CHALMERS

Agent-based Transport Models as a Tool for Evaluating Mobility

ÇAĞLAR TOZLUOĞLU

Department of Space, Earth and Environment
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2022

THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Agent-based Transport Models as a Tool for Evaluating Mobility

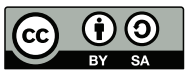
ÇAĞLAR TOZLUOĞLU

Department of Space, Earth and Environment
Division of Physical Resource Theory
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2022

Agent-based Transport Models as a Tool for Evaluating Mobility
ÇAĞLAR TOZLUOĞLU

Department of Space, Earth and Environment
Division of Physical Resource Theory
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000



Chalmers Reproservice
Göteborg, Sweden 2022

ABSTRACT

The transportation system is undergoing fundamental transformations through emerging technologies. Some of these innovations have the potential to contribute to the sustainable transformation of the transportation system, such as electric vehicles (EVs) and shared autonomous electric vehicles (SEAVs). Before enacting policies to support these technologies or limit the use of undesirable ones, decision-makers need to better understand these innovations and the consequences of the policy to be implemented. This insight can be provided with models that are capable of reflecting the dynamics of new mobility, and interactions of travelers with each other and the infrastructure. This thesis describes the development of the Synthetic Swedish Mobility (SySMo) model that represents the travel behavior of an advanced synthetic population of Sweden, using an agent-based framework. The SySMo model provides a scaffold to build decision support tools through which present and future mobility scenarios can be analyzed and thus aid decision-makers in formulating informed policies.

The SySMo model comprises a series of modules that utilize a stochastic approach combined with Neural Networks, a machine learning technique to generate a synthetic population and behaviorally realistic daily activity-travel schedules for each agent. The model first generates a synthetic replica of the population characterized by various socio-economic attributes using zone-level statistics and the national travel survey as input data. Then, daily heterogeneous activity patterns showing activity and trip features are assigned to each individual in the population with a high spatio-temporal resolution. To assess the SySMo model performance in each module, in-sample evaluations (i.e., comparing the model outputs with input data to measure the similarity of the results) and out-of-sample (i.e., comparing the model outputs with data never used in the model) evaluations are performed. The current model offers a valuable planning and visualization tool to illustrate mobility patterns of the Swedish population. The methodology can also be broadly applied to other regions with other relevant data and carefully calibrated parameters.

Keywords: Agent-based modeling; Activity-based modeling; Activity generation; Machine learning; Daily activity pattern

APPENDED PUBLICATIONS

This thesis consists of an extended summary and the following appended papers:

Paper A Ç. Tozluoğlu, S. Dhamal, Y. Liao, S. Yeh, F. Sprei, M. Marathe, C. Barrett and D. Dubhashi (2022a). Synthetic Sweden Mobility (SySMo) model documentation.

Paper B Ç. Tozluoğlu, S. Dhamal, S. Yeh, F. Sprei, Y. Liao, M. Marathe, C. Barrett and D. Dubhashi (2022b). The heterogeneous travel activity of a synthetic population.

Author contributions

Paper A: Conceptualization: Çağlar Tozluoğlu (Ç.T.), Swapnil Dhamal (S.D.), Yuan Liao (Y.L.), Sonia Yeh (S.Y.), Frances Sprei (F.S.), Devdatt Dubhashi (D.D.), Madhav Marathe (M.M.), Christopher Barrett (C.B.); methodology: S.D., Ç.T., S.Y., F.S.; software: S.D., Ç.T.; validation: Ç.T., S.D.; data curation: Ç.T., S.D.; writing - original draft: S.D., Ç.T.; writing - review & editing: Ç.T., S.Y., F.S., Y.L.; project administration: S.Y., F.S.

Paper B: Conceptualization: Çağlar Tozluoğlu (Ç.T.), Swapnil Dhamal (S.D.), Yuan Liao (Y.L.), Sonia Yeh (S.Y.), Frances Sprei (F.S.), Devdatt Dubhashi (D.D.), Madhav Marathe (M.M.), Christopher Barrett (C.B.); methodology: S.D., Ç.T., S.Y., F.S.; validation: Ç.T., S.D.; writing - original draft: Ç.T., S.D.; writing - review & editing: Ç.T., S.Y., F.S., Y.L.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Frances Sprei, and Sonia Yeh for giving me guidance in my research. I would also like to thank them for creating such an excellent research atmosphere with their intellectual and scientific support and always reserving time for discussion. Your guidance made this thesis possible.

I would like to thank Swapnil Dhamal for joining the SySMo model development journey and accompanying me in numerous online coding sessions during the Covid time. I would also like to thank Yuan Liao providing feedback and advice on my research at many points.

Thanks to Madhav Marathe, and Christopher Barrett from the University of Virginia for their advice and valuable comments at the different stages of my research. Thanks to Peo Nordlöf from Trafikverket for providing mentorship with his years of industry experience.

I would also like to thank my colleagues at FRT. Throughout my work, I was surrounded by great people with whom I spent many hours of cheerful coffee breaks. Special thanks to Ankit, Avi, Ahmet, Jinxi, Ella, Gavin and all of my friends for making days excellent with tea parties, ping-pong matches, and various events.

Most importantly, I would like to thank Aynur (Mom), Mustafa (Dad), Çağrı (Brother), and Melda (Sister-in-law) for their endless support. Last but not least, I would like to express my sincere and heartful gratitude to my wife Ezgi for encouraging me to start this journey, and her love and support made this work easier for me.

Göteborg, May 2022
Çağlar Tozluoğlu

CONTENTS

CHAPTER 1

Introduction

The transportation system is undergoing fundamental transformations through emerging technologies. Increased urbanization, growing population, and the unsustainable nature of the current transportation system make these changes necessary. Micromobility, electric vehicles (EVs), and shared autonomous electric vehicles (SEAVs) are some of the innovations that have the potential to contribute to the sustainable transformation of the transportation system. Moreover, emerging technologies that offer services to existing and new user groups can also change people's attitudes and behaviors concerning mobility [1, 2]. When assessing these potential changes, decision-makers should be supported by models that are capable of reflecting the dynamics of new mobility, and interactions of travelers with each other and the infrastructure. Considering that the urban population will more than double its current size in the next 30 years [3], and actions must be taken in the transport sector as soon as possible to limit global warming [4], the need to better understand these changes grows ever more urgent.

Agent-based modeling (ABM) framework equipped with activity-based travel demand generation approach is one possible way to serve this need. Researchers commonly use ABM to model travel behavior, and most agree that the activity-based approach provides a rigorous view of the transport model in creating travel demand. [5]. Rasouli and Timmermans [6] argues that the widespread use of big data sources and the growth of computing power have enabled a faster development of activity-based models toward integrating sub-models, capturing the dependencies between trip chains, higher temporal and spatial resolution, and behavioral realism.

The Synthetic Swedish Mobility (SySMo) model is a framework that simulates transportation in large geographic regions based on the

agent-based modeling (ABM) approach. SySMo provides a scaffold for building decision support tools through applying state-of-the-art methods rooted in recent advances in transportation modeling and computer science. The developed tool assists policymakers in identifying key drivers of innovations and user trends and formulating informed policies. The model first generates a synthetic replica of the Swedish population with socio-economic attributes. After that, the heterogeneous activity-travel schedules showing activity and trip features are added to the synthetic population. So (a) it provides a platform to create various future scenarios, and (b) it does not violate any privacy issues since it is completely synthetic.

Scope and contributions

This licentiate thesis broadly deals with the modeling of human travel behavior at a high spatio-temporal resolution to evaluate today's and future mobility using an agent-based modeling framework. The main scope of the thesis is the development of a model that realistically simulates Sweden's transportation system and human transportation behavior, focusing on the following question:

- How can we generate a synthetic population that is a statistically accurate representation of the Swedish population with certain attributes, to use in ABM?
- How can we model the activity-travel behaviors of individuals in the developed synthetic population?
- How can we maintain heterogeneity, a fundamental feature of human activity behavior, in the population?
- How can we incorporate state-of-art methods in computer science into the modeling process to increase the realism of the simulation?

In order to simulate human travel behavior, the developed model first generates a complete synthetic population with the activity-travel pattern for Sweden using the current data and data structures. Although some studies focus on a small region or a particular mobility behavior such as long-distance travel Canella et al. [7] and Márquez-Fernández et al. [8], a synthetic population replicating the entire

Swedish population does not exist in the current literature. The model address this gap. Furthermore, the proposed methodology employs a novel approach that utilizes the advantages of machine learning techniques. Neural networks, a machine learning technique, capture the correlations between individuals' attributes and their activity sequences with high predictive ability in complex data sets. Paper A explains the details of the novel methodology that generates the synthetic population with mobility patterns.

The proposed methodology generates heterogeneous daily activity schedules showing activity type, start-end time, duration, and sequence for the Swedish synthetic population and creates realistic daily plans of the individual mobility. Traditionally, disaggregated transportation models provide homogeneous daily activity patterns within the sub-population. Homogeneous activity schedules may be a reasonable simplification for many applications, but will be inadequate to assess the effects of policy efforts linked to significant behavioral changes. Paper B describes the methodology producing heterogeneous activity patterns in a synthetic population.

Disposition of this thesis

The thesis consists of four chapters followed by the papers; "The Synthetic Sweden Mobility (SySMo) Model Documentation" and "The Heterogeneous Travel Activity of a Synthetic Population". It is organized as follows: Chapter 2 begins with introducing key concepts in the transportation field and the existing transport modeling approaches. Chapter 3 provides a brief summary of the appended papers with the main results. Chapter 4 ends with a general reflection upon my research and future directions for my research.

CHAPTER 2

Background

This chapter presents a brief overview of transport modeling tools with a focus on the agent-based modeling approach. To provide a broader perspective, we start by describing the fundamental concepts of transportation and transportation modeling. The second section introduces the agent-based transport modeling approach and its components; population synthesis, travel demand generation, and simulation models.

2.1 Transportation modelling

Transportation models are commonly used to explore a wide variety of questions concerning human mobility behavior within the transportation system. Models are established to analyze the entire transportation system or specific components and produce quantitative outputs from the analysis, such as changes in the number of passengers in a public transportation system or peak hours on a road network. Before discussing transportation models and modeling approaches, it is worth outlining what transportation is. According to Black (2004), transportation can be defined as:

> "Transportation is concerned with the movement of goods and people between different locations and systems used for this movement. Included in the former would be the journey to work, trade flows between nations, commodity flows within a single nation, passenger flows by various modes, and so forth, and those factors that affect these flows. In general, movement within a single industrial firm or building, or the migration of population, is not included in this area." [p13, 9]

Cascetta [10] defines a transportation system as a set of interacting or interrelated elements working together that generate the travel demand in a particular area and produce the supply of transport services to meet the travel demand. While the demand for transportation is determined by the main factors varying by time and geographical areas, such as demographics, economic activities, transport options and their service prices and quality, and land use [11], the supply of transportation is determined by users' demand for transport as well as the technical aspects of physical transport supply and the given decisions regarding the presentation of the supply. To make informed transportation policies, decision-makers need insight with predictions over the transportation system. Transport models serve this purpose; more specifically, demand models are used to predict the use of transport services today or alternative future scenarios.

Transportation models reproduce an abstract copy of the transport system as a whole or a particular subsystem using mathematical methods based on specific theories. [12]. To date, numerous transport models have emerged to make inferences about the transportation system. The first operational model developed dates back to the 1950s and was used to analyze transportation-related investments in the USA [13, 14]. Transportation models have shown significant development over time from smaller models reflecting car mode only in the peak hour to more advanced models covering multimodal transport modes for 24 hours using the disaggregated modeling approaches [15]. Along with the development in modeling techniques over time, the application of the outputs produced by the models has been diversified. For example, while models were mainly used for making investment plans, they have also begun to use them in various areas such as policy-making for demand management [16], environmental pollution measurement [17], or calculating energy needs [8, 18].

2.1.1 Modelling of travel demand

This section discusses the commonly used approaches in travel demand models. To design and evaluate transportation systems, it is crucial to predict the travel demand and its variation in space and time. Travel demand models typically consist of a combination of submodels, such as mode choice models or population flow prediction models. These submodels are used to forecast various aspects of the

trips. Cascetta [10] formulates the travel demand model organized as flows between two points or regions as a function of population's socio-economic characteristics and transport infrastructures features, according to the given travel characteristics such as travel purpose or travel mode. That is;

$$\mathbb{D}_{o,d}(K_1, K_2, ...K_n) = f(SE, T, \beta) \tag{2.1}$$

Here, the travel demand flow from origin o to destination d with K_n characteristics is denoted as a function. SE shows the socio-economic variables of the decision-makers in the transportation system, and T, the level-of-service attributes of the transportation supply system. β specifies the model parameters regarding the travel flow between o, and d. Depending on the model adopted to explain the travel flow, the parameters to be used differ (see more in Chapter 8 in the book [10]).

Aggregate and Disaggregate Modelling

Transportation models are designed either to represent the travel behavior of each individual in a population separately or to represent the population as a whole. Depending on the representation, they are called aggregated or disaggregated models. The aggregated models estimate travel behavior between two spatial regions, such as municipalities or zone defined by dividing the city into smaller areas [19]. The disaggregated models simulates the travel behaviors of a single decision-maker (an individual) or a group of decision-makers having similar characteristics (a family) [10]. The aggregation level of the model is defined depending on various factors such as the modeling objectives and the data structure, the devoted time, and the domain knowledge. The disaggregated model results may need to be presented in an aggregated way to be helpful in planning and policy-making processes. While the aggregation process seems relatively easier to get the flows by combining individuals' travel, disaggregating flows requires more rigorous work [19]. A comprehensive description of aggregated models can be found in the books by Dios Ortúzar and Willumsen [12] and Daskin [20].

Modelling approaches

In the literature, there are two common methods to model people's travel demands, trip-based and activity-based modeling approaches. The trip-based modeling approach is the first to emerge technique and calculates travel demand using trips as the unit of analysis. The early applications assume trips occur independently of previous and subsequent trips and forecast the travel demand between zones. A few studies have adopted a tour-based approach taking into account other trips within the tour by developing the trip-based applications [21–23]. In the tour-based modeling approach, the unit of analysis is tours that are defined as the trips from home to one or more locations and then back home [24]. There are also a limited number of disaggregated trip-based modeling studies that independently estimate each individual's travel [25]. Although the trip-based modeling approach has progressed over time, it lacks a valid explanation of the underlying causes of travel behavior.

The activity-based travel demand models adopt a holistic approach that considers travel demand in connection with individuals' activity patterns. This modeling approach aims to jointly deduce the activity schedules of individuals and the travels between the activities for a specific time period (usually one day) [26, 27]. With this concept, the inadequacy of trip-based modeling in reflecting behavioral realism is overcome by the activity-based modeling approach that presents the underlying reason for travel.

The conceptual framework of activity-based modeling consists of two key ideas. First, the travel demand originates from participating in activities and is a derived need [28]. Travel is only undertaken when the utility to be gained from participating in the activities exceeds the disutility caused by the trip. People mostly do not travel for the sake of travel with the possible exceptions such as travel for tourism . Second is the space-time prism concept imposing the temporal and spatial constraints that decision-makers face participating in spatially distributed opportunities [29].

To provide a better understanding of travel behavior and explore different aspects of activity-based modeling, many studies have been conducted. Pas [30], and Hanson [31] investigate the correlation between activity-travel patterns and socioeconomic attributes such as age, gender, and employment status. Kitamura [32] identifies the interdependence of activity locations within the activity sequence.

Golob and McNally [33], and Pooley et al. [34] deal with interactions between household members and their activity patterns. Here, we cover only the principal ones (see more in the paper [5]). The different methodologies in the application of the activity-based approach are presented in the section about the activity-based travel demand modeling in section 2.2.2.

2.1.2 Four-step transportation model

The four-step transportation model (FSM) is a traditional trip-based modelling approach that has been widely used in the transportation modeling field [12, 35]. These models are a primary tool for evaluating large-scale infrastructure projects. It generates aggregated travel flows by the defined travel characteristics between regions in a certain time period. Most FSMs are developed to simulate peak hours or an average day. The overall framework of the FSM contains four successive, and independent steps: trip generation, trip distribution, mode choice, and traffic assignment. The steps can be defined as:

- **Trip generation:** The first step predicts the number of trips produced (started) and attracted (end) in each zone. Using zone level statistics, the production and attraction numbers are modeled independently at an aggregated level. Trip production is deduced by using variables such as population, number of households, income level, number of cars, and residential density. Trip attraction is calculated by using variables such as office space, number of retail buildings, number of employees, and student capacity.
- **Trip distribution:** The objective of the second step is to match trip starts and ends. The best-known technique to calculate travel flow between regions is gravity models. It uses a function that distributes the number of trips between two locations inversely proportional to their distance. This step gives the origin-destination (OD) matrices.
- **Mode choice:** The total trip numbers between zones are distributed among the transportation modes in this step. Discrete choice models, such as the nested logit model, are often used to deduce modal split. This step produces mode-specific OD matrices from the matrices produced in the previous step.

- **Trip assignment:** The last step considers the assignment of trips to a transport network such as road network or public transport network to simulate travel flow. It produces outcomes regarding aggregated travel behavior of the population and the network's performance.

