans. Lastly, Section 3.4 presents the statistical metrics used to evaluate our model.

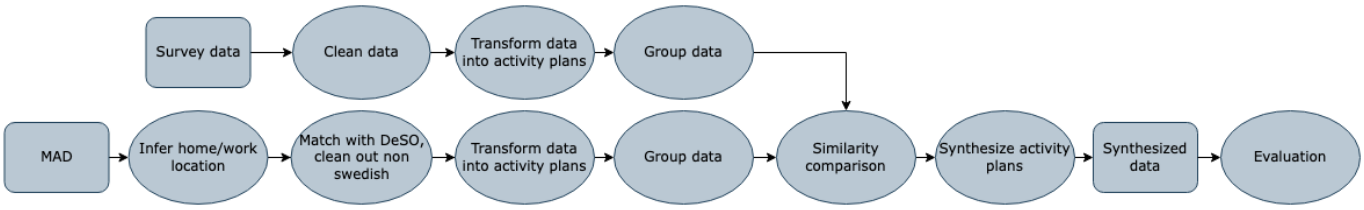

Figure 3.1: Flowchart of the analytical framework.

3.1 Data sources

In this thesis, three data sources are utilized. The first and primary is MAD, retrieved from a large group of Swedish mobile phone devices. The second dataset is the Swedish National Travel Survey from 2011 to 2016, which is used as a complement to the MAD dataset and for pattern matching and validation purposes. The third dataset is census data from the statistics agency in Sweden, providing information on demographic statistic areas.

3.1.1 Mobile Application Data

MAD consists of GPS records collected from users interacting with various mobile applications. We focus on stays, defined as specific geolocations where an individual spends a period of time, e.g., above 15 min in this study. The mobility traces used to identify stays were gathered from smartphone users in Sweden aged 18 and above. The data collection occurred from the first of June 2019 to the end of December 2019, resulting in 25 million GPS time records per day from 1 million devices. Assuming each device corresponds to one individual, this dataset represents approximately

10% of the Swedish population. Nevertheless, it is essential to note that a single individual may possess multiple devices.

The MAD, as provided, underwent two pre-processing steps [37]. Firstly, stays were identified using the Infostop algorithm, chosen for its robustness against large datasets and measurement noise and its ability to conduct a multi-user analysis simultaneously [7], [37]. Detecting stays is essential when measuring movement and studying mobility data. To validate and distinguish a stay from transient locations when the user moves, we set a minimum duration of 15 minutes and require additional geolocation to be recorded within three hours. Furthermore, the maximum allowed distance between two consecutive records had to be within 30 meters to be treated as the same stay.

Lastly, a three-step data filtering process was implemented to enhance data quality: 1) dropping stays above 12 hours, 2) requiring identified individuals to have more than seven active days, where an active is defined as a day with at least one recorded stay, and 3) ensuring the number of unique locations exceeded two.

Attribute	**Description**
User ID	Unique ID of the device
Local time	Timestamp for when the user ID was first shown at a unique location
Location	User-specific ID of the location
Hours start	Time of the day, in hours and minutes
Duration	Time between arrival and departure
Latitude	Latitude of the location
Longitude	Longitude of the location

Table 3.1: Attributes in the MAD.

Additionally, stays during weekends and holidays[1] were excluded from the final MAD as this thesis focuses solely on synthesizing mobility data for regular weekdays. Table 3.2 shows the records in the applied used dataset:

Total rows	Total users	Median rows per user	Median active days per user	Median unique locations visited per user
13 493 110	322 477	22	14	5

Table 3.2: Overview of the statistics in the MAD, after pre-processing.

3.1.2 Swedish National Travel Survey

The second dataset, the Swedish National Travel Survey (2011-2016), is provided by Transport Analysis [10], the official statistics authority of transport and communication in Sweden. In the survey, the complete activity pattern of participating individuals is extracted for weekdays, with holidays and weekends excluded for this

[1]Summer holiday 23/6-10/8, Christmas vacation 22/12 until the end of December.

thesis. It is important to note that the survey relies on one-day travel diaries, capturing a snapshot of individuals' activities on specific weekdays, thus offering a detailed and focused perspective on daily routines and travel behavior.

The original survey data are all the trips reported by the survey participants on their record days. After removing incomplete records regarding trip purposes, origin, destination information, etc., we continue transforming the trip-level data into activities. We keep the individuals where their recorded day starts and ends at the same place, as required by common transport agent-based simulation, e.g., MATSim [38].

For this thesis, the transformed survey dataset serves two purposes: 1) the input to the generative model for synthesizing activity plans of MAD users and 2) the ground truth data for comparing the results of the synthesized activity plans regarding essential activity patterns. The dataset includes a large amount of information about each individual's activities and mobility. All the relevant information from the Travel Survey used in this thesis is summarized in Table 3.3.

Attribute	**Description**
Participant ID	Unique ID of the survey participant
Activity	Home, work[2], or other location
Start time activity	Timestamp for when the activity started
End time activity	Timestamp for when the activity ended
Duration	Time between arrival and departure
Hour start	Start time of the activity (in minutes from 00:00)
Hour end	End time of the activity (in minutes from 00:00)
Zone	The DeSO area (3.1.3) of where the activity occurred
Distance	Self-reported travel distance, previous to current location
Commute	Self-reported distance between home and work

Table 3.3: Attributes in the dataset for the Swedish National Travel Survey.

Two cleaning steps are implemented in the survey dataset. In cases where participants have not reported one or more distances traveled, and the trip is between home and work with a commuting distance reported, this distance is added. Even after the commuting distance is added, participants who still have one or more missing distances are then dropped. Additionally, users outside the ten most common activity sequences are excluded, as indicated in Table 3.4. This step removes the risk of the model overfitting unusual activity sequences while also ensuring that the sizes of each activity sequence group are large enough to be reliable in the evaluation. Following this cleaning step, there are 18 106 survey participants for analysis.

[2]This also includes school activities.

Sequence	Percentage of survey participants	Cumulative percentage of survey participants
H-O-H	23.64 %	23.64 %
H-W-H	18.99 %	42.63 %
H-O-O-O-H	8.44 %	51.07 %
H-O-O-H	7.36 %	58.43 %
H-H	7.28 %	65.71 %
H-W-W-O-H	5.23 %	70.94 %
H-W-W-H	5.19 %	76.13 %
H-O-O-O-O-H	3.64 %	79.77 %
H-O-O-O-O-O-H	2.39 %	82.16 %
H-W-W-W-H	2.36 %	84.52 %

Table 3.4: The top ten activity sequences, representing 84.52% of the total survey data.

3.1.3 Demographic Statistical Regions of Sweden

Demographic Statistical Regions of Sweden (DeSO) is a dataset provided by the Swedish Statistics Agency [39], and it categorizes Sweden into 5 984 DeSO zones, with populations ranging from 700 to 2 700 inhabitants. The dataset includes socioeconomic data about the population in each zone. However, for the purposes of this thesis, only two key aspects are considered; the total number of inhabitants in each zone and whether the zone is located in a rural (A), suburban (B), or urban area (C) with the original code digit in the brackets.

Area type	Share
Rural	15.40 %
Suburban	8.71 %
Urban	75.89 %

Table 3.5: The distribution share of the three area types.

3.2 Inferring home and work locations

To estimate home and work locations[3] of our MAD users, we apply the temporal rule proposed in [28]. This rule recognizes the most visited location during weekday nighttime (7 p.m. to 6:59 a.m.) as the estimated home location for each individual. For estimating work[4] location, the first step is removing the estimated home locations for each user, as having the same location for both home and work would complicate the final results. Subsequently, the opposite hours, minus the user's commute time, are examined, leaving a time span of 9 a.m. to 4:59 p.m. as the period of

[3]To ensure privacy, the location data used is at the zone level with a spatial resolution of approximately 100 meters, and it is not employed to identify any specific individual.

[4]School locations are also included, as it is a fixed location visited during the day. The same definition is applied for the survey data.

interest for being at work. The location with the highest frequency of stays during weekdays in this period is considered the work location. Additionally, two kinds of filtering are applied to enhance the validity of the estimated home and work location. The location must be stayed at for at least five times, with a total duration of 300 minutes. Locations that meet these criteria are included, and MAD users without an estimated home location are discarded.

To further improve the quality of estimated home locations, the temporal visitation pattern of each user's home location is compared with the visitation pattern of home locations in the survey. A day is divided into 30-minute intervals. Following this, the home location is analyzed by counting how many times the user remains at this location during these intervals. Each interval is then normalized by dividing it by the interval with the highest number of stays, resulting in a score between zero and one. A score of one indicates that the time slot is among the most common times for the person to stay at the home location. The same calculation is applied to the survey data with a modification. Instead of analyzing each individual separately, the visitation frequency aggregates all participants, treating them as repeated observations of one virtual individual. The combined visitation pattern is then used as the ground truth.

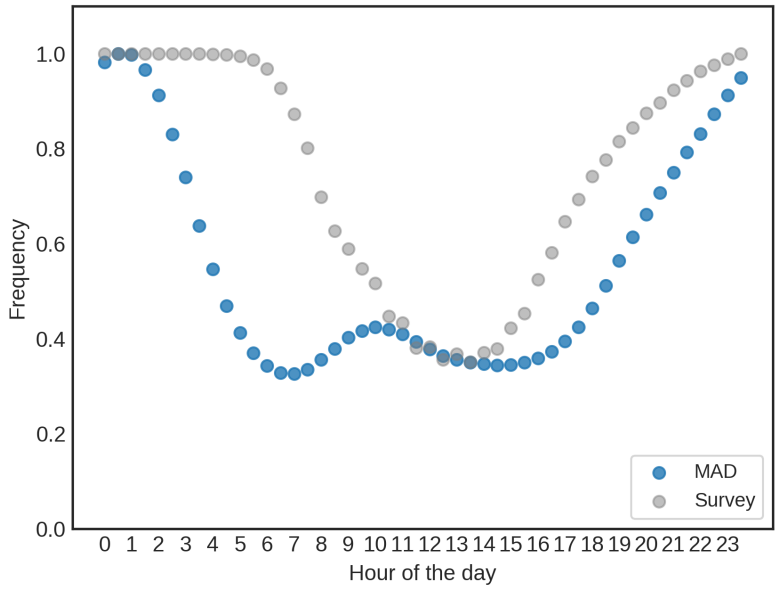

Figure 3.2: The temporal visitation pattern of the home location for MAD users compared to survey participants.