FSMs have been heavily criticized for inadequately presenting human travel behavior, although widely used in modeling [25, 36]. The main criticism is that FSM adopting a trip-based approach, lacks behavioral foundations associated with the creation of travel demand [37]. Furthermore, in FSM, spatial and temporal inter-dependencies between trips in the same trip chain are also disregarded since each trip is independently predicted [35]. Another shortcoming is that FSM inadequatly reflects the inter-dependencies of the different characteristics of an individual's travel such as time, mode, and location. The behavioral inadequacies of traditional FSM with an aggregated modeling approach makes it less sensitive to evaluate complex transportation policies related to specific times of the day or specific travel behaviors. For instance, most FSM is not capable of predicting responses to travel demand management (TDM) policies such as strategies to increase car occupancy [38]. FSM is better suited for infrastructure measures than behavioral measures.

2.2 Agent-based transport modeling

This section will explain the use of agent-based modeling to model the transportation system. Agent-based modeling (ABM) is a general framework that models a system by dividing it into individual actors interacting with each other and the environment according to their characteristics, based on predefined rules [39, 40]. After the 2000s, this approach started to be used frequently in modeling activity-travel behavior [5], as well as in other fields such as telecommunication technologies [41] power markets [42], and financial systems [43]. The agent-based structure with a disaggregated modeling approach in which each actor and their relationships are modeled separately, has made ABMs a valuable tool in modeling transportation systems.

To model a transportation system with an ABM framework, one first needs to generate the agents, which are the system's main components. The agents, a synthetic representation of individuals in the

population, are created with various attributes that affect their interactions with other agents and the environment. The behavior rules set the limits of the agents' actions. The agents act by their attributes and the rules in the given environment. Most agent-based transport models build these components by following the workflow comprising population synthesis, travel demand generation, and execution of agents' daily plans. Activity-based modeling, where transportation demand is generated assuming people are traveling to participate in activities, fits well with agent-based modeling [44]. Activity-based modeling is one of the most commonly used methods to generate travel demand for agents in ABMs

There are three main steps of agent-based modeling, shown in Figure 2.1:

- *The population synthesis module* creates the population in the modeled area with certain attributes.
- *The activity generation module* creates an activity-travel schedule to each agent in the population.
- *The multi-agent travel simulation module* executes the assigned schedules to agents in the transport network.

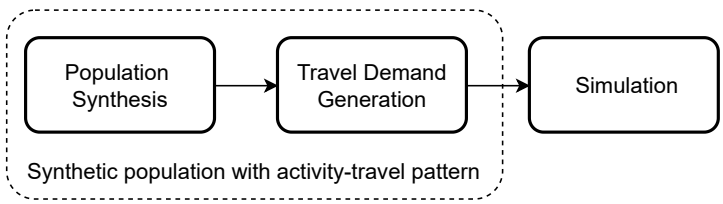


Figure 2.1: An overview of Agent-based transportation model workflow.

2.2.1 Population synthesis

Population synthesis generates a statistically representative population with their characteristics living in a particular geographical area. Any agent-based transportation models require an initial synthetic

population being the fundamental input. Rich [45] argues that the population synthesis step needs to create a representative picture of the people in a given base year and area, be adequately detailed concerning socioeconomic variables to meet the requirements of the transportation model, and identify individuals and their relationships. For modeling transportation, the agent and their relations usually represent individual people grouped by households [46]. Some synthesis methods are dynamic which also deal with the generation of the projection of the population for the future using fertility and mortality rates [47]. However, in this thesis, we will only consider the generation of the people with static methods in a base year.

The population synthesis methods have two categories, reweighting and synthetic reconstruction [48]. The reweighting methods aim to reproduce the population using various techniques that assign weights to micro-data obtained from a national survey. Different ways have been developed under this category, such as combinatorial optimization or generalized regression. The second category, synthetic reconstruction, generally refers to the iterative proportional fitting (IPF) technique and has widespread usage in the transportation modeling field (see, e.g., Smith et al. [49], Frick [50], Arentze, Timmermans and Hofman [51] and Guo and Bhat [52]).

The IPF technique makes a reference sample consistent with known statistics called marginals or control totals. A reference sample is created by using an initial frequency cross-table of all relevant attributes. Let $x = (x_1, x_2, x_3)$ denote attributes of agents in a population and let $N(z, x)$ denote the target values for attribute x in zone z. To estimate $n(z, x_1, x_2, x_3)$, $\forall z, x_1, x_2, x_3$ (i.e., the number of agents belonging to every combination of zone z, attribute 1 x_1, attribute 2 x_2, attribute 3 x_3), IPF is used with the known target values $N(x_1, x_2)$ and $N(x_1, x_3)$. In particular, the following sequence of update rules is iterated until a desired level of convergence is reached.

$\forall z, x_1, x_2, x_3$

$$n(z, x_1, x_2, x_3) \leftarrow \frac{N(x_1, x_2)}{\sum_{x_{1'} \atop x_{2'}} n(z_d, x_{1'}, x_{2'}, x_3)} n(z, x_1, x_2, x_3) \tag{2.2}$$

$$n(z, x_1, x_2, x_3) \leftarrow \frac{N(x_1, x_3)}{\sum_{x_{1'} \atop x_{3'}} n(z_d, x_{1'}, x_2, x_{3'})} n(z, x_1, x_2, x_3) \tag{2.3}$$

Here, the initialization value of $n(z, x_1, x_2, x_3)$ can be deduced by dividing the total number of the population by the number of combinations of the attribute set. Equation 2.2 and 2.3 drive the numbers obtained in an iteration toward the target numbers of (x_1, x_2, x_3) at the zone. This operation is repeated until the expected convergence level is reached. IPF is a suitable technique for both estimation of values and maintaining the known correlation structures between attributes.

2.2.2 Activity based travel demand modelling

Generating the daily travel demand for each individual is a crucial component of agent-based modeling in transport. The activity-based approach being a specific type of travel demand modeling, is often utilized in ABMs [53]. These methods are used to model various components of the travel demand such as activity sequence, activity location, activity duration, and transport mode by using the activate-based approach. The conceptual background behind this approach is explained in section 2.1.1. Here, we will discuss different methods using the activity-based approach, concentrating on computational process models, which will be explained below.

The methods used to develop activity-based models are grouped under three categories: constraint-based models, econometric models, and computational process models [6]. While some models stick to a single modeling method, there are also models utilizing multiple activity-based modeling methods such as TASHA [54], ADAPTS [55].

Constraint-based models are used in the activity-based modelling approach to generate travel demand. These models do not aim to predict activity-pattern but rather to evaluate whether a given activity schedule is possible in a particular space-time context [56]. All possible activity schedules are first generated using a combinatorial algorithm, and the schedules are then given as input to the models. The constraint-based model checks the feasibility of the activity schedule by the start-end time of activities, the activity duration, the location of the activities, and travel time with the used transport mode. These models have some shortcomings: (i) the choice of travel behavior under uncertainty is disregarded, (ii) space-time criteria are defined by a deterministic approach determining fixed opening times, the maximum speed limit, and so on [6, 57]. PESASP [56], CARLA [58], MAGIC [59] and GISICAS [60] are some of the examples to the

constraint-based of modeling approach.

The second stream of activity models is **econometric models** (utility-maximizing models). These models are conceptualized based on the theory that individuals constantly desire to maximize their utilities from their choices. These models utilize a series of discrete choice models (particularly nested logit models) to represent individuals' travel decision-making processes and deduce the activity schedule that provides the maximum utility to the individuals from their activity travel choices [6, 61]. One of this type's best-known models developed by Ben-Akiva and Bowman [62],[63] formalizes its methodology with five nested model as follows: (i) decision-makers choose a travel pattern including not traveling and accordingly, (ii) time of primary tour of day, (iii) primary destination and travel mode, (iv) time of secondary tour of day, (iii) secondary destination and travel mode. Activity travel pattern is defined by primary tour type categorized into the home, work, school, or other, and its frequency and secondary tour. CEMDEP [64], PCATS [65], and NYMTC [66] are some of the examples.

Although econometric models are widely used in modeling travel behavior, they are criticized for being unrealistic. These models assume that all decision-makers are rational utility maximizers. The assumption means that people think about the consequences rationally and choose the one that gives them the most benefit in every decision. However, this assumption is not always valid. Individuals also make decisions that are less beneficial or whose benefits are unknown.

Computational process models (rule-based models) are one of the most recent modelling techniques in the activity-based modeling approach. A set of heuristic rules are used to model travel behaviors instead of applying the assumption that individuals always attempt to maximize their utilities. The activity-travel schedule of the people is generated through the application of these rules at various decision stages. Depicting travel behavior based on deterministic rules is however seen as a limitation of this modeling system since it does not deal with the uncertainty in human mobility. SCHEDULER [67], TASHA [54], and AMOS [68] are some early examples of these models.

Machine learning techniques such as neural networks, support vector machines, or decision trees have begun to be used in recent applications to extract rules from data. Since these techniques provide a higher predictive capability to identify and differentiate complex patterns of human mobility, the use of machine learning in activity-

based models receives increasing attention over the past decade [69].

Some of these models employing machine learning techniques are briefly reviewed here. ALBATROSS [70] is one of the first implementations of a rule-based machine learning approach using decision trees. The model uses the assumption that individuals make plans based on their priorities of activities. These activities groups are either fixed such as work or flexible such as daily shopping (See more in [71], and [72]). AgentPolis is an open-source simulation framework using neural networks, where individuals can dynamically replan their activities at any point in time [73]. Hafezi, Liu and Millward [69] proposes a modeling framework to explore and understand activity pattern clusters. Twelve clusters of homogeneous daily activity patterns were defined using a fuzzy C-means (FCM) clustering algorithm. The clustered data were used to deduce dependencies between activity type, activity sequence, and socio-demographic characteristics of individuals [74]. Individual daily activity schedules that consist of activity type and sequence were modeled by [75]. Model parameters were calculated using support vector machines (SVM). Recently, a data-driven activity scheduler (DDAS) using supervised machine learning methods was introduced by [76]. DDAS sequentially generates the activity schedule that consists of activity type, start-end time, location, and mode choice via four separate models.

The use of machine learning techniques makes activity generation relatively easy compared to traditional methods that depend on expert knowledge [76]. These techniques are widely used in many fields with their predictive abilities, robustness, and flexibility, but their use in predicting activity-travel behavior is not as common as in other fields [77]. Further, the spatio-temporal transferability of models using machine learning techniques has not been adequately tested. One general methodology that is applicable to any region does not exist yet. Another limitation of this technique is that many ML techniques lack interpretability and are designed as a black box. The lack of interpretability makes it difficult to inspect and understand how the algorithm predicts travel activity behaviors.

2.2.3 Simulation of travel and activity plans

In this section, the agent-based simulation of transportation is explained. The simulation models provide very detailed information

regarding each agent's trips, including the decision processes by moving individuals on networks based on their activity schedules. The travel demand described in the previous step is fed into simulation models, which handle route choice in the network, simulation of travels, and generation of interactions between agents and the environment. The simulation models also provide insight into the complex supply-demand relationship in the transportation system.

The most commonly used agent-based transportation simulation tool is Multi-Agent Transport Simulation (MATSim). It has the capacity to simulate high computational large-scale projects in a competitive time. Some of the recent implementations of MATSim transportation simulations are: (i) analysis of long-distance travel behavior of a fleet of vehicles converted to all-electric vehicles [78], (ii) evaluation of the impact of autonomous vehicles at different levels on people's mobility [79], (iii) providing an understanding of the complex relationship between supply and demand in carsharing systems [80].

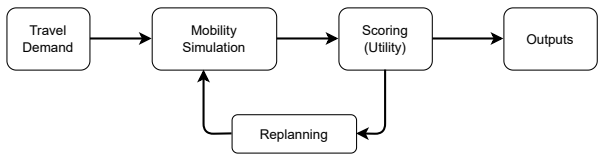


Figure 2.2: Mobility simulation steps by The Multi-Agent Transport Simulation (MATSim) tool. Source: Figure 1.1 from MATSim book [p4, 81].

The daily travel behavior of agents is simulated by executing daily activity plans using MATSim. The concept of a co-evolutionary algorithm is applied to optimize activity plans during the execution. This means each agent attempts to optimize its activity plan in an iterative process while they interact with agents in the model environment, and compete for the limited time-space slots [82]. The iterative process with the co-evolutionary algorithm results in a user equilibrium Horni, Nagel and Axhausen [81].

The MATSim algorithm consists of five steps with a loop, initial demand, mobility simulation, scoring, re-planning, and analysis steps

(Figure 2.2). It begins by feeding MATSim with the travel demand generated from the population's daily activity schedules. In every iteration, each agent performs its activities depending on its activity schedule using the transportation infrastructure, and then a score corresponding to their performance is calculated. The score combines activity utility and travel disutility [83]. The calculated scores are stored. A certain share of the agents are allowed to replan their activity-travel schedules, and the loop restarts. After a certain number of iterations, the model outputs are obtained at a higher spatio-temporal resolution.

CHAPTER 3

Summary of Appended Papers

3.1 Synthetic Sweden Mobility (SySMo) Model Documentation (Paper A)

The model documentation describes the methodology of the Synthetic Sweden Mobility (SySMo) Model consisting of population synthesis, activity generation, and location and mode assignment components.

3.1.1 Introduction

The Synthetic Swedish Mobility (SySMo) model is a large-scale transportation model developed with an agent-based modeling (ABM) approach. SySMo provides a scaffold to build decision support tools that play a part in identifying key drivers and areas where decision-makers need to improve measures to achieve their goals, such as climate targets or other societal aims. The model first generates a synthetic replica of the Swedish population with certain attributes associated with human transport behavior. Thereafter, the heterogeneous activity-travel schedules showing activity and trip features are added to the synthetic population. The model provides a platform to create various future scenarios while not violating any privacy issues since it is completely synthetic. E.g., SySMo model enables the analysis of innovations such as electric vehicles (EVs) and shared autonomous electric vehicles (SEAVs) that can play a role in the sustainable transformation of future transport systems.

The adopted agent-based modelling framework in SySMo corresponds to the disaggregated modeling approach, where each actor is modeled separately. As well as many other advantages, such as modeling actors' interaction with each other and the transportation infrastructure, ABM provides high spatio-temporal resolution. Another

fundamental feature of SySMo is that it employs an activity-based approach to generate the travel demand. This approach relies on the assumption that people travel to participate in activities and fits well into the agent-based modeling framework, in which each person's travel behaviors are modeled separately. Such a model constructs a complete activity schedule consisting of activities to be performed at different places at different times for all members of a population. The SySMo model is the first of its kind to depict the entire Swedish population and its travel behavior at a disaggregated level.

3.1.2 Methodology

The *Synthetic Sweden Mobility (SySMo) Model* consists of three key components: population synthesis, activity generation, and location and mode assignment. Figure 3.1 shows the model workflow describing how these three components are connected and the breakdowns under each component.

The first component is the population synthesis, where all agents are generated in three sub-steps. Each agent with basic attributes (age, gender, civil status, residential zone (DeSO), household size, and the number of children<6) is deduced at first. Subsequently, households are created using age, civil status, and household size attributes. The modeling process in this component is completed by computing the advanced attributes (employment and student statuses of agents, car ownership, and personal income) using a novel method that combines machine learning, iterative proportional fitting, and probabilistic sampling.

In the second component, the activity schedules characterized by activity sequence, type, duration, and start-end times are assigned to each agent. Based on the travel survey, a set of activity types showing daily activity participation is deduced. Following that, the daily total duration of each activity type for each individual is determined by ensuring that durations collectively satisfy consistency constraints. In the next step, an activity sequence is assigned to every individual by matching with a person from the travel survey based on the similarities between their attributes and activity types' durations. Ultimately, activity schedules are created for each individual.

The location and mode assignment component assigns locations to all activities in the sequences and travel modes to access

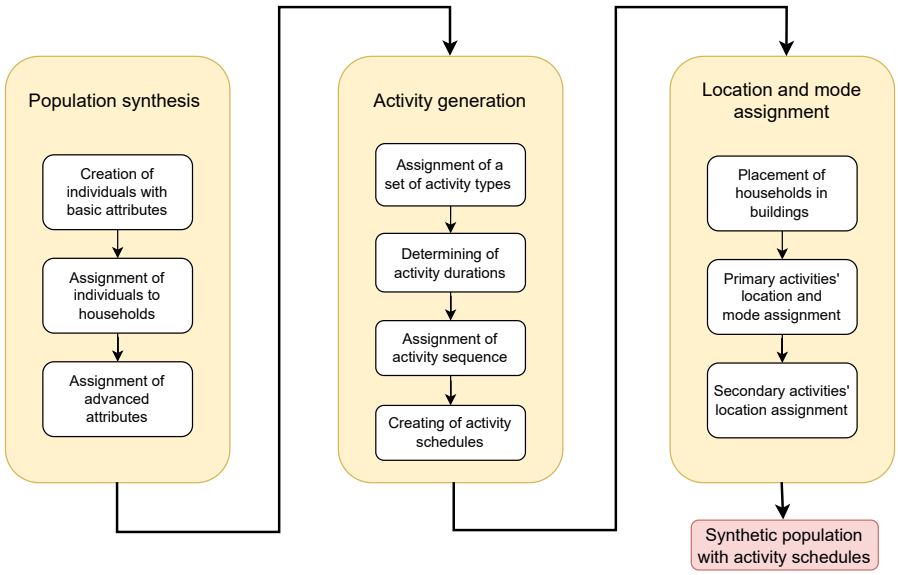


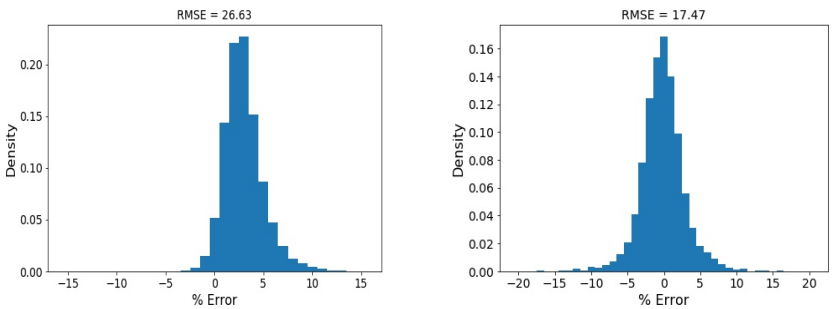
Figure 3.1: Methodology overview of *Synthetic Sweden Mobility (SySMo) Model*. Yellow rectangles: three main components of SySMo model with procedures of the calculations; pink rectangle: the final outputs, a spatially explicit agent-based mobility model.

activities. First, each household is spatially placed in a residential building, broadly classified into detached houses and apartment buildings. Thereafter, the locations of the primary activities and travel modes between activities are assigned using Origin-Destination (OD) matrices from Trafikverket (Swedish Transport Administration)'s Sampers model, or a variant of the gravity model based on the Swedish national travel survey. Finally, locations for the secondary activities whose locations depend on the locations of the primary activities are assigned, using a variant of the gravity model based on the Swedish national travel survey.

3.1.3 Results

This section briefly summarizes the SySMo model's results from population synthesis and location and mode assignment components. The results of the activity generation component are presented under Section 3.2 (Paper B).

The created population is first validated against data from Statistic Sweden [84]. We compute the percentage difference in the number of individuals with respect to basic attributes and advanced attributes in each DeSO zones (the average population is 1 706 people in each zone) and the distribution of the mean-square error (RMSE). Figure 3.2 a) shows the percentage error in the number of employees in each DeSO zones. The error is between -3% and 3% in more than 55 percent of the DeSO zones, and the RMSE is 26.63. Figure 3.2 b) shows the percentage error in the number of in the number of cars in each DeSO zones. The error is between -3% and 3% in more than 76 percent of the DeSO zones, and the RMSE is 17.47. Since these attributes are advanced attributes (i.e. derived based on the basic attributes) the error is slightly higher than basic attributes.



(a) Employees **(b)** Cars

Figure 3.2: The percent error in the number of employees in each DeSO zones(a) and the percent error in the number of cars in each DeSO zones(b).

To evaluate the location and mode assignment component and the previous steps, we looked at the predicted total distance traveled annually by travel modes. The passenger and goods transport statistics from Trafikanalys [85] are used to compare with our results. To calculate the actual road (network) distances, we multiply the car driver, and car passenger modes' Euclidean distances by $\sqrt{2}$ (see more regarding the calculation in the paper [86]). Data from the travel survey were also used for comparison. The comparison of annual total passenger kilometre show that our model results are very close to the Survey and Trafikanalys data (Table 3.1).

Table 3.1: Annual total passenger kilometres by mode in 2018 (in billions km)
In the Trafikanalys column, the numbers calculated using the old technique are on the left side, and on the right side are from the new technique.

Mode	SySMo weighted by weekdays and weekends	Trafikanalys	Survey
Car Driver+Passenger	98	95 - 116	113
Public Transport	24	26	30
Bike	3	2.8 - 3.1	3.3
Walking	4	2.0 - 3.7	3.8

3.2 The Heterogeneous Travel Activity of a Synthetic Population (Paper B)

The paper is concerned with the heterogeneous travel activity generation for the synthetic population defined in SySMo.

3.2.1 Introduction

Agent-based modeling (ABM) framework equipped with an activity-based travel demand generation approach is a pervasively adopted method by modelers to model travel behavior. Buliung and Kanaroglou [5] claim that the activity-based approach provides the most rigorous view of the transportation model in generating travel demand. The proper implementation of the activity generation component still plays a crucial role toward developing a model that accurately represents the population's mobility pattern.

Many models have developed so far utilizing the activity-based approach (e.g., Miller and Roorda [54], Arentze and Timmermans [70], Hafezi, Liu and Millward [74], Allahviranloo and Recker [75] and Drchal, Čertickỳ and Jakob [76]. Although the previous studies give accurate results in representing the population's travel behavior in an overall picture, they are inadequate in reflecting the heterogeneity within sub-populations. The human travel behavior is highly complex, and it will be an oversimplification to assume that this behavior is

homogeneous within a particular group characterized by various attributes. For instance, the studies exploring activity travel behaviours of senior people show the heterogeneity within the sub-population and provide better understanding of travel behavior [87, 88].

In this paper, we propose a novel methodology to capture the heterogeneity in activity generation among individuals in a synthetic population. Using machine learning in conjunction with probability models enables to maintain heterogeneity by sampling from the derived probability distributions of the attributes constituting the daily schedules. So that, while the model captures the overall distribution of attributes in the population, it allows for more targeted and precise studies of people's mobility with the heterogeneous structure.

3.2.2 Methodology

The activity generation framework has four major steps: assignment of a set of activity types, determination of the duration of each activity type, assignment of the activities sequence, and creation of activity schedules for each individual. Figure 3.3 illustrates the proposed activity generation workflow through four main steps and their connection.

The methodology's first step comprises the assignment of a set of activities showing the activity participation of each individual during a day using a machine learning method, neural network classifiers (NNC). The considered activity types are home (h), work (w), school (s), and other(o) activities. In the second main step, we deduce the daily total activity duration for each activity type in three sub-steps. Here, we train NNCs to jointly predict the broad duration classes that classify an individual's total activity time for different activities as low, moderate, or high. The total daily travel time range (TT) for each agent is then deduced. Based on the predicted broad duration classes and travel time range, an hourly duration for each activity type is estimated using NNCs. The hourly durations are assigned such that they collectively meet the constraint (eq. 3.1) implying the sum of the duration of the activity types within 24 hours minus the range of the day's total travel time.

$$24\text{h} - TT_{\text{lower limit}} \leq \sum_{n}^{h,w,s,o} \text{the duration of n} < 24\text{h} - TT_{\text{upper limit}} \quad (3.1)$$

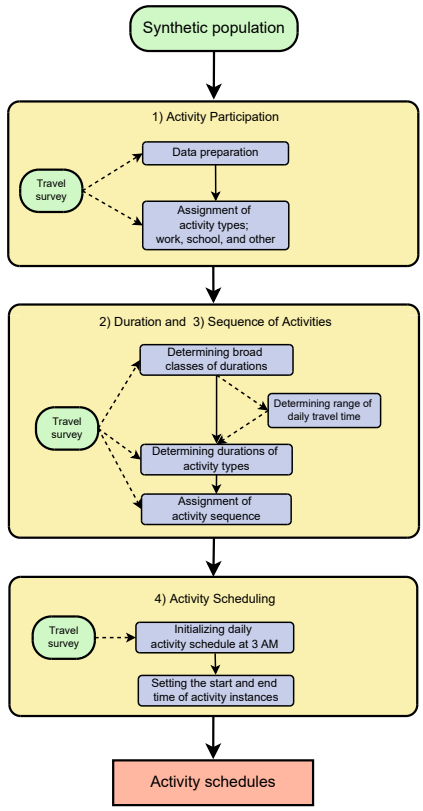


Figure 3.3: Methodology overview of the activity generation module of *Synthetic Sweden Mobility (SySMo) model*. Yellow rectangles: major steps of the activity generation; purple rectangles: steps of the calculations; green ellipses: input data; pink rectangle: model outputs of activity schedules for each individual.

The third step assigns an activity sequence to each individual through matching with an individual from the national travel survey using the agents' attributes and travel times. Finally, in the last main step we calculate the duration of activity instances in the schedules and create activity schedules containing activity type, activity sequence, and start and end times of activity instances for each individual.

3.2.3 Results

The proposed activity generation modeling framework is applied to the Swedish synthetic population created within the SySMo model. We reproduce heterogeneous daily activity schedules, including the synthetic population's activity type, start-end time, duration, and sequence. To evaluate the model results, we first compare the produced distributions by the model with the travel survey using distance metrics. We employ the Hellinger (H) and Jensen–Shannon (JS) distances having values in the range [0,1], where 1 means the maximum distance i.e., completely different distributions. Our calculated Hellinger and Jensen–Shannon distances are in the range [0.07,0.24] and [0.10,0.20], respectively, from various comparisons such as the activity duration distribution by gender, or income groups, the activity end-time distribution by activity type, and so on. These results show that the model accurately predicts distributions regarding the features of the activity schedule.

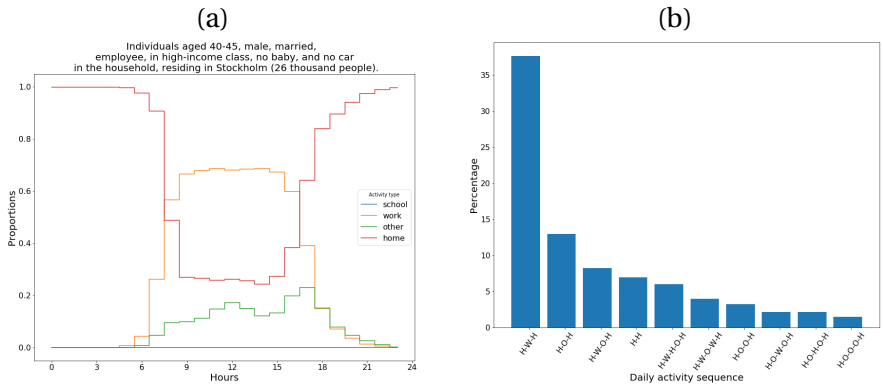


Figure 3.4: Activity pattern of the synthetic agents; aged 40-45, male, married, employee, in high-income class, no children ≤ 6 years old in household, and no car in household, residing in Stockholm. (a): Aggregated activity pattern of the sub-population by activity type, (b): Percentage of 10 most frequent daily activity sequences in the the sub-population (26 thousand agents in total).

Figure 3.4 illustrates one of the main results of the model depicting the heterogeneous activity patterns of a particular population group. The plotted set contains individuals with an age range of 40-45, male, married, employed, in high-income class, with no children ≤ 6 years

old in household, no car in household, and residing in Stockholm. In the figure, while Panel a shows aggregated activity patterns of the population set with the share of participation in different activity types during a day, Panel b depicts the frequency of the 10 most frequent daily activity sequences in the population. Even though the frequency of the activity pattern (H-W-H) is more than 35 percent, this is still below half of the agents and other activity sequences are present within the group.

CHAPTER 4

Discussion and outlook

This thesis contributes to the literature on agent-based modeling, a state-of-the-art method in transportation modeling. The focus is on the development of a sensitive tool to serve informed policy-making by evaluating future mobility vehicles. One of the advantages of the model is the large-scale advanced synthetic population with socio-economic attributes. Other studies are so far limited to small regions (i.e., Stockholm [7]) or focusing on the part of mobility behaviours (i.e., long distance trips [8]). Our methodology produces a statistically accurate representation of the whole population. Furthermore, the proposed methodology keeps the correlation between agent's attributes and mobility patterns using the advantages of ML and IPF methods. Having such a large population with attributes provides flexibility in scenario generation and also allows accurate traffic simulation. For instance, in one particular scenario, people's mobility can be analyzed by assigning only electric vehicles for commuting trips, while in another scenario, people whose income level is above a certain limit can be studied.

The SySMo model generates the agents with the activity-travel pattern for the base year 2018 since the most recent data is published in that year. It evaluates new mobility vehicles with today's population and their mobility behaviors on the existing infrastructure. In a developed country, Sweden, one can assume that the infrastructure and mobility patterns will not change in a short time. However, this assumption raises questions regarding the representativeness of the population for long-term evaluations. As a basic example, the population of Sweden has increased approximately by 15 percent from 1990 to 2020, and it can be considered that the trend will continue in the next years. Future developments of the model will focus on the projection of the population. There are well-known methods (see

more in the papers [45, 48, 89]) to project the population to certain years. Projecting the synthetic population into the future provides more targeted policy-making and more detailed assessments of the future.

ABM is described as a model producing system-wide outputs through behaviors of autonomous agents performing cooperative or competitive interactions with one another (see more in Chapter 2). Modeling the agents requires a lot of detail based on their two foundational characteristics, autonomy and interaction, while the practice of modeling actors involves many abstractions [5, 90]. The agents' interactions are implicitly modeled in our model. The methodology described so far includes the preparation of initial travel demand for all agents. The interaction of the agents and their learning from actions will be carried out together with the scenario analysis during the agent-based travel simulation.

References

[1] L. M. Fulton (2018). Three revolutions in urban passenger travel. *Joule* **2** (4), pp. 575–578.

[2] M. Matyas and M. Kamargianni (2019). The potential of mobility as a service bundles as a mobility management tool. *Transportation* **46** (5), pp. 1951–1968.

[3] *Urban Development* (2022). URL: `https://www.worldbank.org/en/topic/urbandevelopment/overview#1` (Retrieved: 2022-05-10).

[4] P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, D. M. R. van Diemen, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz and J. Malley (2022). *Ipcc, 2022: climate change 2022: mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change*. Tech. rep.

[5] R. N. Buliung and P. S. Kanaroglou (2007). Activity–travel behaviour research: conceptual issues, state of the art, and emerging perspectives on behavioural analysis and simulation modelling. *Transport Reviews* **27** (2), pp. 151–187.

[6] S. Rasouli and H. Timmermans (2014). Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences* **18** (1), pp. 31–60.

[7] O. Canella, G. Flötteröd, D. Johnsson, I. Kristoffersson, P. Larek and J. Thelin (2016). *Flexible coupling of disaggregate travel demand models and network simulation packages (ihop2)*. Tech. rep. Technical report, KTH, Sweco, WSP.

[8] F. J. Márquez-Fernández, J. Bischoff, G. Domingues-Olavarrıéa and M. Alaküla (2021). Assessment of future ev charging infra-

structure scenarios for long-distance transport in sweden. *IEEE Transactions on Transportation Electrification*.

[9] D. A. Hensher, K. J. Button, K. E. Haynes and P. R. Stopher (2004). *Handbook of transport geography and spatial systems*. Vol. 5. Emerald Group Publishing Limited, pp. 13–26.

[10] E. Cascetta (2013). *Transportation systems engineering: theory and methods*. Vol. 49. Springer Science & Business Media.

[11] T. Litman (2017). *Understanding transport demands and elasticities*. Victoria Transport Policy Institute Victoria, BC, Canada.

[12] J. de Dios Ortúzar and L. G. Willumsen (2011). *Modelling transport*. John wiley & sons.

[13] E. Weiner (1997). *Urban transportation planning in the united states: an historical overview*. US Department of Transportation.

[14] S. Wise, A. Crooks and M. Batty (2016). 'Transportation in agent-based urban modelling'. In: *International workshop on agent based modelling of urban systems*. Springer, pp. 129–148.

[15] R. van Nes and G. de Jong (2020). 'Transport models'. In: *Advances in transport policy and planning*. Vol. 6. Elsevier, pp. 101–128.

[16] Y. Xiong, J. Gan, B. An, C. Miao and Y. C. Soh (2016). 'Optimal pricing for efficient electric vehicle charging station management'. In: *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 749–757.

[17] D. Dias, O. Tchepel and A. P. Antunes (2016). Integrated modelling approach for the evaluation of low emission zones. *Journal of environmental management* **177**, pp. 253–263.

[18] M. D. Galus, R. A. Waraich, F. Noembrini, K. Steurs, G. Georges, K. Boulouchos, K. W. Axhausen and G. Andersson (2012). Integrating power systems, transport systems and vehicle technology for electric mobility impact assessment and efficient control. *IEEE Transactions on Smart Grid* **3** (2), pp. 934–949.

[19] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini and M.

Tomasini (2018). Human mobility: models and applications. *Physics Reports* **734**, pp. 1–74.

[20] M. S. Daskin (1985). *Urban transportation networks: equilibrium analysis with mathematical programming methods*.

[21] T. Adler and M. Ben-Akiva (1979). A theoretical and empirical model of trip chaining behavior. *Transportation Research Part B: Methodological* **13** (3), pp. 243–257.

[22] S. ALGERS, J. ELIASSON and J. KOEHLER (2005). Microsimulating the stockholm integrated model system (sims). *PROCEEDINGS OF ETC 2005, STRASBOURG, FRANCE 18-20 SEPTEMBER 2005-RESEARCH TO INFORM DECISION-MAKING IN TRANSPORT APPLIED METHODS/INNOVATIVE METHODS-ACTIVITY BASED MODELS*.

[23] Y. Shiftan (1998). Practical approach to model trip chaining. *Transportation research record* **1645** (1), pp. 17–23.

[24] T. F. Rossi and Y. Shiftan (1997). Tour based travel demand modeling in the us. *IFAC Proceedings Volumes* **30** (8), pp. 381–386.

[25] H. W. David Boyce (2015). *Forecasting urban travel: past, present and future*. Edward Elgar Press.

[26] D. Damm and S. R. Lerman (1981). A theory of activity scheduling behavior. *Environment and Planning A* **13** (6), pp. 703–718.

[27] R. Kitamura (1984a). A model of daily time allocation to discretionary out-of-home activities and trips. *Transportation Research Part B: Methodological* **18** (3), pp. 255–266.