After that, the visitation frequency patterns for MAD users are compared with the travel survey. This is done by calculating the Euclidean distance between them, indicated by Equation 3.1, where for being at home, p represents the frequency value for the MAD user, while q signifies the combined frequency value for the survey data, both within each of the n 30-minute intervals ($n = 48$). A short Euclidean distance indicates a similar visitation pattern, providing a stronger conviction that it is an actual home location found. Conversely, a longer distance suggests that the

identified home might not be correct.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{3.1}$$

Following the calculation of distances for the identified home locations of the MAD users, each user's distance is sorted in ascending order. Subsequently, 20 groups are created based on the sorted distances, with 5% of the users in each group. For instance, group 1 consists of the 5% with the shortest distance, group 2 the 5% with the second-shortest distance, and so on. The average Euclidean distance for each group is then plotted on a graph to identify any distinct elbow point where the distance increases significantly compared to the previous group [40]. Users in groups where the average distance is higher than this identified point are excluded from further analysis. The exclusion is based on the rationale that the increased distance for these users signifies that the found homes have a less similar visitation pattern compared to the survey data. Consequently, the identified home locations are considered less reliable and are excluded from further analysis, ensuring a higher quality in the estimated home locations still present in the data.

Lastly, after the home location is inferred, information is added noting which census zone (DeSO area) the home belongs to, based on the location GPS coordinates. This is done for two reasons: 1) As the focus of this thesis is on individuals in Sweden, if the estimated home location falls outside a DeSO area, it implies the location is outside Sweden, prompting the removal of the user. 2) Add information on which type of area the estimated home location is; rural, suburban, or urban. This information will be used for creating synthetic activity plans.

3.3 Generative model of individual activity plans

The model consists of three steps to synthesize activity plans from MAD users. Firstly, the model divides the day into 48 groups at 30-minute intervals each for both MAD and the survey data. These intervals record the locations visited - categorized as home, work, or other. Secondly, a match is made between the individuals in the MAD and survey data to find the activity sequence twin from the survey participants with the most similar activity pattern. Lastly, the location coordinates of the MAD users' home, work, or other locations are simulated based on the activity sequence of their identified survey twin, resulting in synthetic activity plans for the MAD users.

3.3.1 Transform data into activity plans

The information from both datasets is standardized into the same structure to facilitate the comparison between MAD users and survey participants, shown in Equation 3.2. This involves evenly dividing the day into 48 groups at 30-minute intervals (i) in each and recording the type of location each user visits during each interval. The available location types include home, work, or other (h, w, o), as these categories are commonly used for transport agent-based simulation.

$$\begin{aligned} \mathbf{S} &= \{l_i \mid i = 1, 2, ...48\}, l \in [h, w, o] \\ \mathbf{M} &= \{(h, w, o)_i \mid i = 1, 2, ...48\} \end{aligned} \tag{3.2}$$

where $h + w + o = 1$, indicating that at each time slot, the frequency of observing a MAD user being at home, work, and other locations are normalized into 1.

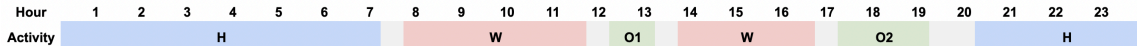

Figure 3.3: Example activity record for a survey participant.

Hour	1			2	...	7			8	9	...	15			16	17	...	22			23
Home	1	1	1	1		0.8	0.8	0.6	0.3	0.1		0	0	0.3	0.3	0.3		0.7	0.7	0.7	0.9
Work	0	0	0	0		0.1	0.1	0.3	0.5	0.6		0.9	0.9	0.5	0.4	0.3		0	0	0	0
Other	0	0	0	0		0.1	0.1	0.1	0.2	0.3		0.1	0.1	0.2	0.3	0.4		0.3	0.3	0.3	0.1

Figure 3.4: Example of a probabilistic activity record for a MAD user.

For survey participants, only one location (l) is typically visited during each interval, given the survey is a one-day travel diary. However, since the day is split into 30-minute intervals, a participant could travel from one location and arrive at the next within one interval. In these cases, the location where the participant spent the most time is selected. If a participant travels throughout the 30-minute interval, the location is unknown with a NaN value assigned. An example of a survey participant activity plan can be seen in Figure 3.3.

For the MAD users, given that data is collected over several months, there is increased variability for each interval. Rather than having one primary location type at each time slot, the probability of each location type for each slot is recorded $(h, w, o)_i$. For example, if a person with ten stops during a 30-minute interval visited their estimated home location five times, work three times, and other locations twice, the probabilities of activity participation would be Home: 0.5, Work: 0.3, Other: 0.2, exemplified in Figure 3.4. Due to the uncertainty in duration for MAD, consideration is not given to the number of minutes spent, as each visit is treated equally. NaN is assigned as the value if no location is visited during an interval.

3.3.2 Search for activity plan twins

To establish a similarity between the survey participant and the MAD user that could be paired up, we organized them into groups based on three variables: home location, commuting distance between home and work, and average travel distance between home and other locations. Each of these variables features multiple levels. A unique combination of levels across these variables identifies a group.

In the distance categories, each variable is divided into quantiles based on the average travel distance to the activity in question. For example, a survey participant

in the "Short" group in the work variable belongs to the 0-25% with the shortest commuting distance. "No Information" is recorded if the individual has not traveled to a particular activity type. For the survey participants, only one activity sequence, H-W-W-O-H, containing just 5.23% of the participants, has distance information for both distances to the other and work locations. In contrast, for MAD users, the number of individuals having both distances is 37.04%. To address this discrepancy, MAD users with distance information to both work and other locations can match with survey participants missing either of them. In other words, these MAD users can be matched with three groups of survey participants; those with distance information for both work and other locations, those with distance information only for work location, and those with distance information only for other location.

- If the home location is in an urban or non-urban (suburban/rural) area (extracted from the census zone information)[5].
- The distance between home and work location. Short, Short-Medium, Medium-Long, Long, No Information[6].
- The distance between home and other locations. Short, Short-Medium, Medium-Long, Long, No Information[7].

After grouping, we pair survey participants with MAD users with similar activity patterns within each group. These "twins" are identified using a Jaccard similarity comparison, which assesses the similarity between two objects based on categorical variables [41]. For a given time interval, if the two objects have the same categorical variable, a count of one is included; otherwise, nothing is added. The combined count is then divided by the number of variables examined. The resulting score ranges from 0 to 1, where 1 indicates an exact match between the two objects, and 0 signifies no shared similarities.

In our case, the comparison is modified by replacing categorical variables with a probabilistic variable for the MAD users, as mentioned in 3.3.1. In each of the 48 time slots, the user has the probabilistic location information (unless a NaN is recorded for the MAD user, then the time slot is disregarded). The score added is the probability of the MAD user being in the location the survey participant visited during that slot. For example, if the survey participant is at home and the MAD user's probability of being there is 0.2, this value is added to the final calculation.

In Equation 3.3 the calculation is presented, with S denoting the survey participant and M representing the MAD user. The probability value is added from all time intervals where the participant and the user match. This aggregated score is divided by the total number of intervals where we have data on the MAD user, $n - M_{nan}$, where n notes the total number of intervals (48) and M_{nan} indicates the number of intervals for the MAD user without any stays. The result is recorded, and the

[5]For MAD, 81.63% of users have homes located in urban areas, compared to 75.25% for the survey participants.

[6]Distance categories based on the quantile of travel distance for each dataset. Short 0-24%, Short-Medium 25-49%, Medium-Long 50-74%, Long 75-100%.

[7]See above footnote.

survey participant with the highest score is identified as the twin.

$$J(\mathbf{S}, \mathbf{M}) = \frac{|\mathbf{S}_h \cap \mathbf{M}_h| + |\mathbf{S}_w \cap \mathbf{M}_w| + |\mathbf{S}_o \cap \mathbf{M}_o|}{|n - M_{nan}|} \tag{3.3}$$

3.3.3 Synthesize activity plans

To generate the synthesized activity plans, we start with the twin traveler's activity sequence and the MAD users' recorded locations. The Home and Work in the sequence are directly linked with their identified home and work locations. However, for Other, it is not as simple to directly link a location from the MAD. Each MAD user could have more or less other locations than the activity sequence. To solve this, we apply a temporal matching method. The day is divided into seven-time groups: Night (00.00–05:30), Early Morning (05:31–08:30), Late Morning (08:31–11:30), Lunch (11:31–13:30), Afternoon (13:31–17:30), Early evening (17:31–20:00), Late evening (20:01–23:59).

The categorization of a visit into one of these groups is based on its midpoint time. The probability P of visiting a location within each time group is recorded, shown in Equation 3.4,

$$P(o_k, j) = \frac{V_{k,j}}{V_{\text{total},j}} \tag{3.4}$$

where o_k is the other location, j denotes the time group ($j = 1, 2, ..., 7$), and V is the number of visits by the MAD user. Then, for each Other in chronological order during the day in the activity sequence, a weighted random selection is performed based on the probability recorded ($P(o_k, j)$). If no other location is available during a specific time interval in the activity sequence, a location from a different time group is randomly chosen. After each selection, this chosen other location is excluded from future selection and cannot be reused. In scenarios where the MAD user has visited fewer other locations than required by the activity plan, locations cannot be selected for all stays, and the travel pattern would be impossible to evaluate. In these cases, the synthesized activity plans are dropped.

Once all the locations are linked to the activities in the twin's sequence, the displacements between the location coordinates are gathered and then multiplied by 1.5 [42] to approximate the distance traveled. In the case of round trips, where the activity plan goes from Home to Home or Work to Work, the distance is directly assigned based on the survey twin's data. With this final step, the generation of activity plans is completed.

3.4 Evaluation of the generative model

To assess the proposed generative model, this thesis considers comprehensive attributes of activity patterns in human mobility and compares them between the survey data (treated as the ground truth), and the synthesized activity plans. The following attributes will be evaluated:

(1) The number of trips. The difference in percentage for each trip count, from one to six daily trips.

(2) The distance between activities' locations. The travel distance data are clustered into twelve different groups based on the traveling distance in kilometers.

(3) The top ten activity sequences. The share of users for the selected top ten activity sequences in the survey data with the corresponding user shares in the synthesized activity plans.

(4) The temporal visitation pattern for each activity type, in half-hour intervals.

(5) The time spent on each activity type.

For the metrics (1)–(4), we use the KL divergence indicator, which is a distance metric designed to quantify differences between two probability distributions. KL divergence yields a value equal to or greater than zero and informs how the datasets differ in terms of their probability distributions across the selected categories [43], [44]. A smaller KL divergence implies that the datasets are more similar, while a larger value indicates greater dissimilarity. The mathematical formula for the KL divergence between two probability distributions, P and Q, is defined as:

$$KL(P||Q) = \sum [P(x) \log \left(\frac{P(x)}{Q(x)} \right)] \tag{3.5}$$

where $P(x)$ represents the probability of an event x according to the first distribution, P, while $Q(x)$ represents the probability of the same event according to the second distribution, Q.

In this thesis, the KL divergence serves as a critical tool to examine if the synthesized activity plans share the same distribution as the survey data for the above metrics (1)–(4), providing valuable insights into the validity of our modeling approach and results. Given the KL divergence is suitable for categorical data, for a continuous variable like (5), we examine the average time spent at home, work, and other locations and visually compare the synthesized activity plans with the survey data.

4

Results

In this chapter, Section 4.1 introduces the inferred home and work locations. Section 4.2 presents the results of identified twin travelers and the corresponding synthesized activity plans. Lastly, the evaluation of the model is introduced in Section 4.3.

4.1 Home and work locations

One of the contributions of this thesis is to improve the validity of identified home locations by investigating their temporal visitation patterns in comparison with the survey data (as the ground truth). MAD users with patterns similar to the survey data are retained by measuring the distance between their temporal distributions.

After the initial inferring of home locations based on the temporal visitation rules, the temporal visitation pattern comparison was applied. To identify a distance cut-off point for improving the validity of identified home locations, we examine the elbow point in Figure 4.1, where longer distances correspond to less reliable identified home locations. While the average distance in each group increases at an even pace for most, a noticeable jump is observed between the second-to-last, and the last group. Figure 4.2 illustrates the difference in the visitation patterns for the selected groups, accentuating the increased distance between the groups with the highest, (d), and second-highest distance, (c). The visitation pattern between the middle, (b), and the second-highest distance group, (c), stays relatively consistent, but in the group with the largest distance (d), a clear change in the visitation pattern is observed.

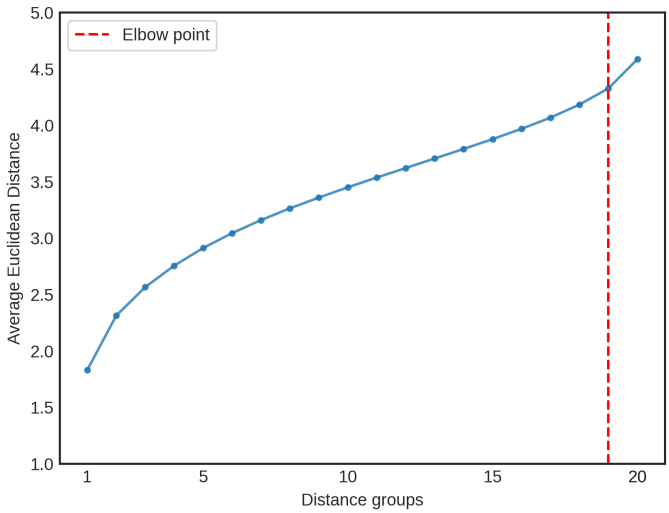

Figure 4.1: The average Euclidean distance of the temporal visitation pattern for each 5% group of MAD users compared with the survey data. The groups are ordered from the most similar to the least from 1 to 20.

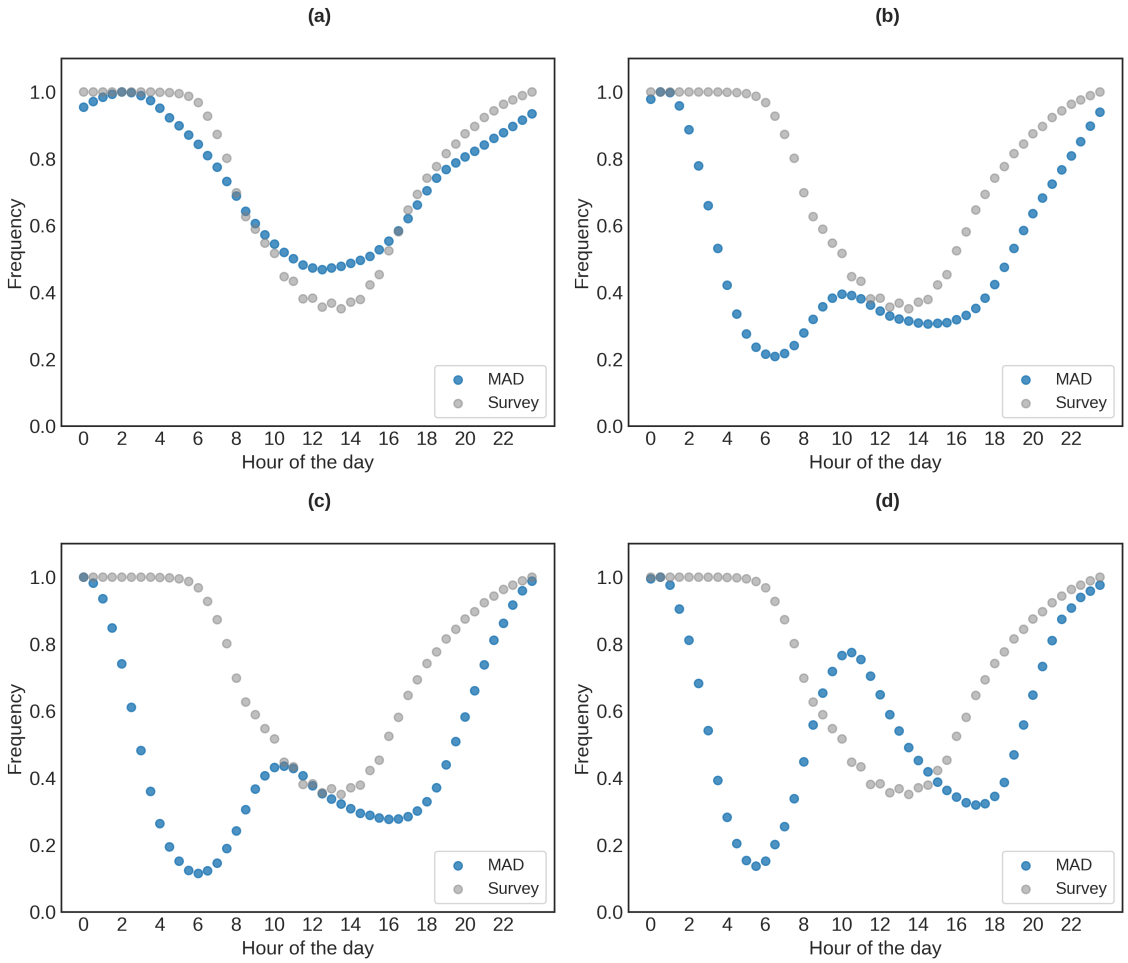

Figure 4.2: Comparison of aggregated user visitation patterns. (a) Distance group 1. (b) Distance group 10. (c) Distance group 19. (d) Distance group 20. The longer the distance, the less reliable the identified home locations for these groups.

After keeping those MAD users who have more reliable home locations identified, we assign DeSO zones to each of them. Home locations outside Sweden are dropped, and the resulting correlation between the number of MAD users in a DeSO zone compared to the actual population is presented in Figure 4.3. After these screening steps, the number of available MAD users drops from 322 477 to 185 008, resulting in a 42.63% reduction.

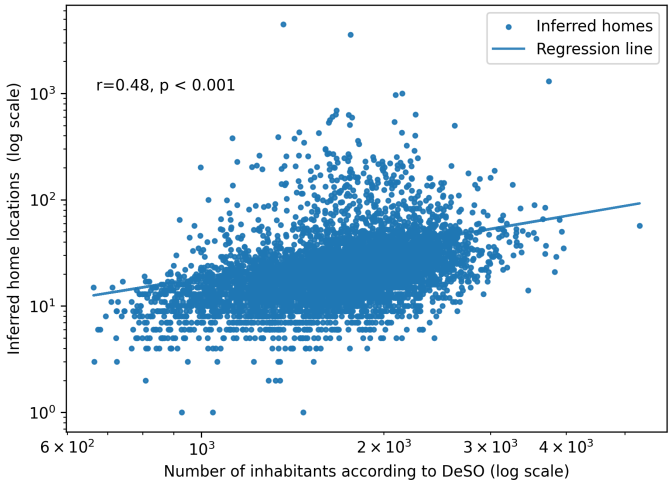

Figure 4.3: The number of inferred home locations compared with the actual population in the DeSO zones, log scale.

Out of the remaining MAD users, 70 533 users have work locations identified from their trajectories. The percentage of MAD users associated with a work location closely mirrors the survey data, as illustrated in Table 4.1. The median commuting distance is also similar; however, the average distance is much larger for the MAD users. In Figure 4.4, it is clear that the two groups have a similar distance in the commute for the first 50%, but for the MAD users, this distance becomes significantly larger for the third quantile compared to the survey users. Examining the even longer commuting distances, almost 17% of the MAD users travel over 100 km, while this number is just 0.68% for the survey participants. The maximum commuting distance reported in the survey data is 800 km. In the MAD, 1.39% of users had commutes exceeding this distance, with the longest recorded commute being 12 653 km. The long tail for the commuting distance of the MAD users suggests variability in the quality of inferred work locations. Further discussion on this topic will be provided in Section 5.2.

Dataset	% of users with a work location	Median commute distance	Average commute distance	Users with 100 km< commuting distance
MAD	38.12 %	10.57 km	91.46 km	16.97 %
Survey	38.06 %	8.00 km	15.13 km	0.68 %

Table 4.1: Commuting statistics. For the survey, the % of users with work location refers to the percentage of participants that have traveled to a work location during their reported day. For MAD users, it is the percentage of users with an inferred work location.

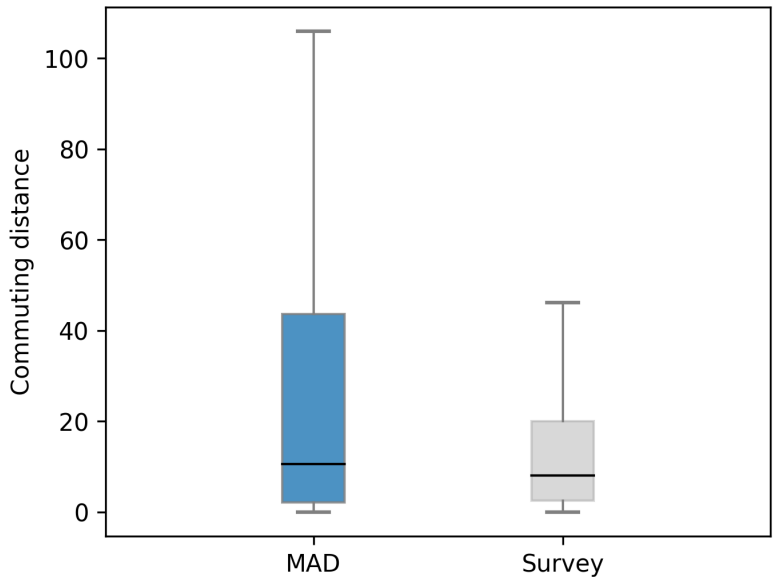

Figure 4.4: The commuting distance for MAD users compared to the commuting distance for the survey participants.