[28] P. M. Jones (2021). 'New approaches to understanding travel behaviour: the human activity approach'. In: *Behavioural travel modelling*. Routledge, pp. 55–80.

[29] T. Hägerstrand (1970). 'What about people in regional science?' In: *Papers of the regional science association*. Vol. 24.

[30] E. I. Pas (1984). The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environment and Planning A* **16** (5), pp. 571–581.

[31] S. Hanson (1982). The determinants of daily travel-activity patterns: relative location and sociodemographic factors. *Urban Geography* **3** (3), pp. 179–202.

[32] R. Kitamura (1984b). Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B: Methodological* **18** (1), pp. 67–81.

[33] T. F. Golob and M. G. McNally (1997). A model of activity participation and travel interactions between household heads. *Transportation Research Part B: Methodological* **31** (3), pp. 177–194.

[34] C. G. Pooley, D. Horton, G. Scheldeman, M. Tight, T. Jones, A. Chisholm, H. Harwatt and A. Jopson (2011). Household decision-making for everyday travel: a case study of walking and cycling in lancaster (uk). *Journal of Transport Geography* **19** (6), pp. 1601–1607.

[35] M. G. McNally and C. R. Rindt (2007). 'The activity-based approach'. In: *Handbook of transport modelling*. Emerald Group Publishing Limited.

[36] R. Kitamura (1996). 'Applications of models of activity behavior for activity based demand forecasting'. In: *Activity-based travel forecasting conference, new orleans, louisiana*.

[37] A. R. Pinjari and C. R. Bhat (2011). 'Activity-based travel demand analysis'. In: *A handbook of transport economics*. Edward Elgar Publishing, pp. 213–248.

[38] D. Gopalakrishna, E. Schreffler, D. Vary, D. Friedenfeld, B. Kuhn, C. Dusza, R. Klein and A. Rosas (2012). *Integrating demand management into the transportation planning process: a desk reference*. Tech. rep., pp. 146–154.

[39] J. M. Epstein and R. Axtell (1996). *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

[40] J. H. Miller, S. E. Page and B. LeBaron (2008). Complex adaptive systems: an introduction to computational models of social life. *Journal of Economic Literature* **46** (2), pp. 427–428.

[41] D. Chen (2009). 'A grid aware large scale agent-based simulation system'. In: *Quantitative quality of service for grid computing: applications for heterogeneity, large-scale distribution, and dynamic environments*. IGI Global, pp. 299–319.

[42] E. Guerci, M. A. Rastegar and S. Cincotti (2010). 'Agent-based modeling and simulation of competitive wholesale electricity

markets'. In: *Handbook of power systems ii*. Springer, pp. 241–286.

[43] D. Sornette (2004). A complex system view of why stock markets crash. *New Thesis* **1** (1), pp. 5–18.

[44] D. Charypar (2008). 'Efficient algorithms for the microsimulation of travel behavior in very large scenarios'. PhD thesis. ETH Zurich.

[45] J. Rich (2018). Large-scale spatial population synthesis for denmark. *European Transport Research Review* **10** (2), p. 63.

[46] K. Müller and K. W. Axhausen (2010). Population synthesis for microsimulation: state of the art. *Arbeitsberichte Verkehrs-und Raumplanung* **638**.

[47] D. Ballas, G. Clarke, D. Dorling and D. Rossiter (2007). Using simbritain to model the geographical impact of national government policies. *Geographical Analysis* **39** (1), pp. 44–77.

[48] R. Tanton et al. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation* **7** (1), pp. 4–25.

[49] L. Smith, R. Beckman, K. Baggerly, D. Anson and M. Williams (1995). *Transims: project summary and status may 1995*. Tech. rep. Los Alamos National Laboratory Report prepared for US Department of ...

[50] M. Frick (2004). Generating synthetic populations using ipf and monte carlo techniques: some new results. *Arbeitsberichte Verkehrs-und Raumplanung* **225**.

[51] T. Arentze, H. Timmermans and F. Hofman (2007). Creating synthetic household populations: problems and approach. *Transportation Research Record* **2014** (1), pp. 85–91.

[52] J. Y. Guo and C. R. Bhat (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record* **2014** (1), pp. 92–101.

[53] J. Castiglione, M. Bradley and J. Gliebe (2015). *Activity-based travel demand models: a primer*. SHRP 2 Report S2-C46-RR-1. Transportation Research Board.

[54] E. J. Miller and M. J. Roorda (2003). Prototype model of household activity-travel scheduling. *Transportation Research Record* **1831** (1), pp. 114–121.

[55] J. Auld and A. K. Mohammadian (2012). Activity planning processes in the agent-based dynamic activity planning and travel scheduling (adapts) model. *Transportation Research Part A: Policy and Practice* **46** (8), pp. 1386–1403.

[56] B. Lenntorp (1977). Paths in space-time environments: a time-geographic study of movement possibilities of individuals. *Environment and Planning A* **9** (8), pp. 961–972.

[57] N. A. Khan (2020). Modelling and microsimulation of activity generation, activity scheduling and mobility assignment.

[58] P. M. Jones, M. C. Dix, M. I. Clarke and I. G. Heggie (1983). *Understanding travel behaviour*. Gower Publishing.

[59] M. Dijst and V. Vidakovic (1997). Individual action space in the city. *Activity-based approaches to travel analysis*.

[60] M.-P. Kwan (1997). Gisicas: an activity-based travel decision support system using a gis-interfaced computational-process model. *Activity-based approaches to travel analysis*.

[61] M. H. Hafezi, H. Millward and L. Liu (2018). 'Activity-based travel demand modeling: progress and possibilities'. In: *International conference on transportation and development 2018: planning, sustainability, and infrastructure systems*. American Society of Civil Engineers Reston, VA, pp. 138–147.

[62] M. E. Ben-Akiva and J. L. Bowman (1998). 'Activity based travel demand model systems'. In: *Equilibrium and advanced transportation modelling*. Springer, pp. 27–46.

[63] J. L. Bowman and M. E. Ben-Akiva (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice* **35** (1), pp. 1–28.

[64] C. R. Bhat, J. Y. Guo, S. Srinivasan and A. Sivakumar (2004). Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record* **1894** (1), pp. 57–66.

[65] R. Kitamura, S. Fujii et al. (1998). Two computational process models of activity-travel behavior. *Theoretical foundations of travel choice modeling*, pp. 251–279.

[66] P. Vovsha and K.-A. Chiao (2006). Development of new york metropolitan transportation council tour-based model. *Innovations in Travel Demand Modeling*, p. 21.

[67] T. Gärling, M.-p. Kwan and R. G. Golledge (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B: Methodological* **28** (5), pp. 355–364.

[68] R. M. Pendyala, R. Kitamura, C. Chen and E. I. Pas (1997). An activity-based microsimulation analysis of transportation control measures. *Transport Policy* **4** (3), pp. 183–192.

[69] M. H. Hafezi, L. Liu and H. Millward (2019). A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation* **46** (4), pp. 1369–1394.

[70] T. Arentze and H. Timmermans (2000). *Albatross: a learning based transportation oriented simulation system*. EIRASS.

[71] D. Ettema, A. Borgers and H. Timmermans (1993). Simulation model of activity scheduling behavior. *Transportation Research Record*, pp. 1–1.

[72] S. T. Doherty (2000). 'An activity scheduling process approach to understanding travel behavior'. In: *79th annual meeting of the transportation research board, washington, dc*. Citeseer.

[73] M. Čertickỳ, J. Drchal, M. Cuchỳ and M. Jakob (2015). 'Fully agent-based simulation model of multimodal mobility in european cities'. In: *2015 international conference on models and technologies for intelligent transportation systems (mt-its)*. IEEE, pp. 229–236.

[74] M. H. Hafezi, L. Liu and H. Millward (2018). Learning daily activity sequences of population groups using random forest theory. *Transportation research record* **2672** (47), pp. 194–207.

[75] M. Allahviranloo and W. Recker (2013). Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological* **58**, pp. 16–43.

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

[77] A. N. Koushik, M. Manoj and N. Nezamuddin (2020). Machine learning applications in activity-travel behaviour research: a review. *Transport reviews* **40** (3), pp. 288–311.

[78] F. J. Márquez-Fernández, J. Bischoff, G. Domingues-Olavarría and M. Alaküla (2019). 'Using multi-agent transport simulations to assess the impact of ev charging infrastructure deployment'. In: *2019 ieee transportation electrification conference and expo (itec)*, pp. 1–6. DOI: 10.1109/ITEC.2019.8790518.

[79] J. Hamadneh and D. Esztergár-Kiss (2021). The influence of introducing autonomous vehicles on conventional transport modes and travel time. *Energies* **14** (14), p. 4163.

[80] L. M. Martıénez, G. H. d. A. Correia, F. Moura and M. Mendes Lopes (2017). Insights into carsharing demand dynamics: outputs of an agent-based model application to lisbon, portugal. *International Journal of Sustainable Transportation* **11** (2), pp. 148–159.

[81] A. Horni, K. Nagel and K. Axhausen, eds. (2016). *Multi-agent transport simulation matsim*. London: Ubiquity Press, p. 618. ISBN: 978-1-909188-75-4, 978-1-909188-76-1, 978-1-909188-77-8, 978-1-909188-78-5. DOI: 10.5334/baw.

[82] K. Nagel and F. Marchal (2003). Computational methods for multi-agent simulations of travel behavior. *Proceedings of International Association for Travel Behavior Research (IATBR), Lucerne, Switzerland*.

[83] K. Nagel, B. Kickhöfer, A. Horni and D. Charypar (2016). *A closer look at scoring*.

[84] *Statistics Sweden* (2020). https://www.statistikdatabasen.scb.se/pxweb/en/ssd/.

[85] *Passenger and goods transport report* (2021). URL: https://www.trafa.se/ovrig/transportarbete/ (Retrieved: 2021-11-10).

[86] C. L. Barrett, R. J. Beckman, K. Maleq, V. A. Kumar, M. V. Marathe, P. E. Stretz, T. Dutta and B. Lewus (2009). 'Generation and ana-

lysis of large synthetic social contact networks'. In: *Proceedings of the 2009 winter simulation conference m*. Winter Simulation Conference. ISBN: 9781424457717.

[87] D. Yang, H. Timmermans and A. Grigolon (2013). Exploring heterogeneity in travel time expenditure of aging populations in the netherlands: results of a chaid analysis. *Journal of Transport Geography* **33**, pp. 170–179.

[88] J. W. Hutchinson (2018). 'Exploring patterns of heterogeneity in activity–travel behaviors of older people'. PhD thesis.

[89] J. Li, C. O'Donoghue et al. (2013). A survey of dynamic microsimulation models: uses, model structure and methodology. *International Journal of microsimulation* **6** (2), pp. 3–55.

[90] J. Odell (2002). Objects and agents compared. *Journal of object technology* **1** (1), pp. 41–53.

Paper A

Synthetic Sweden Mobility (SySMo) Model Documentation

Synthetic Sweden Mobility (SySMo) Model Documentation

Çağlar Tozluoğlu, Swapnil Dhamal, Yuan Liao, Sonia Yeh, Frances Sprei
Department of Space, Earth and Environment
Devdatt Dubhashi
Department of Computer Science
Chalmers University of Technology, Gothenburg, Sweden

Madhav Marathe, Christopher Barrett,
Department of Computer Science
University of Virginia, Virginia, United States

Version 1.0

May 19, 2022

Author Contributions: Conceptualization: Çağlar Tozluoğlu (Ç.T.), Swapnil Dhamal (S.D.), Yuan Liao(Y.L.), Sonia Yeh (S.Y.), Frances Sprei (F.S.), Devdatt Dubhashi (D.D.), Madhav Marathe (M.M.), Christopher Barrett (C.B.); methodology: S.D., Ç.T., S.Y., F.S.; software: S.D., Ç.T.; validation: Ç.T., S.D.; data curation: Ç.T., S.D.; writing - original draft: S.D., Ç.T.; writing - review & editing: Ç.T., ,S.Y., F.S., Y.L.; project administration: S.Y., F.S.

Department of Space, Earth and Environment
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
www.chalmers.se

Contents

List of Figures

List of Tables

Abstract

This document describes a decision support framework using a combination of several state-of-the-art computing tools and techniques in synthetic information systems, and large-scale agent-based simulations. In this work, we create a synthetic population of Sweden and their mobility patterns that are composed of three major components: population synthesis, activity generation, and location assignment. The document describes the model structure, assumptions, and validation of results.

Chapter 1

Introduction

"Synthetic Sweden" is a large-scale agent-based model (ABM) that provides a scaffold on which to build decision support tools to model and analyze future mobility scenarios. It replicates a statistically accurate representation of the real population of Sweden, but is completely synthetic so that (a) it does not violate any privacy issues and (b) it can be modified easily to create alternative scenarios. It is the latter feature that makes the model an ideal tool for modeling and analyzing future scenarios. The modeling tool can be a valuable planning and visualization tool for public and private stakeholders in Sweden. In additional, the methodology can be broadly applied to other regions with new data and carefully calibrated parameters.

Agent-based models (ABM) and activity-based travel demand models are often combined [1]. As well as many other advantages, activity-based demand generation fits well into the paradigm of multi-agent simulation, where each traveler is kept as an individual throughout the entire modeling process. Such a model provides the travel behavior of each individual agent by creating sequences of activities to be performed at different places at different times during a given period of time, such as one day.

The activity-based modeling approach constructs a complete activity plan for each member of a population, and derives the transportation demand from the fact that consecutive activities at different locations are connected by travel via certain modes such as walking, biking, cars, buses, etc. So, the two important aspects of activity-based travel demand modeling are activity generation and location assignment. Activity generation is concerned with the types, start times, and durations of the different activities, along with their sequence. Location assignment dictates the locations of activities and hence, the origins and destinations of trips.

1.1 Model overview

The *Synthetic Sweden Mobility (SySMo) Model* is comprised of three key components: population synthesis, activity generation, and location and mode assignment. We first briefly describe how these three components are connected, then we explain the methodology of each component in detail (Chapters 3-5). Fig. 1.1 shows a schema of the methodology. The key modeling components are connected in the following ways:

1. Population synthesis (Chapter 3)

    (a) Based on DeSO-level (Demographic statistical areas, see Section 2) data regarding age and gender distribution, and municipality-level data regarding the distribution of civil status-age-gender, create a synthetic population with basic attributes: civil status, age, gender using iterative proportional fitting (IPF).

   (b) Based on DeSO-level data regarding the household types and municipality-level data regarding the distribution of number of children per household, assign a household to each individual of the synthetic population; first accounting for adults (singles, couples, others) and then children.

   (c) Based on SCB data and data from travel survey, use machine learning (ML) and IPF to assign advanced attributes to individuals: employment and student statuses, personal and household incomes, car Ownership, etc.

2. Activity generation (Chapter 4)

   (a) Based on the travel survey, assign a set of activity types to each individual using ML

   (b) Based on the travel survey and activity participation of individuals, determine duration of each activity type for each individual while ensuring that durations collectively satisfy certain consistency constraints.

   (c) Assign an activity sequence to every individual by matching with a person from the travel survey based on the similarities between their attributes and activity types' durations.

   (d) Create activity schedules for each individual.

3. Location and mode assignment (Chapter 5)

   (a) Spatially place households in residential buildings, broadly classified into detached houses and apartment buildings.

   (b) Assign locations for the primary activities and travel modes between activities, using Origin-Destination (OD) matrices from Trafikverket (Swedish Transport Administration)'s Sampers model, or a variant of gravity model based on Swedish national travel survey.

   (c) Assign locations for the secondary activities whose locations depend on the locations of the primary activities, using a variant of gravity model based on Swedish national travel survey.

The travel behavior of an individual, as well as the overall population, on a weekend is significantly different from that on a weekday. Thus, modeling the daily travel pattern corresponding to an average day of the week would capture neither a weekday nor a weekend accurately. Hence, in the SySMo model, we model daily travel patterns corresponding to two types of days: an average weekday and an average weekend.

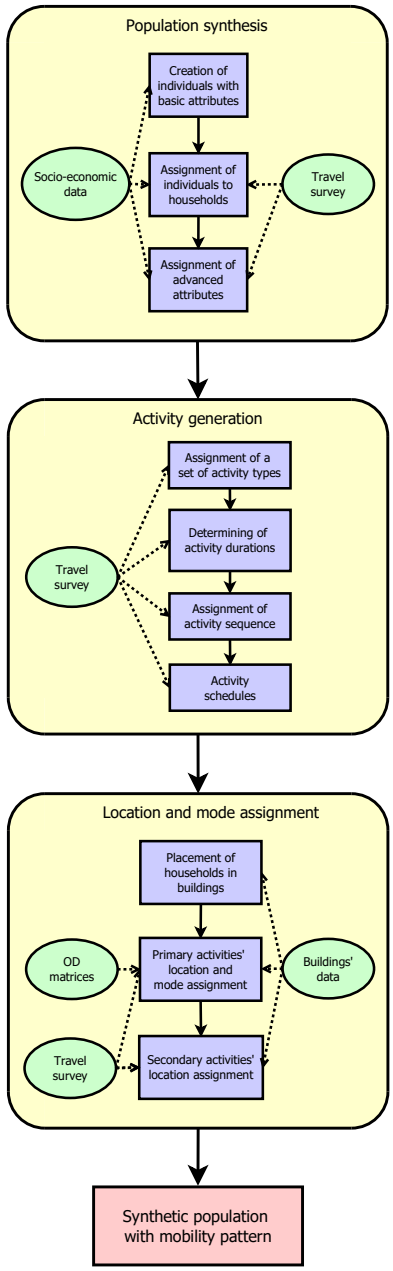


Figure 1.1: Methodology overview of *Synthetic Sweden Mobility (SySMo) Model*. Yellow rectangles: three main components of SySMo model; blue rectangles: procedures of the calculations; green ellipses: input data for modeling and calibration; pink rectangle: the final outputs, a spatially explicit agent-based mobility model.