4.2 Twin travelers and synthesized activity plans

Besides the improved quality of identified home locations using big geodata, the other contribution of this thesis is the generative model that combines MAD and survey data. Specifically, for each MAD user, we search for twin travelers in the travel diary and enrich these twins' activity sequences with MAD users' abundant locations spanning over months.

There are 185 008 MAD users who are linked with a survey twin based on the modified Jaccard similarity score evaluating their activity patterns' similarity. As stated in Section 3.3.3, the activity sequence of the survey twins is filled with locations from the MAD user. Matches where the MAD user have fewer other locations than required in the paired up activity sequence are dropped, leaving 181 143 users for further analysis. For these remaining users, the distribution of their similarity scores in comparison with their twin travelers from the survey is shown in Figure 4.5. It is evident that, even among the twin travelers identified as having the most similar activity patterns to those of the MAD users, some exhibit low similarity scores, indicating a reduced reliability in the pairing process.

We further investigate the temporal visitation patterns of these MAD users of different similarity scores. Figure 4.6 illustrates how users with a low similarity score (c) compare to those with a higher score (d), all MAD users (a), and survey participants (b). MAD users with a highly similar twin traveler exhibit more distinct patterns across the three activity categories. And their activity patterns are more similar to the survey participants.

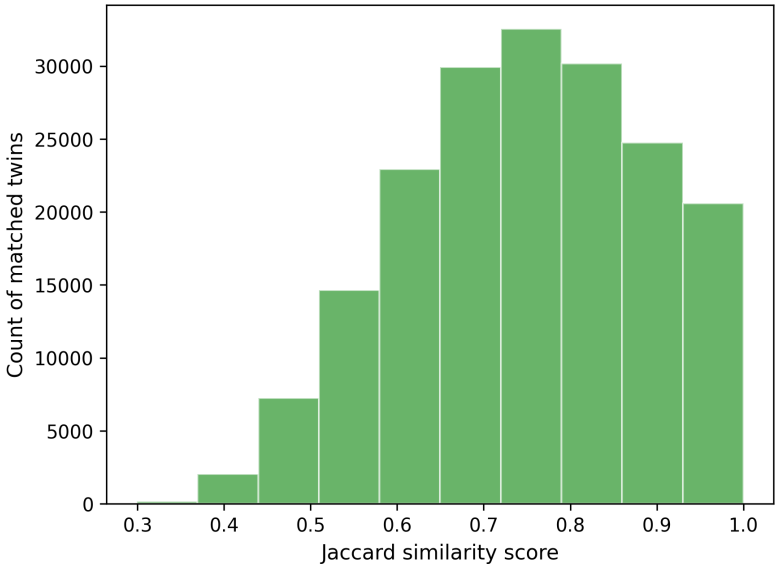

Figure 4.5: The distribution of the similarity scores between the MAD users and the survey twins.

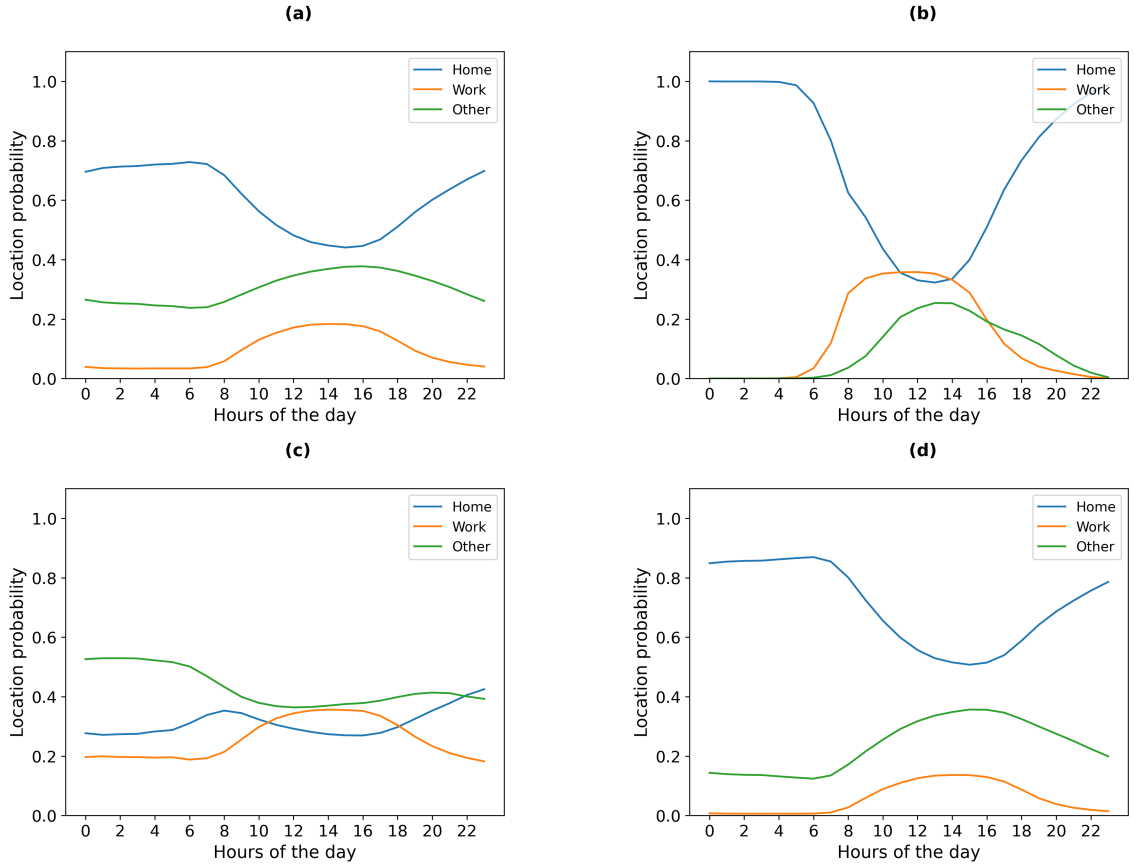

Figure 4.6: Comparison of the probability of visitations for Home, Work, or Other for MAD users vs. survey. (a) All MAD users. (b) Survey participants. (c) MAD users with below 0.5 similarity score with their matched twins. (d) MAD users with above 0.7 similarity score.

To improve the quality of synthetic activity plans for MAD users, we explore the impact of matching similarity on the model performance. The evaluation is carried out at every interval increase of 0.1 in the similarity score, ranging from 0.5 up to 0.9, where the KL divergence results of our evaluation metrics are recorded, shown in Table 4.2, and further illustrated in Figure 4.7. The evaluation result with a cut-off point of 0.7 proved to be the most successful, and we kept the users with a similarity score above this threshold. This leaves us with 113 488 individuals with synthesized activity plans for our final evaluation, as summarized in Table 4.3. One observation from the impact of similarity on the evaluation process is that the more stringent cut-off points of 0.8 and 0.9 significantly decrease the KL divergence results. This may be linked to overfitting, with very high similarity scores being connected to simplistic activity plans, that do not represent the entire survey dataset.

Groups of similarity cut-off:	All	> 0.5	> 0.6	$>$ **0.7**	> 0.8	> 0.9
Total trips	0.066	0.061	0.050	**0.034**	0.035	0.145
Total trip distance	0.269	0.265	0.259	**0.253**	0.279	0.375
Average trip distance	0.283	0.278	0.269	**0.261**	0.277	0.333
Activity sequences	0.134	0.136	0.143	**0.164**	0.273	0.673
Average temporal distribution	0.110	0.108	0.102	**0.090**	0.083	0.090
Average KL divergence	0.172	0.170	0.165	**0.161**	0.189	0.323
Number of activity plans	181 143	173 150	151 288	**113 488**	68 170	28 522

Table 4.2: KL divergence values for different cut-off points of similarity score. For the average temporal distribution value, we averaged the results from the three calculations made; home, work, and other. The bottom two rows show the average KL divergence for the column and the total number of synthesized activity plans used.

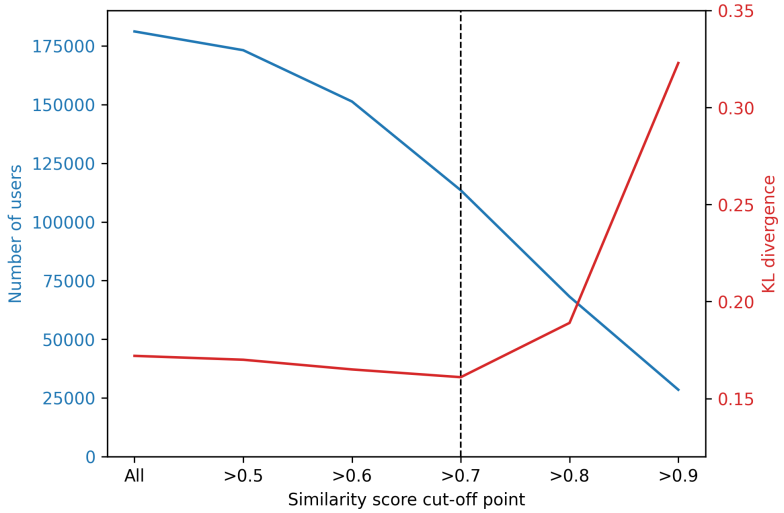

Figure 4.7: A visualization of Table 4.2. The red line shows the average KL divergence and the blue line presents the number of activity plans still available after each cut-off point.

Number of synthesized activity plans	Of which have a work location	Of which have one or more other locations
113 488	19 933 (17.56%)	90 970 (80.16%)

Table 4.3: Total number of synthesized activity plans, the number of plans that have a work location, and the number of plans that have at least one other location.

The number of connected MAD users varies significantly among survey participants. There are 18 106 survey participants in our applied dataset. Of these, only 4 309 participants are identified as twin travelers for at least one MAD user, resulting in the exclusion of 76.20% unused survey participants. The distribution of MAD users per survey participant is heavily skewed to the right, as seen in Figure 4.8. Most participants match with a handful of MAD users, but the tail extends far, with 123 participants associated with over 100 MAD users. Of these, ten participants are linked with over 1 000 MAD users each, and the participant assigned to most MAD users reaches 7 157.

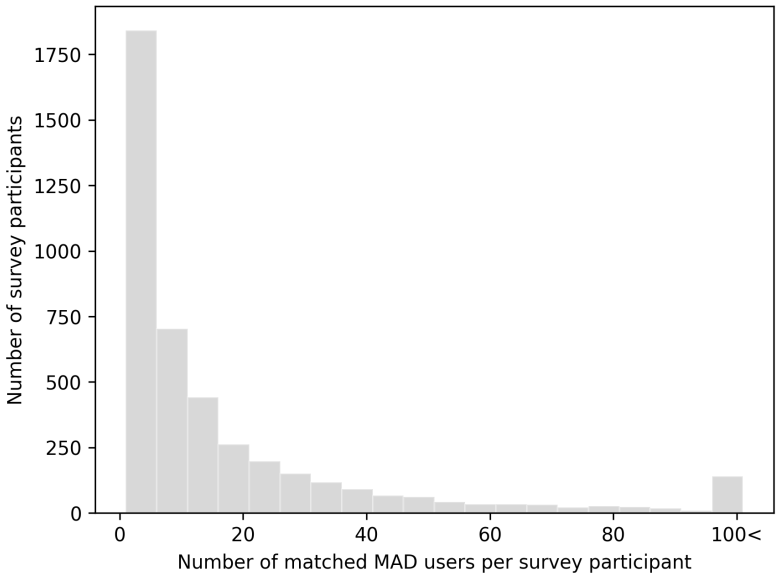

Figure 4.8: The number of MAD users associated with each survey participant. The graph only includes survey participants with at least one connected MAD user.

4.3 Evaluation of the synthesized activity plans

This part presents the evaluation results for each attribute described in Section 3.4. The KL divergence for these are shown in Table 4.4. The closer the score is to zero, the more similar the distributions are between the survey data and synthesized activity plans for the specific evaluated aspect.

	KL divergence
Total trips	0.034
Total trip distance	0.253
Average trip distance	0.261
Activity sequences	0.164
Temporal visitation home	0.036
Temporal visitation work	0.166
Temporal visitation other	0.069

Table 4.4: Model performance indicated by KL divergence. A smaller KL divergence indicates a high similarity, hence better model performance in certain aspects.

The KL divergence for total trips is the lowest among the various aspects, indicating a close similarity between the distribution of the number of trips in the survey data and the synthesized activity plans. Conversely, average trip distance accounts for the highest KL divergence in Table 4.4. While the value of 0.261 is not far from zero, it still reflects notable differences in the distribution of trip distances, which will be discussed further in the upcoming sections.

4.3.1 Total trips

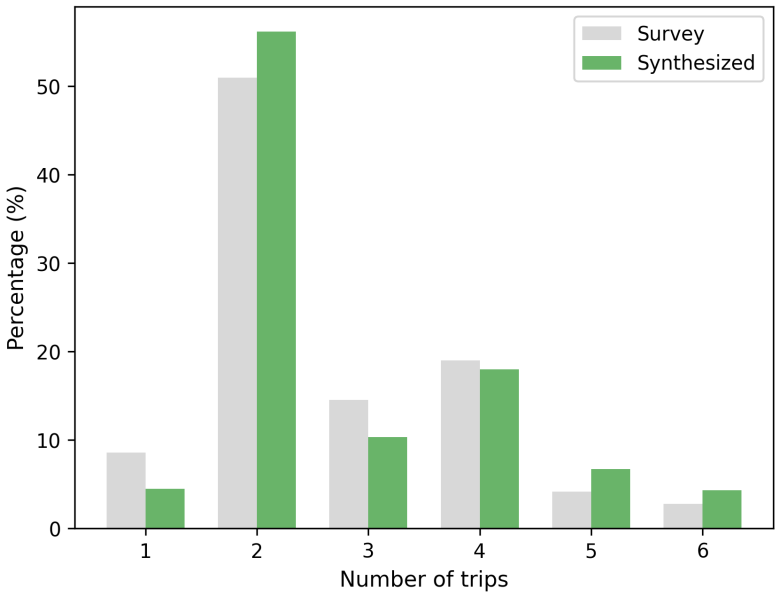

Figure 4.9: Total trip distributions for both the survey data and synthesized activity plans.

The distribution of total trips in the two datasets appears to be fairly even. As illustrated in Figure 4.9, there are no large differences between the distributions for the survey and synthesized data. This observation is further supported by the computed KL divergence of 0.034, indicating high similarity between the distributions. Illustrated in the histogram, approximately 50% of the total trips in the datasets

fall into having two trips. The other two most common trip counts, four and three, are consistent across both datasets. The average total trips is 2.67 for the survey data and 2.79 for the synthesized, indicating a slight tilt in the synthesized data towards a higher number of trips.

4.3.2 Trip distance

The histograms in Figure 4.10 display two different distributions. On the left, the graph represents the total distances traveled. For very short trips (0-5km), the survey data accounts for 24%, and the synthesized activity plans account for 26%. Conversely, for very long trips (>55km), the survey data represents 16%, while the synthesized data represents 45%. In the distances between very short and long trips, there is an apparent underrepresentation in the synthesized data, with the survey data dominating these distance groups. On the other hand, in the average distance trips, the share of very short trips increases to 45% for the survey data and 36% for the synthesized data. However, the opposite trend is observed for very long trips, with the shares decreasing to 4% in the survey data and 30% in the synthesized.

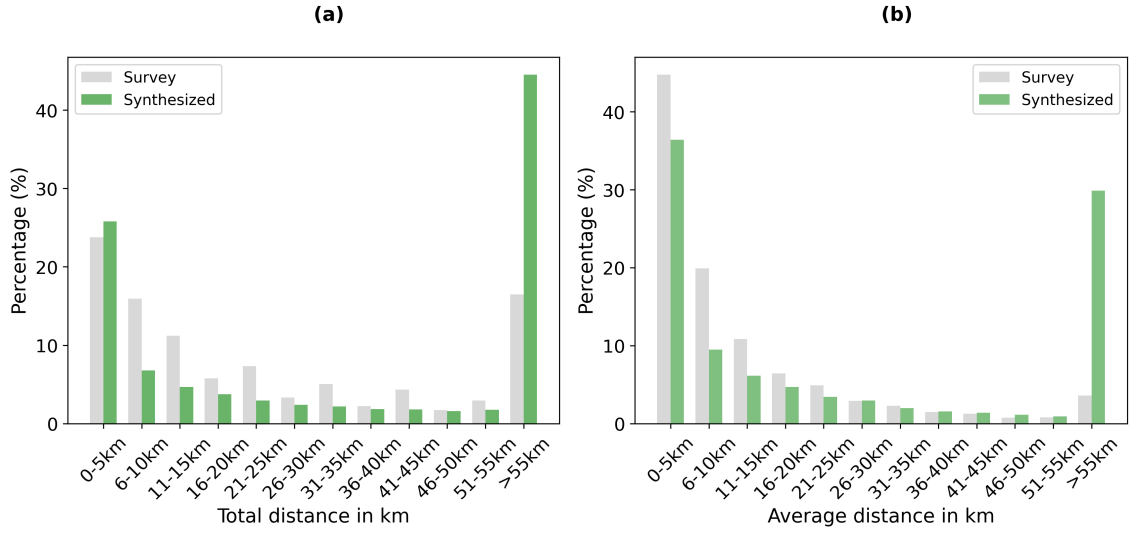

Figure 4.10: Distribution of users' total trip distance (a), and average trip distance (b), during their recorded day.

In general, synthesized activity plans overly represent long-distance trips compared to the survey data. According to the KL divergence presented in Table 4.4, the performance of this metric is 0.253 for total trip distance and 0.261 for average trip distance. This indicates that the datasets are not a perfect match but that the synthesized activity plans still capture a significant portion of the survey data's characteristics.

4.3.3 Activity sequences

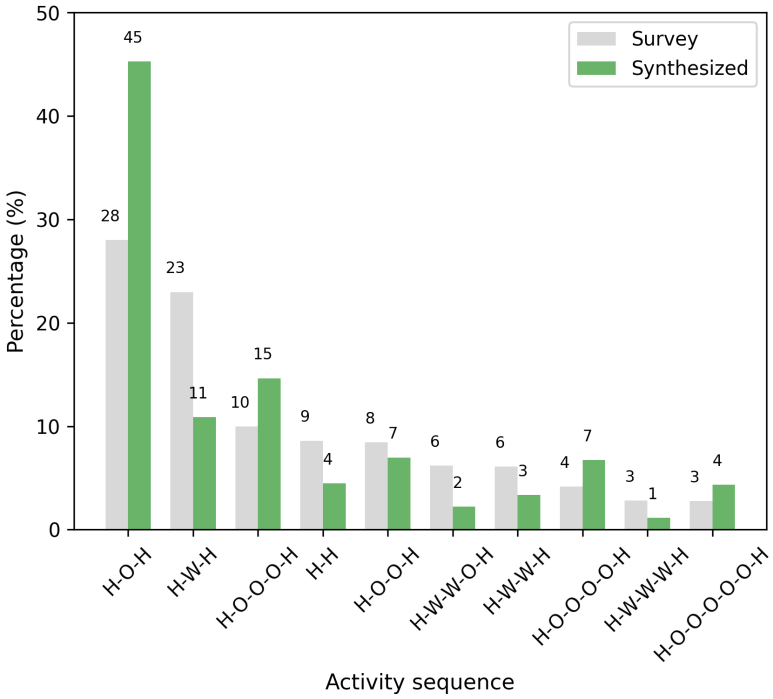

Figure 4.11: Distribution of the activity sequences.

Sequence	Percentage Survey	Percentage Synthesized
H-O-H	28.00 %	45.29 %
H-W-H	22.99 %	10.89 %
H-O-O-O-H	9.99 %	14.64 %
H-H	8.59 %	4.49 %
H-O-O-H	8.46 %	6.98 %
H-W-W-O-H	6.18 %	2.21 %
H-W-W-H	6.09 %	3.34 %
H-O-O-O-O-H	4.14 %	6.71 %
H-W-W-W-H	2.81 %	1.12 %
H-O-O-O-O-O-H	2.76 %	4.33 %

Table 4.5: Percentage distribution of the activity sequences.

Figure 4.11 illustrates the distribution of the ten chosen activity sequences. The most significant disparity between the survey data and the synthesized activity plans is observed in the "H-O-H" sequence, with survey data at 28% and synthesized activity plans at 45.29%. Generally, the synthesized activity plans tend to have a higher percentage share in activity sequences containing Other. Among the six sequences with at least one Other, only "H-O-O-H" and "H-W-W-O-H" are dominated by the survey data. The total percentage for these six sequences in Table 4.5 shows that the survey data accounts for 59.53%, while synthesized activity plans represent 80.16%.