Chapter 2

Data Description

There are four main sources of data for building and calibrating *SysMo*: statistical data from Statistics Sweden (SCB) (Section 2.1), Swedish national travel survey (Section 2.2), Origin-Destination (OD) matrices from Trafikverket (Swedish Transport Administration)'s model – Sampers (Section 2.3), and buildings from Lantmäteriet (Section 2.1). Data from Transport Analysis agency (Section 2.5) is utilised to validate SySMo model. The SCB statistics and the travel survey used to construct the model are used for in sample validation as well (See more in chapter 6). We present a brief description of the data in the sections below. Other data are explained elsewhere in the documentation where suitable.

2.1 Statistical data of Sweden

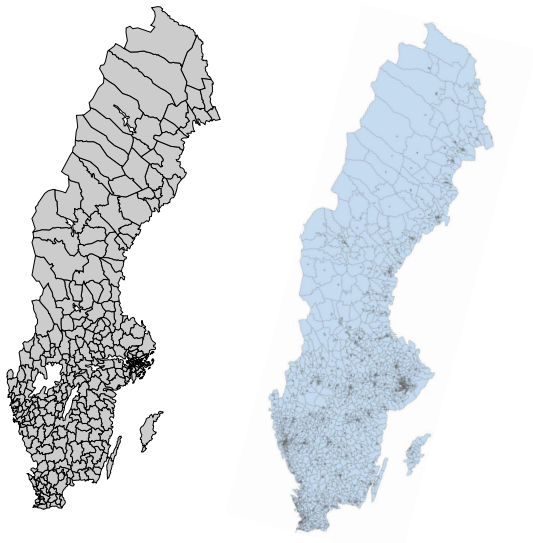
Statistics Sweden (SCB) [2] produces the official statistics at various geographical levels such as municipality or zone system. Fig. 2.1a shows the boundaries of 290 municipalities which act as local government entities. The statistical data at municipal areas are the number of individuals with a given combination of gender, age group, and civil status, number of children belonging to different household types, number of individuals belonging to different income classes, average household income of individuals in a given age group belonging to a given household type, and number of employees by industry types.

SCB also provides data at a zone level called Demographic Statistical Areas (DeSO) [3]. DeSO zones follow municipal boundaries and each municipality consists of a number of DeSO zones, for a total of 5,984 DeSO zones in Sweden (Fig. 2.1b). Each DeSO zone typically has between 700 and 2,700 inhabitants. The data utilized at DeSO zone level are the number of males and females, number of individuals belonging to different age groups, number of households of different types (single, couple, other), number of employees and students, and number of cars.

Sweden is also divided into sq.km. (square kilometer) grids, whose primary purpose is to capture the density of population in different regions. In this grid system, statistics on the registered population are presented in 114161 square areas covering only populated areas within Sweden.

2.2 Swedish national travel survey

The Swedish national travel survey [4] provides the data about the travel behaviour of anonymized individuals in conjunction with data on their socio-economic and geographical characteristics. The survey period is between 2011 and 2016, and consists of around 40000 participants aged 6-84 years. The travel survey was conducted with individuals, not households. However, the survey respondents provide some information regarding the household and its members such as number of people in the household. Activity location information of individuals is deduced from the start and end point of travel and activities are broadly classified as home, work, school, and



(a) municipal areas (b) DeSO zones

Figure 2.1: Swedish nation-wide geographic subdivisions

other.

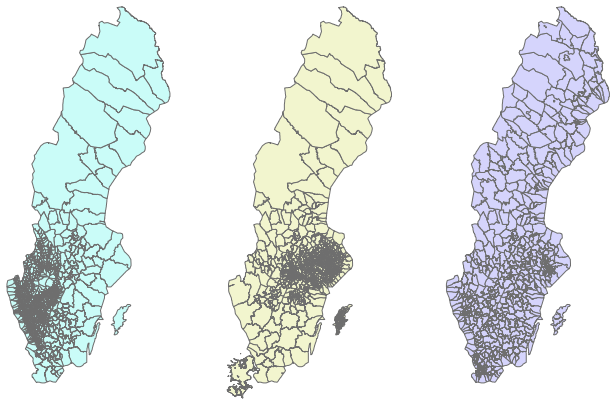
Each participant has one weight V_k according to their socio-demographics and another weight V_d based on the day the participant conducted the survey. These weights directly indicate the representative power of the respondent regarding socio-demographics or travel patterns. The total population can be generated using these weights.

In our model, we use the travel survey to train our ML algorithms and obtain various characteristics of our synthtic population such as employment and studenthood statuses, activity sequence, activity start-end times, activity durations, distances traveled, and trip modes.

2.3 The origin-destination (OD) matrices

Sampers [5], is a national transportation model developed by Trafikverket (Swedish Transport Administration) to do traffic analyses of passenger transport across Sweden. Predicting future traffic flows, evaluating new investments, and analyzing the impact of transportation policies are among the main uses of the model. The travel analyses can be carried out at the national or regional level.

Sampers consists of five regional models that are Palt, Samm, Skåne, Sydost, Väst and a national model covering the whole of Sweden. The national model consists of 682 zones, while the regional models provide data with a higher spatial resolution with a total of more than 10,000 zones. The national model captures only long-distance trips (more than 100 km). Each regional model consists of zones of different sizes. In a core area of a regional model, there are zones with a division into very fine zones. A core area is bordered by a ring area that usually consists of zones that are not as fine. The zones in remote areas representing the rest of Sweden are quite coarse.(Fig. 2.2). From Trafikverket, we received Samm and Väst regional models, which cover the two largest cities in Sweden: Stockholm and Gothenburg respectively, and the national model. These models contain information regarding short and long distance



(a) Väst regional (b) Samm regional (c) National

Figure 2.2: Zone systems of Swedish Sampers transportation model: regional (Väst and Samm) and national.

OD matrices by modes of transport (car, bike, walk, public transport) and by trip purposes (work, business, other, and private). Fig. 2.2 shows the zone systems in the two regional and the national models.

2.4 Buildings

The building data is adapted from the property registers covering all Sweden. It is in vector format and provide by Lantmäteriet [6]. The data contains more than 8.6 million buildings with its location, geometry, and type by usage purpose. We use the data to determine the home locations of the agents and where their activities take place. While assignment of individuals' activity locations at zone levels suffices for an aggregate analysis, we assign all activities to buildings to have higher spatial resolution in SySMo. Assigning the activities performed by agents to the buildings locations makes it possible to do more precise spatial analysis.

For assignment of residential buildings in SySMo, we use two main building types, which we create by combining the subcategories in the data: detached houses and apartments. Along the same lines, work, school, and other main categories are created from the subcategories in the building type and so each building is used for the activity assignment procedure by activity type.

2.5 Data on distance travelled

Transport Analysis is an agency established to produce official statistics on transport in Sweden. To validate the model results in SySMo, we use annual total distances travelled by modes of transport (Transportarbete) [7] generated by using calculation techniques and models. The data is available from 2000 to 2020. After 2016 they publish two values per year since the agency adopted a new method for calculating the total distance travelled, thus both values based on both the old and new method are presented.

The statistics includes the four main modes of transport road, rail, aviation and shipping and their respective subgroups. Road transport is divided into passenger car, bus, motorcycle, moped, bicycle and walking. For rail transport, modes of travel by rail, tram and metro are included. We use the statistics on road and rail modes only to validate the result of SySMo, i.e., they are not used as an input to the model.

Chapter 3

Population Synthesis

The attributes of individuals are classified into basic and advanced. We first synthesize the individuals along with their basic attributes. These consist of age, gender, civil status, and residential zone. We then assign individuals advanced attributes, i.e., employment and student statuses, personal income, and car ownership. Table 3.1 summarizes the variables that represent the different attributes used in the presentation of the methodology.

Table 3.1: Variable for describing individuals.

Variable	Description	Subcategories
g	gender	Male, Female
a	age group	0, 1-6, 7-15, 16-18, 19-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-75, 75-84, 85+
c	civil (marital) status	Single, Couple, Child
z_m	municipality zone	-
z_d	DeSO zone	-
ψ_W	employment status	Employed, Not employed
ψ_S	student status	Student, Not student
ρ^c	personal income class	0, [1, 180K), [180K, 300K), [300K, 420K), [420K, 1M)
n	number of cars owned	0,1,2,3

The procedures and assumptions are described in detail in the sections below.

3.1 Assigning basic attributes

For synthesizing individuals and their basic attributes, data for gender (i.e., number of males and females) and age (i.e., number of individuals belonging to different age groups) are available at the DeSO level. The data for the number of individuals with a given combination of gender, age group, and civil status (single, couple, or child) are available at the municipality level.

Let $N(z_d, a)$ denote the desired number of agents belonging to age group a in DeSO zone z_d. Similarly, let $N(z_d, g)$ denote the desired number of agents belonging to gender g in DeSO zone z_d. Let $N(z_m, a, g, c)$ denote the desired number of agents belonging to the combination of age group a, gender g, and civil status c, in municipality zone z_m. Let A, G, C be the sets of age group, gender, and civil status, respectively. We consider $A = \{0, 1\text{-}6, 7\text{-}15, 16\text{-}18, 19\text{-}24, 25\text{-}29, 30\text{-}34, 35\text{-}44, 45\text{-}54, 55\text{-}64, 65\text{-}75, 75\text{-}84, 85+\}$; $G = \{$'male', 'female'$\}$; $C = \{$'single', 'couple', 'child'$\}$. We use i to denote a typical agent and k to denote a typical household. Let $n(z_d, a, g, c)$

denote the deduced number of agents belonging to the combination of age group a, gender g, and civil status c, in DeSO zone z_d.

The iterative proportional fitting (IPF) procedure is used to deduce $n(z_d, a, g, c), \forall z_d, a, g, c$ (i.e., the number of agents belonging to every combination of DeSO zone z_d, age group a, gender g, and civil status c). In particular, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached (in our implementation, we consider 20 iterations).

$\forall z_d, a, g, c$:

$$n(z_d, a, g, c) \leftarrow \frac{N(z_m, a, g, c)}{\sum_{z_{d'} \in z_m} n(z_{d'}, a, g, c)} n(z_d, a, g, c) \tag{3.1}$$

where $z_m \ni z_d$,

$$n(z_d, a, g, c) \leftarrow \frac{N(z_d, a)}{\sum_{\substack{g' \in G \\ c' \in C}} n(z_d, a, g', c')} n(z_d, a, g, c) \tag{3.2}$$

$$n(z_d, a, g, c) \leftarrow \frac{N(z_d, g)}{\sum_{\substack{a' \in A \\ c' \in C}} n(z_d, a', g, c')} n(z_d, a, g, c) \tag{3.3}$$

Equation (3.1) scales the deduced number $n(z_d, a, g, c)$ on DeSO zone level by the ratio of the desired number $N(z_m, a, g, c)$ on municipality zone level to the number obtained in an iteration on municipality zone level, so as to drive the obtained number towards the desired number. Eqs. 3.2 and 3.3 drive the numbers obtained in an iteration toward the desired numbers of age and gender, respective, at DeSO zone level. The numbers $n(z_d, a, g, c)$ are finally rounded to the nearest integer. Note that since the last step corresponds to scaling with respect to the gender data on the DeSO zone level, the obtained population would be exactly consistent (up to a round-off error) with the gender data on the DeSO zone level.

We initialize the number of agents belonging to a given combination of gender, age group, and civil status on DeSO zone level, by dividing the desired number of agents belonging to that combination on the municipality level into the number of DeSO zones belonging to that municipality. That is,

$$\forall z_d, a, g, c : \quad n(z_d, a, g, c) \leftarrow \frac{N(z_m, a, g, c)}{|z_m|}, \quad \text{where } z_m \ni z_d \tag{3.4}$$

Here, $z_m \ni z_d$ denotes that municipality zone z_m contains DeSO zone z_d, and $|z_m|$ is the size of the municipality zone (i.e., the number of DeSO zones constituting the municipality).

This simulation hence synthesizes $n(z_d, a, g, c)$ number of agents having the combination of corresponding basic attributes, namely, DeSO zone z_d, age group a, gender g, and civil status c.

3.2 Creating households

The second key step in the synthetic population is the creation of households of different types (couple, single, and other) and assigning children to the households. Data on the number of

households of these different types are available for each DeSO zone. A 'couple' household contains a couple with or without children. A 'single' household consists of a 'single' individual with or without children. Any other type of household (e.g., one with multiple singles or multiple couples or a combination of singles and couples) is classified as 'other' household.

3.2.1 'Couple' households

We use a statistical method for matching individuals based on age. In particular, we consider the distribution of the age difference between the two individuals of a 'couple' household. From the national travel survey, we observe the variance (say, σ_a^2) of the age difference between two individuals in a 'couple' household. For each DeSO zone, we sort the list of 'couple' individuals by gender and then divide the list into two even groups. In cases where the number of males and females on the 'couple' individuals list is not equal, the groups contain individuals from both genders. These mixed groups result in some of the 'couple' households comprising individuals of the same gender. But with a small number of exceptions, the two individuals would belong to different genders. Given the group containing half of the 'couple' individuals in a DeSO zone, we sort the first group in ascending order of age. Afterwards we then sort the second group in ascending order of an *age proxy*, which we obtain by sampling a value from Gaussian distribution with the actual age as its mean and the aforementioned observed variance σ_a^2. That is, for an individual i having age a_i belonging to the second group, its age proxy is sampled from $\mathcal{N}(a_i, \sigma_a^2)$. The two ordered groups are then matched one-to-one. Note that we use an age proxy instead of the actual age for the second group, to ensure some disparity in the ages of the matched individuals. Also note that in order to avoid overfitting, we use only the travel survey for tuning the variance, not for precise modeling of matching with respect to age.

3.2.2 'Single' households

Typically, it is much more likely that younger individuals with 'single' status share houses with other singles, than elder individuals with 'single' status sharing houses with other singles. So, we sort the list of 'single' individuals in a DeSO zone in descending order of age and assign household status in that order based on the number of single households at DeSO level. So that elder individuals are given a higher priority of being assigned 'single' households. If the number of singles exceeds the number of 'single' households in the DeSO zone, the younger single individuals could share houses with other single individuals, and hence they would be assigned as 'other' households.

Note that owing to inconsistencies between datasets and procedural errors, the previously assigned civil statuses of certain individuals may get altered post household assignment. For instance, an individual with civil status 'couple' may end up staying alone in a 'single' household, in which case, its civil status is altered to 'single'.

3.2.3 Assigning children

We assign children to households using a two-step method. In the first step, the number of children in each family is determined. From the data regarding the total number of children in each municipality belonging to each household type, we derive the probabilities of a given type of household in each municipality having 0, 1, 2, and 3+ children. Afterwards, we assign number of children to each household by sampling from the corresponding multinomial distribution over $\{0, 1, 2, 3\}$. If the sum of the sampled numbers is less than the number of children in the municipality, some households with sampled value of 3 are randomly assigned a slightly higher value (given that the data is actually 3+ and not exactly 3 children), so that the sum of the sampled numbers equals the number of children in the municipality. If this sum is more than the number of children, we do not do any further processing.

We assign children to households (in other words, matching children with households) in the

second step. The households are sorted in ascending order of the age of the eldest constituent individual. Then, we create a list where each household is replicated by number of children assigned above. We create a second list by sorting the children in the considered municipality in ascending order of an *age proxy*, that is obtained by sampling a value from Gaussian distribution with the actual age as its mean and some variance. These two lists of households and children are matched one-to-one. Thus, all the synthetic agents, including children, are assigned households.

3.3 Assigning advanced attributes

The advanced attributes for the synthetic individuals include employment and student statuses, personal income, and car ownership.

3.3.1 Employment and student statuses

We model the employment status (ψ_W) and student status (ψ_S) of individuals, given their socio-economic attributes, using neural network classifier (NNC). ψ_W is a binary variable corresponding to being employed and ψ_S is a binary variable corresponding to being a student. The classes considered are: neither employee nor student ($\psi_W = 0, \psi_S = 0$), only employee ($\psi_W = 1, \psi_S = 0$), only student ($\psi_W = 0, \psi_S = 1$), and both employee and student ($\psi_W = 1, \psi_S = 1$). The Swedish national travel survey is used for training the classifier. In particular, the features considered are age, gender, civil status, coordinates of the municipality's center, household size (i.e., number of residents in household), and number of children ≤ 6 years old in household. The relevant data available for calibration are the number of employees and students in each DeSO zone. Let $N(z_d, \psi_W)$ and $N(z_d, \psi_S)$ respectively denote the desired number of employees and students in DeSO zone z_d. Let $\mathbb{P}_i(\psi_W = x, \psi_S = y)$ denote the probability that a synthetic agent i's employment status is x and student status is y, where $x, y \in \{0, 1\}$. We obtain the preliminary values of this probability from the output of the neural network classifier, which would correspond to the probability of the agent belonging to the class $(\psi_W = x, \psi_S = y)$. Note that we have, $\forall i$:

$$\mathbb{P}_i(\psi_S = 1) = \mathbb{P}_i(\psi_W = 0, \psi_S = 1) + \mathbb{P}_i(\psi_W = 1, \psi_S = 1) \text{ and}$$
$$\mathbb{P}_i(\psi_W = 1) = \mathbb{P}_i(\psi_W = 1, \psi_S = 0) + \mathbb{P}_i(\psi_W = 1, \psi_S = 1).$$

Similar to IPF, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

For $x \in \{0, 1\}, \forall z_d, \forall i \in z_d$:

$$\mathbb{P}_i(\psi_W = x, \psi_S = 1) \leftarrow \frac{N(z_d, \psi_S)}{\sum_{j \in z_d} \mathbb{P}_j(\psi_S = 1)} \mathbb{P}_i(\psi_W = x, \psi_S = 1) \tag{3.5}$$

For $y \in \{0, 1\}, \forall z_d, \forall i \in z_d$:

$$\mathbb{P}_i(\psi_W = 1, \psi_S = y) \leftarrow \frac{N(z_d, \psi_W)}{\sum_{j \in z_d} \mathbb{P}_j(\psi_W = 1)} \mathbb{P}_i(\psi_W = 1, \psi_S = y) \tag{3.6}$$

For $x, y \in \{0, 1\}, \forall i$:

$$\mathbb{P}_i(\psi_W = x, \psi_S = y) \leftarrow \frac{\mathbb{P}_i(\psi_W = x, \psi_S = y)}{\sum_{x', y' \in \{0,1\}} \mathbb{P}_i(\psi_W = x', \psi_S = y')} \tag{3.7}$$

Equation (3.5) scales the probabilities so that the sum of probabilities of being a student, over all agents in a given DeSO zone, is consistent with the desired number of students in that DeSO zone. Similarly, Equation (3.6) scales the probabilities so that the sum of probabilities of being an employee, over all agents in a DeSO zone, is consistent with the desired number of employees in that DeSO zone. Equation (3.7) ensures that for every agent, the probabilities of belonging to the four classes sum to 1. A class is hence assigned to every agent using multinomial sampling corresponding to the deduced probabilities. Thus, every agent is assigned its employment and student statuses. Note that this would capture heterogeneity in population since similar agents can have different employment and student statuses.

3.3.2 Personal income

We first model the personal income class (ρ^c) of agents using neural network classifier, given their socio-demographic information. The 5 classes considered in terms of Swedish krona (SEK) are: $I = \{$ 0, [1, 180K), [180K, 300K), [300K, 420K), [420K, 1M) $\}$. The partitions are based on the Swedish national income quartiles; also we consider the upper limit to be SEK 1M in our model. The Swedish national travel survey is used for training the classifier. The features considered include features used for modeling employment and student statuses as well as employment and student statuses themselves.

The relevant data showing the number of individuals for all classes in each municipality is available for calibration. Let $N(z_m, \rho^c = x)$ denote the desired number of individuals in municipality zone z_m belonging to income class x. Let $\mathbb{P}_i(\rho^c = x)$ denote the probability that a synthetic agent i's income class is x, where $x \in I$. We obtain the preliminary values of these probabilities from the neural network classifier's output. Similar to the procedure for deducing employment and student statuses, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$\forall x \in I, \forall z_m, \forall i \in z_m : \mathbb{P}_i(\rho^c = x) \leftarrow \frac{N(z_m, \rho^c = x)}{\sum_{j \in z_m} \mathbb{P}_j(\rho^c = x)} \mathbb{P}_i(\rho^c = x) \tag{3.8}$$

$$\forall x \in I, \forall i : \mathbb{P}_i(\rho^c = x) \leftarrow \frac{\mathbb{P}_i(\rho^c = x)}{\sum_{x' \in I} \mathbb{P}_i(\rho^c = x')} \tag{3.9}$$

Equation (3.8) scales the probabilities so that the sum of probabilities of belonging to an income class, over all agents in a given municipality zone, is consistent with the desired number of individuals belonging to that income class in that municipality zone. Equation (3.9) ensures that for every agent, the probabilities of belonging to the different classes sum to 1. An income class is hence assigned for every agent using multinomial sampling corresponding to the deduced probabilities.

3.3.3 Car ownership

Car ownership is the number of cars owned by each agent. In order to design our methodology for assigning car ownership, we make a practically reasonable assumption that an agent would be able to drive only if the agent owns at least one car, and an agent can own a maximum of 3 cars (which would hold true for almost all agents in practice). If an agent does not own a car, he/she cannot be a car driver, but can be a car passenger. The number of cars owned by a household would be equal to the sum of the number of cars owned by its constituent agents. Note that we assign cars to agents and not to households; this helps avoid the problem of choosing the agent(s) who would drive the car(s) in the household.

We use a neural network classifier trained on the national travel survey, with the features being

the employment and student statuses, personal income, and the features that were used for modeling employment and student statuses.

Let $\mathbb{P}_i(n)$ denote the probability that a synthetic agent i owns n cars, where $n \in \{0, 1, 2, 3\}$. So, the expected number of cars owned by an agent i is $\sum_{n'=1}^{3} n' \mathbb{P}_i(n')$. We obtain the preliminary values of these probabilities from the neural network classifier's output. We now calibrate the preliminary probabilities using the data on the total number of cars for each DeSO zone. Let $N_c(z_d)$ denote the desired number of cars in DeSO zone z_d, as per the real data. Since the expected number of cars in a DeSO zone should be equal to the sum of the expected number of cars owned by agents in that DeSO zone, we need to update the aforementioned preliminary of the probabilities so that their sum in a DeSO zone equals the desired total number of cars in that DeSO zone. Hence, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$\forall i, \forall n \in \{1, 2, 3\} : \mathbb{P}_i(n) \leftarrow \frac{N_c(z_d)}{\sum_{j \in z_d} \sum_{n'=1}^{3} n' \mathbb{P}_j(n')} \mathbb{P}_i(n), \quad \text{where } z_d \ni i \tag{3.10}$$

$$\forall i, \forall n \in \{0, 1, 2, 3\} : \mathbb{P}_i(n) \leftarrow \frac{\mathbb{P}_i(n)}{\sum_{n'=0}^{3} \mathbb{P}_j(n')} \tag{3.11}$$

Here, $z_d \ni i$ means that agent i belongs to DeSO zone z_d. Hence, each agent is assigned a certain number of cars using multinomial sampling corresponding to the deduced probabilities.

Chapter 4
Activity Generation

The activity generation has four major steps as listed below and illustrated in Fig. 4.1:

- Assign a set of activity types to each individual
- Determine the duration of each activity type for each individual
- Sequence the activities for each individual
- Create activity schedules

The first main step is the assignment of activity types namely home, work, school and other to the individuals. It includes 2 sub-steps. At first, the requisite data sets are prepared in the required format. Thereafter, the participation of individuals in activities is assigned.

The second main step includes the calculation of activity duration and sequencing. First, broad duration classes for all activity types are jointly deduced and overall travel time in a day is determined. Second, duration of activity types are calculated. In the next main step, an activity sequence is assigned to each individual by matching with an individual from the travel survey possessing similar socio-economic attributes and the same set of activity participation, based on the similarities between the duration of their activity types.

The last main step is activity scheduling. First, in order to provide a temporal organization at the extremes of the schedule, the duration, start and end time of the activity taking place at 3 am is calculated. After this step, a preliminary activity schedule is generated by distributing the total duration of each activity type among all the activities instances of this activity type.

Since the travel patterns on weekdays and weekends are significantly different, we model daily travel patterns corresponding to two types of days: an average weekday and an average weekend. Hence, while training and calibrating our model for a day of a given type (weekday or weekend), we consider individuals from the travel survey who were surveyed for the travel behavior corresponding to that type of day.

4.1 Activity types

For each agent in the synthetic population, we assign a set of activity types that the agent could be involved in. We consider four broad types of activities: staying at home, working, studying, and other activities like visiting shops, restaurants, gyms, etc. Throughout this document, we refer to these activity types as *home*, *work*, *school*, and *other*, respectively.

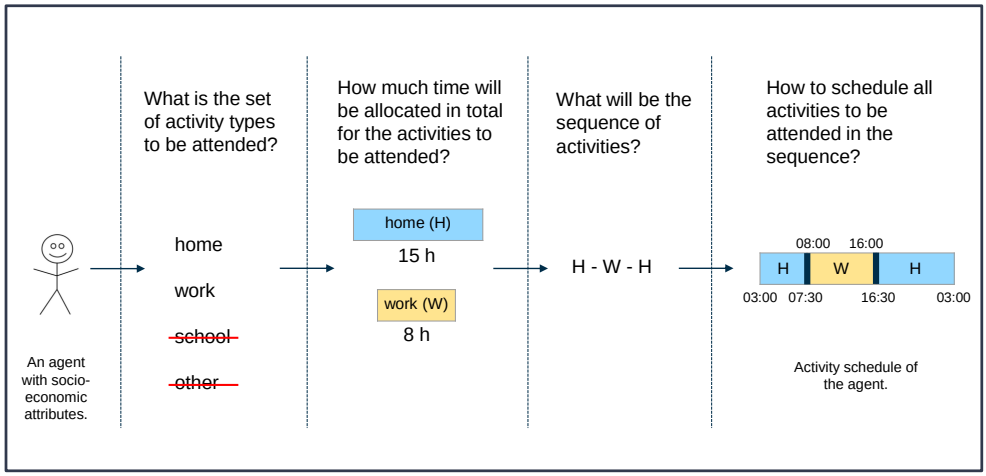


Figure 4.1: The main steps of the activity generation component. Each step in the activity generation component is represented divisions drawn by vertical dashed lines. Activity schedules are generated for agents in the synthetic population.

4.1.1 Data preparation

We first filter out individuals from the Swedish national travel survey whose activity schedules do not meet the requirement of being a daily schedule (e.g., if the sum of the activities' duration exceeds one day). We further assume that every individual visits home at least once in a day and we remove individuals not having a home activity in their daily schedule. Lastly, we filter out individuals whose first and last activities of the day are different. This is done in order to be consistent with the traffic simulation model, MATSim, that we plan to couple with later.

We present our methodology and numerical data corresponding to weekday activity schedules; note that weekend activity schedules can be modeled in the same way. Table 4.1 is a summary of additional variables used in the activity generation module.

Table 4.1: Summary of additional variables used in the activity generation module.

Symbol	Description
H	home activity
W	work activity
S	school activity
O	other activity
t_A	duration of activity type A
θ_A	willingness for activity type A
ψ_W	employment status
ψ_S	student status

4.1.2 Assignment of activity types

We begin by deducing each agent's willingness to participate in work, school, and 'other' activity types.[1] Let the variable capturing the daily duration of an activity type A be t_A, where

[1] As mentioned previously, it is assumed that each individual visits the home at least once a day and each individual is willing to join the home activity. Therefore, our model does not include a separate step to determine an individual's willingness for home activity.

$A \in \{H, W, S, O\}$; H, W, S, O correspond to home, work, school, and other activity types respectively. An individual has willingness for an activity type A ($t_A > 0$) if it is involved in that activity type on the considered day. We denote the willingness for activity type A by θ_A where $A \in \{W, S, O\}$ since H is always = 1. Using neural network classifier (NNC), we model jointly an individual's willingness to work (θ_W), study (θ_S), and 'other' activities (θ_O) given its socio-economic attributes. Modeling over joint classes preserves the correlation between the participation of the different activity types. We consider a total of $2^3 = 8$ classes, since each of $\theta_W, \theta_S, \theta_O$ could be either 0 or 1. We develop four different ML models depending on the employment status (0/1) and student status (0/1). The status considered are: neither employee nor student (0, 0), only employee (1, 0), only student (0, 1), and both employee and student (1, 1). Developing four separate models ensures that non-employees do not participate in work activities and non-students do not participate in school activities.

The Swedish national travel survey is used for training the classifiers; the features considered are age, gender, civil status, coordinates of the municipality's center, household size, number of vehicles owned, income level, and number of children ≤ 6 years old in household. $\mathbb{P}_i(\theta_W = x, \theta_S = y, \theta_O = z)$ is the probability that a synthetic agent i's willingness to work is x, willingness to study is y, and and willingness for 'other' activities is z, where $x, y, z \in \{0, 1\}$. A class is hence assigned for every synthetic agent using multinomial sampling corresponding to the deduced probabilities.

4.2 Activity duration

We determine the daily duration of different activity types using a two-step method applying neural network classifiers and sampling techniques (Fig. 4.2). In the first step, we jointly deduce broad duration classes for the different activity types; this enables us to capture the correlation between the duration of the different activity types. Broad duration classes are the classification of an individual's total activity times for different activities as low, moderate, and high. Using these broad classes and attributes of individuals, we deduce the range of overall travel time in a day or rather the range of time remaining in a day after summing the duration of all activity types. In the second step, using the deduced broad classes of duration of all the activity types and the range of daily travel time, we derive duration of all the activity types. The method proposed here replicates people's heterogeneity in the population by allowing agents with similar attributes to have different activity duration.

4.2.1 Determining the broad classes of duration of activity

The broad classes for duration we consider, are low, moderate, and high.[2] Evidently, the definitions of low, moderate, and high would depend on the activity type. The broad duration classes we consider for the different activity types are as follows (in hours):

- Home: (0,12], (12,18], (18,24]

[2]The purpose of having broad classes for duration is to capture the correlation among the duration of 4 activity types. The sum of the hourly classes is at most 24. A possible distribution of at most 24 hours among the 4 activity types could be represented by a tuple of 4 positive integers. The number of possible tuples is $\binom{24}{4} = 10{,}626$. Clearly, this is an exceedingly high number of classes for travel surveys, which typically consist of a few tens of thousands of individuals. Even accounting for the possibility that many of these joint classes would be vacuous owing to them not corresponding to any individual in the survey, most of the non-vacuous classes would contain just a few tens of individuals. Such classification is clearly not suitable for training a neural network classifier. So, it is important that the number of joint classes is reasonably low, which is why we consider broad classes.

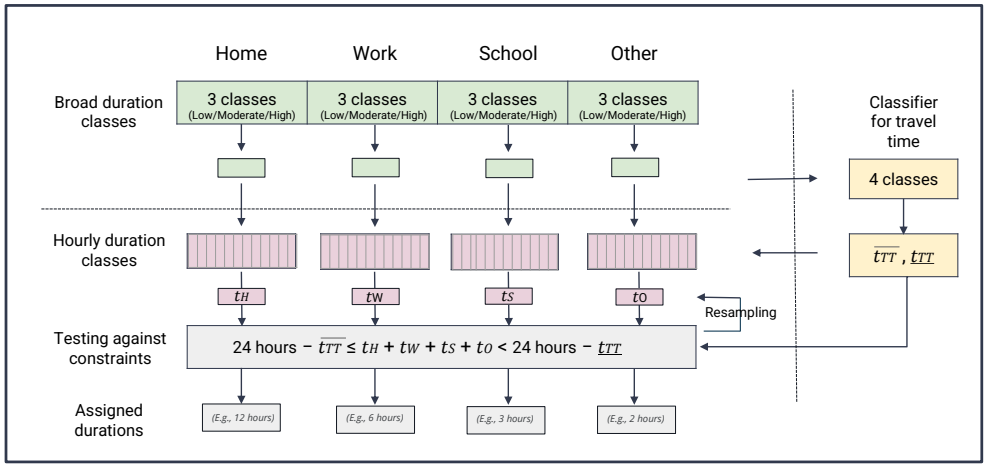


Figure 4.2: The flow chart of activity duration assignment methodology in SySMo. Green rectangles: joint model for broad activity duration, yellow rectangles: model for travel time, pink rectangles: model for hourly activity duration, and gray rectangles: final activity duration satisfying the constraint.

- Work: (0,6], (6,10], (10,24]
- School: (0,6], (6,8], (8,24]
- Other: (0,2], (2,5], (5,24]

Since we have 3 broad classes for each of the 4 activity types, the total number of joint classes is $3^4 = 81$. In order to increase the robustness of the classifiers, we consider different classifiers for different sets of activity types (here, a set for an individual would contain an activity type if the individual has a willingness for that activity type). Since all individuals are assumed to be involved in home activity, a set of activity types is of the form $\{H\} \cup S$, where $S \in 2^{\{W,S,O\}} \setminus \{\}$. Note that we exclude the null set from S since agents with only home and no other activity type, will be assigned a duration of 24 hours for home activity type. That is, we consider 7 different classifiers. Thus, a classifier trained using survey individuals with a given set of activity types, is used to deduce the joint class for an agent in the synthetic population with that particular set of activity types. Similar to the previously described classifiers, the national travel survey is used for training and the features considered are socio-economic attributes and employment/studenthood statuses. The classifier produces the probabilities of an agent belonging to the joint classes; the broad classes of duration of activity types are hence assigned using multinomial sampling.

4.2.2 Determining the range of daily travel time

In order to deduce more specific duration of the different activity types for an agent, we estimate the daily total travel time for that agent. The sum of the specific acitivity duration is then set equal to 24 hours minus the day's total travel time. Note that agents with only home activity type are assigned zero daily travel time. We consider 4 classes for estimating daily travel times, namely (in hours): (0,0.5], (0.5,1], (1,2], 2+. These classes are approximately based on the four quartiles for non-zero daily travel time in the travel survey.