On the other hand, when calculating the total percentage for activity sequences with at least one Work, survey data is at 38.07%, and synthesized activity plans drops to 17.56%.

Furthermore, from Table 4.4, the KL divergence of the distribution of activity sequences is measured at 0.164. This value is closer to zero compared to the KL divergence observed for average trip distance in Section 4.3.2. This indicates that the distribution of activity sequences is closer to the ground truth data than the average trip distance.

4.3.4 Temporal visitation

When comparing the temporal visitation patterns of the synthesized agents to the survey participants, each activity type is examined individually, and only activity plans that contain the activity type are included.

Figure 4.12 presents three temporal visitation patterns of Home, Work, and Other. In graph (a), there is a delay in when the synthesized agents leave their homes, with more staying at home during the day. In the evening, the synthesized activity plans tend to indicate longer durations of staying out. A similar delay pattern as in the home visitation pattern is observed in the work (b) graph, where the synthesized activity plans show later arrivals and longer stays at the work location compared to the survey participants. For the visitation distribution of other locations, the synthesized activity plans show a broader temporal spread, with agents traveling to these locations earlier in the day and staying later.

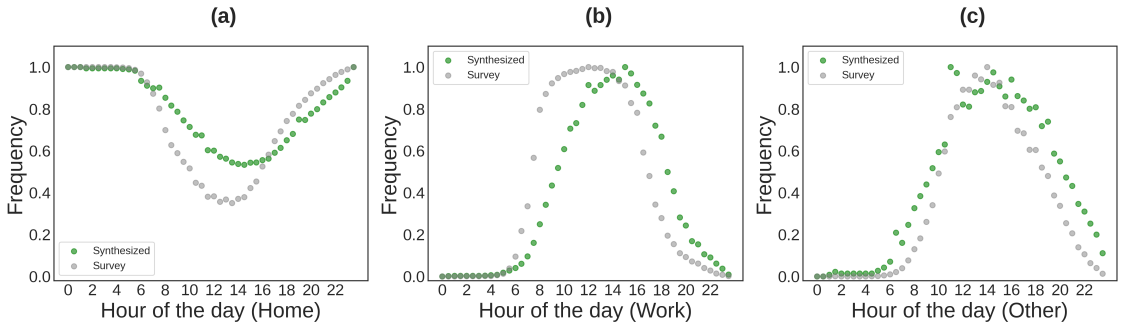

Figure 4.12: The temporal distribution of the synthesized and survey data for stays at (a) home, (b) work, and (c) other locations.

Table 4.4 presents the KL divergence, illustrating differences in temporal visitation patterns for home (a), work (b), and other (c) locations between the synthesized and the survey data. The KL values were calculated by averaging the scores across all time intervals for each respective activity type. For home, the divergence is 0.036, indicating nearly identical patterns in the two datasets. The value slightly increases to 0.069 for other locations, while work shows a more significant increase at 0.166. Figure 4.13 visualizes a 24-hour day, highlighting the time intervals where differences in KL divergence between the synthesized and the survey data are most noticeable. The graph for work is the most volatile, showing spikes from the morning to the middle of the day and another spike towards the end of the day. In comparison, the

curves for home and other locations demonstrate a lower span for KL divergence throughout the day, with a few clear fluctuations at certain times.

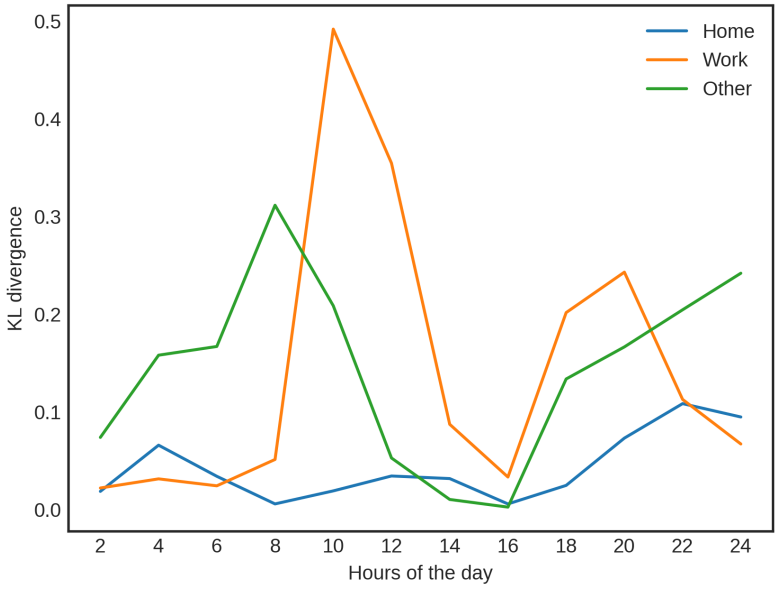

Figure 4.13: KL divergence for each two-hour interval. The hours on the x-axis indicate the end time of each interval.

4.3.5 Time duration of activities

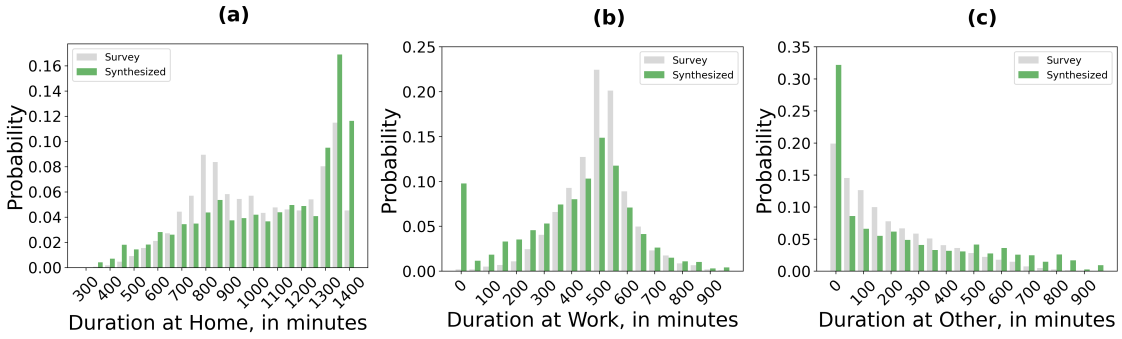

Figure 4.14: Distribution of time duration for (a) Home, (b) Work, and (c) Other activities.

Figure 4.14 (a) shows that synthesized activity plans have more cases where the majority of the day is spent at the home location, specifically from 1300 to 1440 minutes. Conversely, the synthesized data is notably underrepresented from 700 to 1000 minutes. This suggests that the synthesized activity plans lack agents who spend a large proportion of their day outside the home, for example, being at work or engaging in longer Other activities. In Figure 4.14 (b), the time durations at work are presented, and a peak is observed at 500 to 599 minutes, representing approximately 43% of the survey participants. While the synthesized activity data also peaks around the same duration, at around 27%, the peak is less distinct, and

there is a clear leftward skew in the data. In this direction, the synthesized activity plans exhibit a clear outlier, with almost 9% of the data recording a time duration between 0 and 50 minutes, whereas the survey data shows nearly no occurrences in this range.

Furthermore, the duration for other locations displays a right-skewed distribution, as shown in Figure 4.14 (c). This distribution is also evident in the synthesized activity plans, but with varying percentages for each duration category. In the 0 to 50 minutes range, survey data represents 20%, while synthesized activity plans are overrepresented, accounting for nearly 33%. For durations between 50 and 450 minutes, the synthesized activity plans are underrepresented. However, beyond 500 minutes, this relationship changes, and there is a higher number of synthesized plans than expected. From 800 minutes, the survey data has almost no recorded durations, while the synthesized activity plans continue to estimate such extended durations.

Table 4.6 summarizes the difference in duration spent at activities, confirming the patterns from Figure 4.14 where the synthesized activity plans have a more prolonged duration at home, less time at work, and more on other locations. Note that the average duration is only calculated for synthetic activity plans and survey participants' records who have visited the respective location type.

	Average time duration		Median time duration	
	Survey	Synthesized	Survey	Synthesized
Home	1045.75	1099.94	1029.00	1169.00
Work	514.84	430.55	527.50	475.00
Other	215.49	262.67	160.00	165.00

Table 4.6: Average and median time duration for users in each activity (minutes).

5

Discussion

This thesis combines Swedish national travel survey data with big geolocation data from mobile applications to synthesize activity plans. The model identifies home and work as anchor locations and excludes unreliable homes by comparing them to corresponding locations in the survey data. The user data is transformed into activity plans, and we find matching twins between the datasets using a modified Jaccard similarity. Synthesized activity plans are created by combining survey twins' activity sequences with extensive geolocation data from the mobile app users. We validate against the survey data and apply the KL divergence to assess the similarities between the two datasets.

For this chapter, we begin by highlighting the limitations of the MAD and the survey data in Section 5.1. In the following Section 5.2, we discuss the inferred home and work locations, while Section 5.3 delves into the choices made in the methodology and its outcomes. The validity of the results and the performance of the model are discussed in Section 5.4. Finally, in Section 5.5, we explore and discuss the potential future work.

5.1 Data limitations

This thesis combines big geodata with travel survey data, aiming to leverage their complementary strengths. The motivation behind this approach stems from the inherent shortcomings of each dataset. In the survey data, the recorded activity plan typically reflect a regular day that starts and ends at the home location, limiting the exploration of more diverse activity plans. Existing literature highlights that travel survey data is characterized by low sampling rates, short collection durations, and under-reporting of trips [3], [45]. Additionally, it tends to overlook the majority of infrequent long-distance trips [46].

In MAD, there is a more diverse selection of locations visited, spanning a much larger area compared to survey data [47]. It also covers a longer time frame, with the data collection stretching over several months. However, inherent issues with the data, such as sparsity and lack of information beyond the time and coordinates visited, complicate the tasks of creating reliable activity plans at the individual level. There is also a bias towards leisure locations, and a risk that work and other more mundane activities are underrepresented [48]. Another noteworthy bias in the MAD is a shift in the time of activities towards the later part of the day. This likely stems

from the collection practice, where a stay is only registered when a person uses their phone at a location. Since this event can happen at any time during the visit, the commencement of the stay in the data is often delayed, resulting in later activity records in the dataset.

Beyond these built-in issues, the compared data originates from different time periods. The survey data spans from 2011 to 2016, whereas the MAD is collected from June to December 2019. Although there is no indication that the travel patterns in Sweden have changed between these collection periods, potential differences could impact the results, particularly given the seasonal variations in the collected data. However, since there is no tracking of location types beyond home, work, and other, any potential variations in visited locations throughout the year are not expected to exert a major influence on the results.

5.2 Validity of identifying anchor locations from big geodata

Home and workplace locations serve as crucial anchor points [49] and significantly influence one's daily mobility [50]. In highly anonymized datasets like MAD, these locations must be inferred from user stays. While it is widely accepted to use simple temporal rules to estimate these locations [28], [50], the literature often lacks validation of the results. In this thesis, we introduce an additional step in the home inferring process by examining the visitation frequency patterns to enhance the reliability of the inferred home locations. The results indicate that a proportion of the inferred home locations diverges significantly from the expected visitation pattern, and inclusion of these locations in future analyses might lead to issues. Consequently, we conclude that incorporating additional verification steps can be an important way to improve the quality of analysis based on anonymized big geodata.

Upon analyzing the outcomes of the inferred home locations, a notably high correlation is observed between these inferred homes and the actual population distribution within the DeSO zones. However, there are clear outlier zones with a much larger estimated population than expected. The zones in question are all located in city centers, and within them are locations where many people pass by, for instance, central stations or shopping malls. Since these are locations many people visit, they have a disproportionate amount of visibility in the dataset, leading to issues when estimating locations of importance for our users. One potential solution could involve excluding these zones from the home estimation, or utilizing information from the graph as a weight distribution mechanism to reduce the importance of the activity plans extending from these areas. Nevertheless, this thesis does not address these issues, and solutions regarding the disproportionate spread of the inferred population would need to be explored more in future work.

Regarding the inferred work locations, the percentage of users with a work location is nearly identical to the survey participants. However, the commuting distance is significantly higher for a large portion of the inferred work locations compared to the survey data. When estimating the work location, only temporal aspects were used

(the most visited location outside the inferred home between 9 a.m. to 5 p.m.) to select the location. The large number of inferred work locations with an abnormally long commuting distance indicates that an additional aspect of distance should be included in the selection process to enhance the quality of the results. Another solution is to implement a similar temporal visitation profile screening as used in this thesis when we inferred home locations to increase the validity of the inferred work locations.

However, it is crucial to consider the trade-off between quantity and quality when implementing more stringent filtering steps. Applying stricter rules to infer locations of interest results in fewer data points and users available for future analysis. With more lenient rules, the amount of data increases, but the decrease in quality might lead to complications in the final results. A systematic investigation to identify the optimal balance between these factors should be conducted to further the results of this thesis. One final consideration is that the inference process in this thesis targets individuals with 8 a.m. to 5 p.m. jobs at a single location, sleeping at home during the night. Individuals working night shifts or in multiple locations cannot be accurately identified using the applied temporal rules.

5.3 Model designs

The thesis proposes a generative model for creating synthetic activity plans, that aims to strike a balance between the abundant geolocations and broad population coverage in MAD, and the completeness of activity plans in the survey data. The proposed model demonstrates strong transferability to similar large-scale geodata sources and ubiquitous survey data. The synthesized activity plans generated by the model provide greater variation in trip types compared to available survey data. It also generates over five times the amount of activity plans, increasing the dataset's diversity in terms of locations and regions. Compared to the unaltered MAD, the synthesized plans shows a more interesting side of the users activities and allows the data to be applied in more advanced analytical settings, as in ABM. However, certain elements of the design warrant additional enhancement.

5.3.1 Diversity in twin searching

There is a clear issue in the low diversity of survey participants used. A vast majority of the participants did not get linked to a single MAD user (13 797 out of the 18 106), while ten matched with over 1 000 users each. The lack of variety of survey participants used in synthesizing activity plans leads to less variation in travel patterns. The issue of some users receiving a very high number of matches has, in some cases, resulted in uncommon activity patterns being clearly overrepresented, a problem that will be explored more in Section 5.4. Both of these issues could be addressed by improving the way of using the survey data when searching for twin travelers. Rather than selecting the survey participant with the highest score, the twin could be randomly chosen from all participants who exceed the established similarity score threshold.

5.3.2 Work activities

Another issue in the twin matching process is that the survey participants with only home and other locations in their activity plan are matched at an unreasonably frequent rate. In both the MAD and survey data, around 38% of the participants are connected to a work location. However, after the twin matching, less than 18% of the synthesized activity plans have a work location, as seen in Table 4.3. This indicates an issue with MAD users' work locations not being present enough in the matching process. It could be due to a lack of recorded visits in the inferred work location, or that the work location visitation patterns for MAD users are skewed, making it less likely to match with survey participants that have a more regular visitation pattern to work. To address these issues, one approach is to introduce additional stays throughout the workday, enhancing the visibility of work locations for MAD users during typical working hours. Another suggestion is assigning weights to the work location stays for MAD users, giving the work location a higher probability in the matching process without changing the time of day the MAD user visits the location.

5.3.3 Other activities

The final step of the synthesizing process is to select other locations from the MAD user to incorporate into the activity sequences of twin travelers. The selection criteria are primarily temporal, where we randomly choose the other locations based on visitation frequency within the time frame of interest in the activity sequence. Although we consider distance constraints in searching for the twin travelers as described in Section 3.3.2, the spatial context can be improved in this thesis. Acknowledging the significance of distance in travel patterns, incorporating a spatial aspect into the selection process would enhance the model. A maximum travel distance boundary could be established, where all locations beyond a certain distance from the anchor locations are excluded. Another solution is to include the distance from the previous location in the activity sequence as a variable in the location selection. For example, one study [51] proposes calculating the total energy consumption for various types of travelers and using it as a supplement to a Markov process [52] when selecting destinations within a reasonable distance.

5.4 Model performance

As indicated by the KL divergences in Table 4.4, the model performance illustrates how closely the synthesized activity plans align with the survey data in the different evaluated aspects. Our model largely agrees with the survey in these key attributes. For instance, in the case of total trips, the distribution closely mirrors the survey data, a similarity which is further supported by its KL divergence approaching zero. However, while the model performs reasonably well in many aspects, there are a few deviations and caveats that will be discussed in this section.

5.4.1 Long-distance trips

The proportion of synthesized activity plans that include long distance trips significantly exceeds the corresponding data in the survey for both total and average trip distance. It is crucial to note that the survey data is not the definitive ground truth. As mentioned in 5.1, this source involves careful sampling to statistically represent the true population, but it tends to underestimate both the number of trips [3], [45] and the amount of longer trips [46]. This weakness could be a contributing factor to the differences in the results.

Another possible reason for this disparity could be the absence of a maximum travel distance boundary in our model as discussed in Section 5.3.2 and 5.3.3. However, it is important to remember that the other locations in the synthesized activity plans are derived from actual locations in the MAD, and that MAD users have traveled to them at some point. The valuable observations from big geodata sources such as MAD, where trips occur over several days, justify significantly longer travel distances than those observed in a one-day travel diary. For instance, traveling back and forth from Gothenburg to Stockholm within one day is uncommon, while traveling between these two cities on different days is more reasonable. This allows one of the strengths of the MAD, the diversity of the data, to be fully utilized.

5.4.2 Sequences and temporal patterns of synthesized plans

The distribution of activity sequences indicates an overestimation of synthesized activity plans containing only Home and Other, while those with Work are underestimated. This disparity arises from the twin matching process, as discussed in Section 5.3, where the model often fails to match MAD users with survey participants who have a work location and includes too many that travel only to other locations. When focusing exclusively on sequences within each category, those containing only Home and Work or Home and Other, the results are promising, and the order of the synthesized sequences match the surveys. For example, with the Home and Other sequences, H-O-H is the most common sequence for both datasets, H-O-O-O-H is the second most common, etc.

The distribution of time durations for the three activity types tends to overrepresent extreme values, and the model does not align with the peaks observed in the survey data. As mentioned in Section 5.3, there are survey users who matched with over 1 000 MAD users each, being overly represented in the synthesized activity plans. One particularly noticeable occurrence is the 0 to 50 minute interval spent at work. While less than 0.1% of survey participants fall under this duration, 10% of the synthesized activity plans are in that range. A similar issue arises with users spending more than 800 minutes at other locations. In the ground truth data, only a very small percentage of people do this, but in the synthesized data, there are almost 10%. These outliers underscore the overfitting problem towards certain survey participants and suggest the need for creating a higher variation in twin matching to address this discrepancy.

Finally, upon examining the probability results related to the temporal distribu-

tion for home and work locations, a subtle shift toward the afternoon and evening becomes evident in the synthesized activity plans. Additionally, there is also a noticeable discrepancy during the middle of the day with how many plans are at home. This divergence may be attributed to the scarcity of data for certain MAD users, amplifying the influence of the home location in the matching process, resulting in a significant proportion of plans being at home the vast majority of the day. For the former issue, we estimate that this stems from the inferred home and work locations. If these locations are misidentified, it could result in MAD users staying at these places during uncommon times of the day, causing deviations from the survey visitation pattern.

5.5 Future work

There are two main directions to enhance the model in this thesis. The first one is to improve the quality and diversity of the single-day activity plans. The second one is to extend the travel patterns to include several days, making it possible to extract more complex sequences. These enhancements would facilitate greater utilization and more effective engagement with the complexity of the big geodata such as the MAD.

5.5.1 Improve one-day generative model

For the first direction, some improvements are mentioned earlier in the discussion. One improvement could be synthesizing home and work location stays. Another improvement involves assigning weights to existing stays for MAD users, which could yield more realistic activity plans by reducing the biases embedded in the passive data collection mechanism in big geodata like MAD. This, in turn, increases the probability of MAD users matching with working survey participants and should improve the overall results.

In addition, one could implement a random selection method for eligible survey twins when combining survey data with big geodata. This approach has the potential to enhance the diversity of matched twins while also reducing the risk of specific participants disproportionately influencing the results.

Finally, we could improve the activity plan creation by better-assigning work and other locations. For example, those beyond a certain distance threshold from the inferred home location could be excluded. A distance weight could be applied for the remaining locations to reduce the probability of selecting a location that is unreasonably remote considering its preceding one, either as the inferred work location or the selected other location. This adjustment would decrease the number of long-distance trips in the synthesized activity plans, making the total and average distance traveled better approximate the reality.

5.5.2 Extend one-day to multi-day activity plans

The second direction, synthesizing activity plans spanning several days, requires substantially modifying the model proposed in this thesis. Since the survey participants only provide activity plans for a single travel day, increasing connected days would complicate using the survey data as a straight template for the synthesized activity plans. One study suggests combining the survey data with public transport smart-card data to extend the travel pattern to several days [53]. Alternatively, the structure of activity plans could be synthetically created starting from the mechanisms of human mobility instead of directly using empirical data, for example, by using mobility studies as [7]. While this introduces a significant increase in model complexity, the resulting activity plans would more accurately mirror the actual mobility patterns of the population. Such an enhancement could substantially improve the current survey data used in transportation planning.