A neural network classifier is trained using the travel survey, the features being the socio-

economic attributes, the employment and studenthood statuses, the set of activity types, and the broad classes of duration deduced above. The classifier outputs the probability distribution over the 4 classes for each agent; a class is hence assigned using multinomial sampling. Note that since the '2+ hours' class is unbounded and since the number of surveyed individuals with more than 6 hours of the day's total travel time is negligible, we interpret this class as (2,6] hours when assigning to agents in the synthetic population. Thus, we obtain the class, and hence, the range of daily travel time for each agent. If the class assigned to an agent is $\left(\underline{t_{TT}}, \overline{t_{TT}}\right]$, the lower limit of the range of its daily travel time is $\underline{t_{TT}}$ and the upper limit is $\overline{t_{TT}}$.

4.2.3 Determining duration of activity types

Now that we have deduced the broad classes of duration of all activity types and the range of daily travel time for each agent in the synthetic population, we determine the duration of the different types of activities with a higher resolution. The sum of the duration of the activity types should be between 24 hours minus the range of the day's total travel time $\left(\underline{t_{TT}}, \overline{t_{TT}}\right]$. That is,

$$24 \text{ hours} - \overline{t_{TT}} \;\leq\; t_H + t_W + t_S + t_O \;<\; 24 \text{ hours} - \underline{t_{TT}} \tag{4.1}$$

We achieve this in two steps. First, we deduce the preliminary probability distribution over hourly duration of each activity type, by considering 24 hourly classes per activity type. Then, we sample the duration of all types of activities such that they collectively satisfy Constraint (4.1).

We now explain how we deduce the preliminary probability distributions over the 24 hourly duration classes for the different activity types. An hourly duration class is of the form $[T, T+1)$ hours, where $T \in \{0, 1, \ldots, 23\}$. We model the hourly duration of an agent's given activity type using neural network classifier, given its socio-economic attributes, employment and studenthood statuses, willingness for the activity types, broad classes of duration of the activity type, and the class corresponding to daily travel time. When modeling the hourly duration of an activity type, we consider 3 different classifiers for the 3 different broad duration classes of each activity type. Each classifier is trained using survey individuals with a particular broad duration class. We thus obtain the preliminary probability distribution over the 24 hourly duration classes for the 4 activity types, for each agent in the synthetic population.

Next, we explain how we obtain the duration of all activity types such that their sum satisfies Constraint (4.1). There are fundamentally two ways to achieve this, namely, the mathematical way[3] and the simulation-based way. In our implementation, we employ a simulation-based approach. For an agent, we sample the hourly duration of the 4 activity types from the aforementioned preliminary probability distributions. Then, numbers that are sampled uniformly at random in [0,1) are added to each of the sampled hourly activity duration to introduce idiosyncratic variances and generate a final duration. If Constraint (4.1) is satisfied for an agent, the four activity types are assigned the sampled duration. On the other hand, if the constraint is not satisfied, we repeat the sampling for the hourly duration and the idiosyncratic variances procedure. We run the redrawing of samples for a fixed large number of iterations (30 iterations) so that Constraint (4.1) is satisfied for a large fraction (99%) of agents, and hence a large fraction of agents are assigned duration of the four activity types. However, in order to ensure

[3]In the mathematical approach, one would need to create a truncated joint distribution of the hourly duration of the four activity types, which can be obtained by combining the distributions of the activity types' duration and truncating to satisfy Constraint (4.1). The hourly duration can then be sampled from this truncated joint distribution, followed by adding a few minutes to the hourly duration so as to introduce a natural idiosyncratic variance, while ensuring that Constraint (4.1) is not violated.

that no agent violates the constraint, in principle, it could take infinite iterations of redrawing of samples. We hence employ a simple heuristic procedure that trims or adds sampled times for achieving this and thus assign the activity duration satisfying the constraint to the remaining agents.

4.3 Activity sequencing

We now generate the sequence of activities for each agent in the synthetic population. While there are several ways to generate an activity sequence by matching individuals with distinct sequences, most approaches employed in the literature can be broadly classified into: (a) directly based on socio-economic attributes, e.g., [8] and (b) based on proxy parameters, e.g., [9] where the proxy parameters are daily activity duration. We employ the approach of having daily activity duration as proxy parameters.

The approach is based on the assumption that individuals with similar socio-economic attributes and activity type duration, would have similar activity sequences. This means that an synthetic agent in our model would be assigned the activity sequence of the individual in the travel survey that is most similar to them. In this approach, similarity between two individuals is measured using Euclidean distance between their attributes and duration. Note that while similarity between two sets of activity duration (t_H, t_W, t_S, t_O) could be quantified since duration have the same unit (namely, time unit), it is not clear how similarity between two sets of socio-economic attributes (e.g., age, gender, etc.) could be quantified since these attributes do not have the same unit and are not directly comparable. In our model, however, an individual's activity duration are themselves deduced from its socio-economic attributes, and so, the activity duration act as a proxy for the socio-economic attributes. We hence measure the similarity between two individuals based on the Euclidean distance in the 4-dimensional space, between their activity duration' tuples, namely, (t_H, t_W, t_S, t_O).

We employ a two-step method to assign the daily activity patterns to the agents. We first determine candidate individuals in the travel survey and then find the most similar individual among the candidates using activity duration. Since in our approach, the duration of the four activity types act as a proxy, and are in a sense, encoding of the socio-economic attributes, some information is lost during this encoding. It is hence important to specifically ensure that the two individuals being compared are not very different with regard to their socio-economic attributes and have the same set of willingness for the activity types. So, for a given agent in the synthetic population, we consider a set of candidate individuals from the travel survey who have the same set of willingness for the activity types and have as many similar socio-economic attributes as possible.

For having as many similar socio-economic attributes as possible, we gradually filter candidate individuals based on their socio-economic characteristics, while ensuring that the filtered set remains above 50. If after filtering according to an attribute, the size of the candidate individuals' set falls below the considered threshold, we revert back to the set that was before filtering, and the obtained set is considered the final set of candidate individuals. Following the creation of the set of candidate individuals from the survey, for a synthetic agent, we choose the individual who is the most similar to the considered agent with regard to the Euclidean distance between their activity duration' tuples (t_H, t_W, t_S, t_O). We then assign to the synthetic agent, the activity sequence of the chosen individual from the survey. It should be remembered that the assigned activity sequences also capture the heterogeneity in the population, as the process of assigning activity duration capture the heterogeneity in the population, and activity duration are used as a proxy parameters. To avoid overly complicated and repetitive activity sequences,

we simplify adjacent activity instances in the assigned sequence. We first deduplicate home, work, school activity types , that is, if two adjacent activity instances in the sequence are of the same type, we merge them into one instance of that activity type. For instance, $-W$-$W-$ would be converted to $-W-$. For activity type *other*, we consider up to 3 consecutive activities, unlike the deduplication method followed for home, work, and school activities.[4]

4.4 Activity scheduling

With the duration of the different activity types and the activity sequence at hand, we are now ready to generate the activity schedule for each agent in the synthetic population. We first deduce the start and end times of the activity that takes place at 3 AM. Thereafter, we distribute the total duration of an activity type among its individual instances in the activity sequence, so as to provide the temporal order of all instances, hence generate an activity schedule. Note that, we assume the day to start and end at 3 AM, since a minimum number of individuals are travelling and thus a maximum number of individuals are at an activity at this time according to the travel survey.

Modeling the start and end times of the 3 AM activity instance accurately is important for a number of reasons. Firstly, it facilitates the arrangement of remaining activities during a day using activity sequences and duration, as the head and tail of the sequence is defined. Secondly, for most individuals, the start time of the 3 AM activity instance would be in the evening and the end time would be in the morning; so they would help in capturing the morning and evening peak in traffic patterns.

4.4.1 Concretizing the 3 AM activity

The 3 AM activity type for an agent is directly obtainable from its deduced activity sequence, as the first/last activity type. Let $a_{3\text{AM}}$ denote the 3 AM activity instance and $t_{a_{3\text{AM}}}$ be its duration. Let $T^s_{a_{3\text{AM}}}$ and $T^e_{a_{3\text{AM}}}$ denote the start and end times of the 3 AM activity instance. In order to deduce $T^s_{a_{3\text{AM}}}$ and $T^e_{a_{3\text{AM}}}$, we first deduce their hourly distributions, using neural network classifiers (with 24 classes each) trained using the travel survey. On similar lines as the determining activity duration procedure, we develop different models by activity type using the travel survey.

For the sampling process, we impose a certain constraint with regard to the amount of time spent for the 3 AM activity instance. It is clear that the amount of time spent for the 3 AM activity should not exceed the total duration of the activity type corresponding to the 3 AM activity. We impose a lower bound such that the mean of the upper and lower bounds equals the deduced time to be spent for the 3 AM activity instance. Let $D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}})$ denote the amount of time spent for the 3 AM activity instance to be sampled. Since we have already deduced the total duration of the activity type $A_{3\text{AM}}$, the fraction of the total duration of the 3 AM activity type that is allotted to the 3 AM activity instance can be denoted $f_{3\text{AM}} = \frac{t_{a_{3\text{AM}}}}{t_{A_{3\text{AM}}}}$. We deduce $f_{3\text{AM}}$ by way of regression using neural network trained using the travel survey. To have a lower bound such that the mean of the upper and lower bounds equals the deduced spent time for the 3 AM activity instance, we formulate the lower bound as $\big(1 - 2(1 - \hat{f}_{3\text{AM}})\big)$. We hence obtain the following constraint:

[4]It is to be noted that simplification of adjacent activity instances is not a requirement of our methodology, but rather a choice we make for our model. In essence, our model considers that if two adjacent activity instances are of the same type, they are either at the same location (e.g., going for a walk or a ride and returning to the same place) or the locations are close to each other. This would help our model be simple enough to analyze, while being detailed enough for modeling mobility.

$$\big(1 - 2(1 - \hat{f}_{3\text{AM}})\big)t_{A_{3\text{AM}}} \;<\; D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}}) \;<\; t_{A_{3\text{AM}}} \tag{4.2}$$

We sample the start and end times of the 3 AM activity instance from their corresponding hourly distributions that we deduced earlier, and add natural idiosyncratic variances to them to obtain times that satisfy Constraint (4.2). We employ a similar approach as the one for sampling activity duration while satisfying Constraint (4.1). For the small fraction of agents whose start and end times of the 3 AM activity instance do not satisfy Constraint (4.2), we employ a simple heuristic procedure to meet the constraint.

Note also that for the particular case of agents for whom the 3 AM activity type occurs only at the start and end of the activity sequence (i.e., there is no instance of $A_{3\text{AM}}$ apart from $a_{3\text{AM}}$ itself), we need to ensure that $D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}})$ equals $t_{A_{3\text{AM}}}$.

4.4.2 Deducing start and end times of activity instances

Now that we have deduced the start and end times of the 3 AM activity instance, the head and tail of the activity sequence are concretized. We proceed to present our approach for distributing the duration of an activity type among its individual instances in the activity sequence, with the help of a running example of an agent whose activity sequence is *H-W-H-W-O-H*. Fig. 4.3 present an illustration of the example. Since the activity type at the two extremes (head and tail) of the sequence is H, the 3 AM activity type is 'home'. We have deduced the start and end times of the 3 AM activity instance and so, we know at what times the first home activity instance ends and the last home activity instance starts.

We now distribute the total duration of each activity type among its different instances in the sequence. For an activity type that is not the 3 AM activity type (for this example, an activity type other than home), we distribute its total duration equally among its instances in the sequence. In the considered example, such activity types are W (work) and O (other). Since we have 2 instances of work and 1 instance of other activity type, the amount of time spent for each of the work activity instance is $\frac{t_W}{2}$ and that for the sole other activity instance is $\frac{t_O}{1}$. For the activity type corresponding to the 3 AM activity instance (home, in this example), the amount from the total activity duration that remains after allotting to the 3 AM activity instance (i.e., $t_{A_{3\text{AM}}} - t_{a_{3\text{AM}}}$, where $t_{a_{3\text{AM}}} = D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}})$), is distributed equally among its instances barring the 3 AM instance. Since we have 1 instance of the home activity type in the sequence apart from the 3 AM one, the amount of time spent for this home activity instance is $\frac{t_H - t_{h_{3\text{AM}}}}{1}$, where $t_{h_{3\text{AM}}}$ is the time allotted to the 3 AM home activity instance.

Our next step is to assign the travel times between adjacent activity instances. Firstly, the daily travel time could be calculated by subtracting the sum of the total duration of the different activity types from 24 hours (i.e., $t_{TT} = 24 - (t_H + t_W + t_S + t_O)$). Note that we are now deducing the daily travel time, while earlier, we had deduced its range in order to feed into Constraint (4.1). We then distribute this total daily travel time equally across the different trips in the activity sequence. In the activity sequence of our running example, since we have a total of 5 trips, the amount of time spent for each of the trips is $\frac{t_{TT}}{5}$. It is worth noting that these are preliminary travel times, and will later be refined based on the assigned activity locations [10] and using an agent-based transport simulation software such as MATSim.

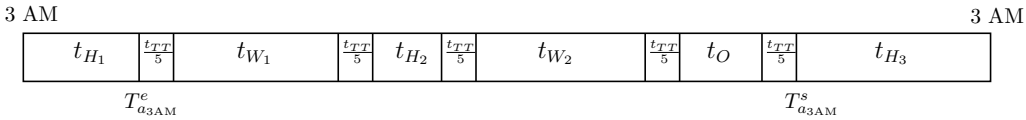


Figure 4.3: Activity schedule of an agent with activity sequence is *H*-*W*-*H*-*W*-*O*-*H*. The daily travel $t_{TT} = 24 - (t_H + t_W + t_S + t_O)$ **.**

Now that we have a temporal arrangement of all activity instances within a day for every agent (that is, the activity sequence along with the start and end times of each activity instance), the daily activity schedules of all the agents in the synthetic population are ready.

Chapter 5

Location and Mode Assignment

This chapter describes the methodology for the mode and location assignments for agents' activities (Fig. 1.1 third box from the top). We first start with home location assignment where we assign building types and residential locations to the households. This is then followed by mode and location assignments to all the non-home activities (broadly classified as work, school, or other activity types). Fig. 5.1 shows a flow chart of activity, mode and location assignment methodology.

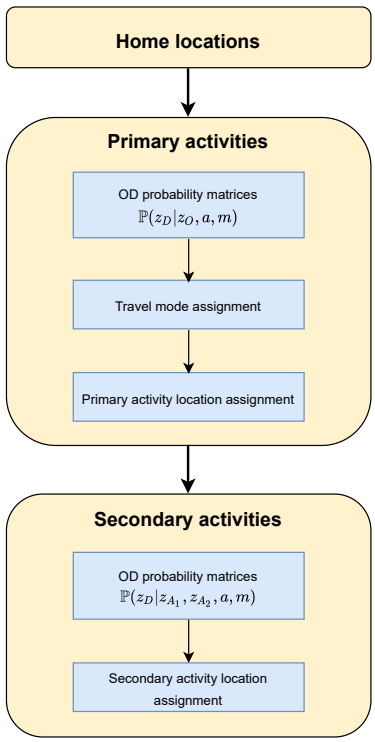


Figure 5.1: A flow chart of activity, mode and location assignment Yellow rectangles: major steps of the activity location assignment methodology; blue rectangles: sub-steps within the main steps.

5.1 Home locations

Up till now individuals and households have been synthesized in DeSO zones (See Chapter 3). In order to maintain the accuracy of the population distribution in the location assignment, we create smaller zones called "virtual zones" from the overlap of the two zone systems DeSO and sq.km. zones.

Let z_v denote a virtual zone being the intersection between a DeSO zone and a sq.km. zone. A building lies in virtual zone z_v if and only if its geometrical center lies in sq.km. zone $z_s \ni z_v$ as well as in DeSO zone $z_d \ni z_v$. Here, $z_s \ni z_v$ and $z_d \ni z_v$ denote that sq.km. zone z_s and DeSO zone z_d contain virtual zone z_v. Let $N(z_d)$ and $N(z_s)$ denote, respectively, the desired populations of DeSO zone z_d and sq.km zone z_s. Let $n(z_v)$ denote the deduced number of agents in virtual zone z_v. We iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$
\begin{aligned}
&\forall z_v : \\
&\quad n(z_v) \leftarrow \frac{N(z_s)}{\sum_{z_{v'} \in z_s} n(z_{v'})} n(z_v), \quad \text{where } z_s \ni z_v \\
&\quad n(z_v) \leftarrow \frac{N(z_d)}{\sum_{z_{v'} \in z_d} n(z_{v'})} n(z_v), \quad \text{where } z_d \ni z_v
\end{aligned}
$$

Next, in a DeSO zone, we assign to each household a virtual zone by way of multinomial sampling where the probability of being assigned a virtual zone is proportional to the aforementioned deduced number of agents in that virtual zone. Let a DeSO zone consists of virtual zones $z_{v_1}, \ldots, z_{v_m}$, and $\mathbb{P}_h(z_{v_p})$ denote the probability of a household h in the DeSO zone being assigned virtual zone z_{v_p}.

$$
\forall h \in z_d : \quad \mathbb{P}_h(z_{v_p}) \leftarrow \frac{n(z_{v_p})}{\sum_{p'=1}^{m} n(z_{v'_p})}, \quad \text{where } z_d \ni z_{v_p} \tag{5.1}
$$

With this procedure, the expected number of agents in a virtual zone will be consistent with the aforementioned deduced number of agents, despite the DeSO zone having households of various household sizes. This can be shown as follows. Let there be q number of households in DeSO zone z_d with household sizes $n(h_1), \ldots, n(h_q)$. Since Eq.5.1 gives the probability of a household being assigned virtual zone z_{v_p}, the expected number of agents in virtual zone z_{v_p} is $\sum_{j=1}^{q} n(h_j)\mathbb{P}_h(z_{v_p})$. In addition,

$$
\sum_{j=1}^{q} n(h_j) = \sum_{p'=1}^{m} n(z_{v'_p}) = n(z_d) \tag{5.2}
$$

Eq.5.2 states that the sum of the sizes (number of individuals) of all households in DeSO zone z_d should be equal to the sum of the number of agents in all virtual zones constituting DeSO zone z_d, which is the number of agents in the DeSO zone z_d. Thus, the expected number of individuals in virtual zone z_{v_p} is $\sum_{j=1}^{q} n(h_j)\mathbb{P}_h(z_{v_p}) = n(z_d)\frac{n(z_{v_p})}{n(z_d)} = n(z_{v_p})$, which is as desired.