6

Conclusion

Studying the mobility patterns of a population yields crucial insights for informed decision-making in infrastructural construction and city planning. The surveys traditionally used in these studies provide a clear yet simplified representation of actual travel behaviors, often missing longer trips and unique travel patterns. The utilization of MAD presents an opportunity to address these limitations, while at the same time being more cost-effective and easier to collect. However, the sparsity and limited information outside the temporal and spatial aspects in MAD necessitate pre-processing before it can be applied in more advanced contexts, such as in ABM. This thesis addresses the issues in two ways. First, we employ standard practice temporal rules to infer home and work locations for the MAD users. A layer of temporal visitation pattern matching is then introduced for home location to enhance the validity of the results. Secondly, we address the sparsity issue by merging the locations of interest from the MAD with activity sequences from survey participants, creating synthesized activity plans without any gaps throughout the day.

The construction of this generative model explores a new area of MAD usage in mobility studies, establishing a baseline from which further improvements and exploration can be conducted. One strength of the proposed generative model for applying big geodata is in the quantity of produced data. The number of synthesized activity plans is more than five times the amount present in the survey. Moreover, the evaluation results indicate that the produced activity plans largely agree with the survey data regarding the essential aspects of human mobility patterns, e.g., trip frequency, distributions of activity sequences, and temporal visitation patterns for the different types of locations. The trip distance in the synthesized activity plans is notably long for a significant proportion of users. However, as detailed in Section 5.4, these types of long-distance trips actually occur, and are something the traditional surveys struggle to collect.

Considering MAD's numerous advantages over traditional surveys, including its ease of collection, the ability to observe a large number of users, and the capacity to account for seasonal variations, utilizing this data in a broader range of mobility studies would benefit transport and urban planning. Continuing the research presented in this thesis and further exploring the introduced methods could provide transport agent-based modeling with a great data source for generating a multitude of agents with flexible and realistic activity plans. These agents have the potential to enhance mobility models, providing decision-makers and city planners with a more robust

foundation for shaping the transportation networks of the future.

Bibliography

[1] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin–destination matrices using mobile phone call data," *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 63–74, 2014.

[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," 2011.

[3] 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 Behaviour and Society*, vol. 1, no. 2, pp. 69–78, 2014.

[4] H. Kavak, J. J. Padilla, C. J. Lynch, and S. Y. Diallo, "Big data, agents, and machine learning: Towards a data-driven agent-based modeling approach," in *Proceedings of the Annual Simulation Symposium*, 2018, pp. 1–12.

[5] N. Marilleau, "An agent based meta-model for urban mobility modeling," in *First International Conference on Distributed Frameworks for Multimedia Applications*, IEEE, 2005, pp. 168–175.

[6] H. Barbosa, M. Barthelemy, G. Ghoshal, *et al.*, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.

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

[8] Y. Liao, "Understanding human mobility with emerging data sources: Validation, spatiotemporal patterns, and transport modal disparity," Ph.D. dissertation, Chalmers Tekniska Hogskola (Sweden), 2020.

[9] R. Kitamura, "An evaluation of activity-based travel analysis," *Transportation*, vol. 15, pp. 9–34, 1988.

[10] Transport Analysis, *Travel survey*, May 2022. [Online]. Available: `https://www.trafa.se/en/travel-survey/travel-survey/`.

[11] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, 2011, pp. 145–156.

[12] C. Guan, J. Song, M. Keith, Y. Akiyama, R. Shibasaki, and T. Sato, "Delineating urban park catchment areas using mobile phone data: A case study of tokyo," *Computers, Environment and Urban Systems*, vol. 81, p. 101 474, 2020.

[13] A. I. Tokey, "Spatial association of mobility and covid-19 infection rate in the usa: A county-level study using mobile phone location data," *Journal of Transport & Health*, vol. 22, p. 101 135, 2021.

[14] J. Oh, H.-Y. Lee, Q. L. Khuong, *et al.*, "Mobility restrictions were associated with reductions in covid-19 incidence early in the pandemic: Evidence from a real-time evaluation in 34 countries," *Scientific reports*, vol. 11, no. 1, pp. 1–17, 2021.

[15] G. Bonaccorsi, F. Pierri, M. Cinelli, *et al.*, "Economic and social consequences of human mobility restrictions under covid-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 530–15 535, 2020.

[16] F. Schlosser, V. Sekara, D. Brockmann, and M. Garcia-Herranz, "Biases in human mobility data impact epidemic modeling," *arXiv preprint arXiv:2112.12521*, 2021.

[17] M. D. Garber, K. Labgold, and M. R. Kramer, "On selection bias in comparison measures of smartphone-generated population mobility: An illustration of no-bias conditions with a commercial data source," *Annals of Epidemiology*, vol. 70, pp. 16–22, 2022.

[18] S. Lai, A. Farnham, N. W. Ruktanonchai, and A. J. Tatem, "Measuring mobility, disease connectivity and individual risk: A review of using mobile phone data and mhealth for travel medicine," *Journal of travel medicine*, vol. 26, no. 3, taz019, 2019.

[19] Statistics Sweden, *Use of mobile phones and applications by gender and age. years 2018-2020.* Available online at `https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__LE__LE0108__LE0108J/LE0108T30/?fbclid=IwAR0R1OVpj8o4LZZb0cKb1avrEHCYLKz_V7j9K5UeKPNYLemCYaKeaGQw1KI`, 2023.

[20] Y. Liao, K. Ek, E. Wennerberg, S. Yeh, and J. Gil, *A mobility model for synthetic travel demand from sparse traces*, 2022.

[21] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete trajectory reconstruction from sparse mobile phone data," *EPJ Data Science*, vol. 8, no. 1, pp. 1–24, 2019.

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

[23] J. Drchal, M. ertick, and M. Jakob, "Data-driven activity scheduler for agent-based mobility models," *Transportation Research Part C: Emerging Technologies*, vol. 98, pp. 370–390, 2019.

[24] E. Arkangil, M. Yildirimoglu, J. Kim, and C. Prato, "A deep learning framework to generate realistic population and mobility data," 2022.

[25] C. R. Bhat and F. S. Koppelman, "Activity-based modeling of travel demand," *Handbook of transportation Science*, pp. 39–65, 2003.

[26] H. Zhou, J. Dorsman, M. Mandjes, and M. Snelder, "Sustainable mobility strategies and their impact: A case study using a multimodal activity based model," *Case Studies on Transport Policy*, vol. 11, p. 100 945, 2023.

[27] H. J. Miller, "Activity-based analysis," *Handbook of regional science*, pp. 187–207, 2021.

[28] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, and L. Bravo, "Evaluation of home detection algorithms on mobile phone data using individual-level ground truth," *EPJ data science*, vol. 10, no. 1, p. 29, 2021.

[29] L. Tongsinoot and V. Muangsin, "Exploring home and work locations in a city from mobile phone data," in *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2017, pp. 123–129.

[30] Y. Zhang, C. Li, Y. Song, Y. Chai, and Y. Fan, "Personalizing the dichotomy of fixed and flexible activities in everyday life: Deriving prism anchors from gps-enabled survey data," *Transportation*, pp. 1–26, 2022.

[31] C. Anda, S. A. O. Medina, and K. W. Axhausen, "Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103 118, 2021.

[32] E.-J. Kim, D.-K. Kim, and K. Sohn, "Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network," *Transportation Research Part C: Emerging Technologies*, vol. 137, p. 103 616, 2022.

[33] G. Badu-Marfo, B. Farooq, and Z. Patterson, "Composite travel generative adversarial networks for tabular and sequential population synthesis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 976–17 985, 2022.

[34] A. Thamm, *Heuristics*, 2023. [Online]. Available: `https://www.alexanderthamm.com/en/data-science-glossary/heuristic/`.

[35] K. Daniel, *Thinking, fast and slow*. 2017.

[36] C.-H. Joh, T. Arentze, and H. Timmermans, "Modeling individuals activity-travel rescheduling heuristics: Theory and numerical experiments," *Transportation Research Record*, vol. 1807, no. 1, pp. 16–25, 2002.

[37] Y. Liao, J. Gil, S. Yeh, R. H. M. Pereira, and L. M. Alessandretti, "Unraveling nativity segregation and its link with transport access: A mobility perspective using big geolocation data in sweden," Working Paper, Aug. 30, 2023.

[38] K. W Axhausen, A. Horni, and K. Nagel, *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.

[39] Statistics Sweden, *Deso - "demographic statistical areas"*, Available online at `https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/regionala-indelningar/deso-\protect\@normalcr\relax--demografiska-statistikomraden/`, 2023.

[40] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, pp. 267–276, 1953.

[41] K. Cui1 He, "Travel behavior classification: An approach with social network and deep learning," *Transportation Research Record*, vol. 2672(47), pp. 68–80, 2018.

[42] H. Cong, *Flows generation for synthetic travel demand*, 2023.

[43] S.-i. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194.

[44] W. Qian, F. Lauri, and F. Gechter, "A probabilistic approach for discovering daily human mobility patterns with mobile data," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part I 18*, Springer, 2020, pp. 457–470.

[45] 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," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 162–177, 2015.

[46] G. Mattioli and M. Adeel, "Long-distance travel," in Elsevier, 2021.

[47] M. Janzen, M. Vanhoof, Z. Smoreda, and K. W. Axhausen, "Closer to the total? long-distance travel of french mobile phone users," *Travel Behaviour and Society*, vol. 11, pp. 31–42, 2018.

[48] C. Liao, D. Brown, D. Fei, X. Long, D. Chen, and S. Che, "Big data-enabled social sensing in spatial analysis: Potentials and pitfalls," *Transactions in GIS*, vol. 22, no. 6, pp. 1351–1371, 2018.

[49] H. J. Miller *et al.*, "Time geography and space-time prism," *International encyclopedia of geography: People, the earth, environment and technology*, vol. 1, 2017.

[50] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, 2017.

[51] W. Wang and T. Osaragi, "Daily human mobility: A reproduction model and insights from the energy concept," *ISPRS International Journal of Geo-Information*, vol. 11, no. 4, p. 219, 2022.

[52] D. W. Stroock, *An introduction to Markov processes*. Springer Science & Business Media, 2013, vol. 230.

[53] S. A. O. Medina, "Inferring weekly primary activity patterns using public transport smart card data and a household travel survey," *Travel Behaviour and Society*, vol. 12, pp. 93–101, 2018.