We then proceed to assign a specific residential building to households. The residential buildings are broadly classified into detached houses, apartment buildings, and buildings of other or unknown types. A detached house can accommodate one household, while an apartment building can accommodate multiple households. If there is no apartment building in a virtual zone, we treat a building of other or unknown type as an apartment building (i.e., it can accommodate multiple households).[5] The average household size for detached houses (2.7) and apartment buildings (1.9) differ greatly [11]. The correlation between households and types of residence is established via household size (number of individuals constituting a household) and buildings are assigned to households in each virtual zone.

5.2 Overview of activity mode and location assignment

This subsection provides an overview of the methodology for work, school, and other activities travel mode and location assignment. The mode and location assignment begins with OD probability matrices for assigning modes and locations of primary and secondary activities. The OD matrix estimation procedures vary by an unique SySMo's zone system that combines the zone systems of three models constituting the Swedish transportation model Sampers: the Väst and Samm regional models and the national model (Section 2.3). These models provide short and long distance OD matrices by mode of transport and trip purpose (work, bussiness, other, and private). In the regional models, the zones inside the corresponding region are small, while being large outside of the region. In the national model, all the zones are moderately sized. Fig. 5.2 shows the SySMo zone system where the zones are small inside the Väst and Samm regions, and moderate outside these regions.

Activities are categorized into primary and secondary activities. Primary activities are critical activities whose locations are determined independently of the locations of other activities except 'home' [12, 13, 14]. Such activities comprise of work and school. In activity sequences in which an agent does not participate any primary activity from a home activity to the next home activity, 'other' activities are also categorized as a primary activity. Secondary activities are activities between primary activities.[6] Their location depend on the location of the primary activities that are adjacent to them in the activity sequence. For instance, if an agent visits a shopping center while traveling from work towards home, it is categorized as a secondary 'other' activity. The modes we consider in our model are car as driver (car), car passenger (carP), public transit (PT, which includes buses, trams, and trains), bike, and walk.

For each origin zone and activity type, we deduce the distribution of the modes and destination zones using one of the following: (a) Sampers OD matrices, (b) IPF, or (c) gravity model. The methodology consists of different procedures according to origin and destination zones, and the distance of trips. It is summarised in a schematic form in Table 5.1.

For example, the table entry corresponding to origin z_{V_1} (a zone in the Väst region) and destination z_{O_1} (a zone belonging to neither Väst nor Samm region) are long distance trips from Väst to Other regions in Table 5.1(i.e., $\mathcal{L}$). The flow for this particular OD pair (shown in Table 5.2 with the activity type as 'work', 'Starting/ending in Väst/Samm region' is '✓' and 'Distance class' is 'Long' is obtained by way of IPF using both national and regional models.

The procedures for using Table 5.2 to calculate mode and location assignments are briefly explained here and will be explained in more details in the sections below.

[5] This is useful if in a virtual zone the number of households exceeds the number of detached houses and there is no apartment building to accommodate the remaining households. While this might be rare, it is important for the model's completeness sake.

[6] 'Other' activities that cannot be categorized as primary activities could be viewed as secondary activities.

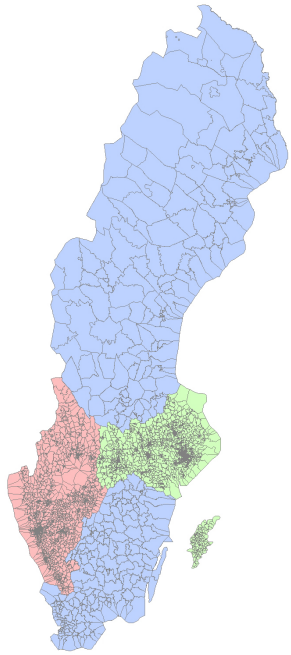


Figure 5.2: The zone system used in SySMo. Pink: zones according to Väst regional model, green: zones according to Samm regional model, and blue: zones according to the national model.

Table 5.1: A schema of short vs. long distance trip definition by SySMo's zone system for work/other trips. The colors correspond to different estimation methods described in Table 5.2.

$\mathcal{S}$: Short distance trip, $\mathcal{L}$: Long distance trip; For $y \in \{$ Väst, Samm, and Other $\}$, zones z_{y_1} and z_{y_2} are close to each other, z_{y_2} and z_{y_3} are close to each other, z_{y_1} and z_{y_3} are far from each other.

		Väst			Samm			Other		
		z_{V_1}	z_{V_2}	z_{V_3}	z_{S_1}	z_{S_2}	z_{S_3}	z_{O_1}	z_{O_2}	z_{O_3}
Väst	z_{V_1}	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{L}$						
	z_{V_2}	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$		$\mathcal{L}$			$\mathcal{L}$	
	z_{V_3}	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{S}$						
Samm	z_{S_1}				$\mathcal{S}$	$\mathcal{S}$	$\mathcal{L}$			
	z_{S_2}		$\mathcal{L}$		$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$		$\mathcal{L}$	
	z_{S_3}				$\mathcal{L}$	$\mathcal{S}$	$\mathcal{S}$			
Other (O)	z_{O_1}							$\mathcal{S}$	$\mathcal{S}$	$\mathcal{L}$
	z_{O_2}		$\mathcal{L}$			$\mathcal{L}$		$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
	z_{O_3}							$\mathcal{L}$	$\mathcal{S}$	$\mathcal{S}$

- For the cases where we use the OD matrices directly from the national and Väst and Samm regional models, the mode distribution as well as the distribution of zones for activity location (i.e., destination zones) are taken directly from the models. Sampling from these distributions, we assign mode and destination zone (given the mode used and travel time) of the activity to each agent.
- For the cases for which we employ IPF, we use the combination of OD matrices from the regional and the national models for long distance trips at finer zone levels. Once we obtain the IPF's output, the distributions over modes and zones for activity locations are calculated based on the similar procedures previously mentioned.
- For the cases corresponding to primary activity types for which we use gravity model, the methodology comprises the following steps: mode-based gravity model, mode distribution, potential mode usage, mode assignment, and destination assignment.
- The cases corresponding to secondary activity types is modelled with a different methodological treatment, using a gravity model.

5.3 Primary activities

In this step, we assign the activity location for each primary activity and the travel mode. We first compute the origin-destination (OD) probability matrices for each activity type and mode. We then determined the mode of transportation between activities. This is followed by the activity location assignment at the building level performed by using the agent's home location, primary activity type, travel mode, and travel time.

5.3.1 OD probability matrices

The objective of forming OD probability matrices is to deduce the probability of an activity location being in a zone z_D, given the origin (home) zone z, activity type a, and mode m. We have 15 different types of OD probability matrices by each primary activity type and mode. A matrix corresponding to activity type a and mode m can be visualized as containing elements $\mathbb{P}(z_D|z,a,m)$ in row z and column z_D, where z is the origin zone and z_D is a candidate destination zone. By definition, $\sum_{z_D}\mathbb{P}(z_D|z,a,m)=1$ and so, it is a probability (or stochastic) matrix. The methodology employed to form OD probability matrices consists of different procedures

Table 5.2: Summary of sampling methods for estimating the flows in the OD matrices by activity type, starting/ending regions and distance class. The definition of distance class by starting/ending region for work/trip trips are defined in Table 5.1.

Activity type		Starting/ending in Väst/Samm region?	Distance class	Multinomial sampling
Primary	Work / Other	✓	Short	Väst and Samm regional models
			Long	IPF based on the national and Väst and Samm regional models
		✗	Short	Gravity model based on Väst and Samm regional models
			Long	National model
	School	✓/ ✗	Short	Gravity model based on Väst and Samm regional models
Secondary	Other	✓/ ✗	Short / Long	Gravity-like model using the travel survey $\mathbb{P}(k\|j,i) \propto s_k e^{\beta t_{jk}+\gamma t_{ki}}$

according to origin and destination points, and the trip distance. These procedures can be seen in the following.

Work/other trip | short distance | Väst or Samm region.
As presented in Table 5.2, we obtain the probabilities for short distance trips that start or end in Väst or Samm region, using the OD matrices corresponding to the Sampers regional models. Specifically, if the regional model matrix corresponding to activity type a and mode m is $M_r^{a,m}$, and its entry corresponding to origin zone z_{r_o} and destination zone z_{r_d} is $M_r^{a,m}(z_{r_o}, z_{r_d})$, the probability is obtained as

$$\mathbb{P}(z_{r_d}|z_{r_o}, a, m) = \frac{M_r^{a,m}(z_{r_o}, z_{r_d})}{\sum_{z_{r_{d'}}} M_r^{a,m}(z_{r_o}, z_{r_{d'}})} \tag{5.3}$$

which forms the entry for origin zone z_{r_o} and destination zone z_{r_d} in the OD probability matrix corresponding to activity type a and mode m.

Work/other trip | long distance | neither Väst or Samm region.
Concerning long distance trips that neither start nor end in Väst or Samm region, we obtain the probabilities using the OD matrices corresponding to the Sampers national model. If the national model matrix corresponding to activity type a and mode m is $M_n^{a,m}$, and its entry corresponding to origin zone z_{n_o} and destination zone z_{n_d} is $M_n^{a,m}(z_{n_o}, z_{n_d})$, the probability is obtained as

$$\mathbb{P}(z_{n_d}|z_{n_o}, a, m) = \frac{M_n^{a,m}(z_{n_o}, z_{n_d})}{\sum_{z_{n_{d'}}} M_n^{a,m}(z_{n_o}, z_{n_{d'}})} \tag{5.4}$$

which forms the entry for origin zone z_{n_o} and destination zone z_{n_d} in the OD probability matrix corresponding to activity type a and mode m.

Work/other trip | long distance | Väst or Samm region.
For the long distance trips that start or end in Väst or Samm region, we use iterative proportional fitting (IPF) using Sampers OD matrices from both regional and national models. The purpose of performing IPF is to combine the long distance trips given in a small-sized zone within the region in the regional models and in a moderate-sized zones in the national model.

From the regional models corresponding to activity type a and mode m, we know $M_r^{a,m}(z_{r_a}, z_{r_b})$ where either zone z_{r_a} (a small-sized zone) or zone z_{r_b} (a large-sized zone) belongs to Väst or Samm region. Also, from the national model, we know $M_n^{a,m}(z_{n_p}, z_{n_q})$ where either zone z_{n_p} or zone z_{n_q} (moderate-sized zones) belongs to Väst or Samm region. The Väst (or Samm) model's zones staying within the region Väst (or Samm) are partitions in the national model. Also, the national model's zones outside the Väst and Samm regions are partitions in the Väst and Samm zones. Hence, let $z_{r_a} \in z_{n_p}$ and $z_{n_q} \in z_{r_b}$. We need to deduce a new matrix whose elements are $M_s^{a,m}(z_{r_a}, z_{n_q})$, since z_{r_a} and z_{n_q} are the smaller sized zones in their respective regions. Fig. 5.3 presents an illustration of the aforementioned idea. The IPF procedure is initialized as follows:

$$\forall z_{r_a}, z_{n_q}: \quad M_s^{a,m}(z_{r_a}, z_{n_q}) \leftarrow \frac{M_r^{a,m}(z_{r_a}, z_{r_b})}{|z_{r_b}|}, \quad \text{where } z_{r_b} \ni z_{n_q}. \tag{5.5}$$

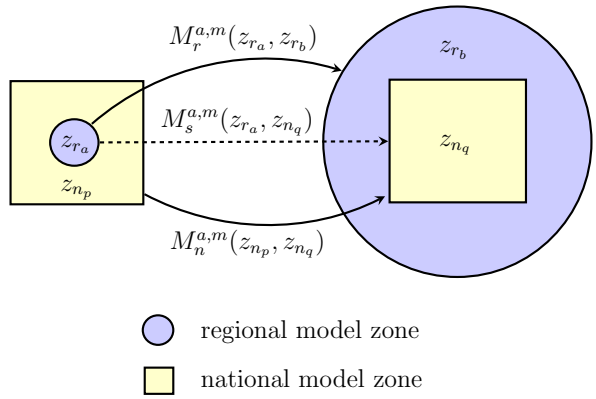


Figure 5.3: An abstract illustration of regional and national model zones, and OD matrices' values to be used for IPF (arrows point from origin to destination; solid arrow means that the value is available from Sampers OD matrices; dotted arrow means that the value is to be deduced)

Here, $z_{r_b} \ni z_{n_q}$ denotes that regional model zone z_{r_b} contains national model zone z_{n_q}, and $|z_{r_b}|$ is the number of national model zones constituting the regional model zone z_{r_b}. In order to deduce $M_s^{a,m}(z_{r_a}, z_{n_q}), \forall z_{r_a}, z_{n_q}$, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$\begin{aligned}
&\forall z_{r_a}, z_{n_q} : \\
&M_s^{a,m}(z_{r_a}, z_{n_q}) \leftarrow \frac{M_n^{a,m}(z_{n_p}, z_{n_q})}{\sum_{z_{r_{a'}} \in z_{n_p}} M_s^{a,m}(z_{r_{a'}}, z_{n_q})} M_s^{a,m}(z_{r_a}, z_{n_q}) \\
&M_s^{a,m}(z_{r_a}, z_{n_q}) \leftarrow \frac{M_r^{a,m}(z_{r_a}, z_{r_b})}{\sum_{z_{n_{q'}} \in z_{r_b}} M_s^{a,m}(z_{r_a}, z_{n_{q'}})} M_s^{a,m}(z_{r_a}, z_{n_q})
\end{aligned} \tag{5.6}$$

Note that just as we deduce $M_s^{a,m}(z_{r_a}, z_{n_q})$ using $M_r^{a,m}(z_{r_a}, z_{r_b})$ and $M_n^{a,m}(z_{n_p}, z_{n_q})$, we can deduce $M_s^{a,m}(z_{n_q}, z_{r_a})$ using $M_r^{a,m}(z_{r_b}, z_{r_a})$ and $M_n^{a,m}(z_{n_q}, z_{n_p})$. After deducing these new matrices via IPF, we obtain (OD) probability matrices employing a similar method used to create in the previous procedures.

Work/other trip | short distance | neither Väst or Samm region.
For short distance work and primary 'other' trips that neither start nor end in Väst or Samm region, we use gravity models. We have different gravity models (mode-specific gravity model [15]) in which the parameters corresponding to each activity type and mode are calibrated independently using OD matrices from two regional models.

School trips | short distance
In the Sampers model school trips are modelled as other trips and not separately. Parameters corresponding to school activity are also calibrated using the 'other' trips while developing the gravity model for SySMo, since school trips are integrated into 'other' trip in Sampers' OD matrices. We thus apply a mode-specific gravity model to all regions for this trip type. In our model school trips can only be short-distance, unlike other activity types.

Table 5.3: Gravity model parameters for primary activity types

Activity type	Car	CarP	PT	Bike	Walk
Work	−0.14	−0.14	−0.08	−0.59	−1.67
School, Primary Other	−0.21	−0.21	−0.12	−0.93	−2.10

Gravity model We apply an exponential decay function in the model presented in Equation (5.8). It has been observed that the gravity models with exponential decay in the distance capture short distance trip distributions very well [16]. Let $\mathbb{P}(z_d|z_o, a, m)$ denote the probability that an agent's activity location is z_d, given that its home location is z_o, activity type is a, and mode used is m. In what follows, let the parameter corresponding to activity type a and mode m be denoted by β_m^a. Let $d(z_o, z_d)$ denote the spherical distance between zones z_c and z_b; we define $d(z_o, z_o)$ to be the radius of zone z_o. Let $s_{z_d}^a$ denote the *attraction potential* of zone z_d for activity type a. $s_{z_d}^a$ could be simply assumed to be equal to the population of zone z_d. With all the variables defined, the gravity model in its probabilistic form can be expressed as:

$$\mathbb{P}(z_d|z_o, a, m) \propto s_{z_d}^a e^{\beta_m^a d(z_o, z_d)} \tag{5.7}$$

$$= \frac{s_{z_d}^a e^{\beta_m^a d(z_o, z_d)}}{\sum_z s_z^a e^{\beta_m^a d(z_o, z_d)}} \tag{5.8}$$

Table 5.3 presents the calibrated values of the parameters. A more negative value of parameter β_m^a means that the probability drops rapidly with an increase in distance. We see that across all modes, the values of parameter β_m^a for school and primary 'other' activity types are more negative than for work activity type. Also, across all the presented activity types, the value of β_m^a for walk is more negative than that for bike, followed by that for car and carP, while the value for PT is the least negative.

Once the preliminary probabilities are obtained using the above procedure, they could be fine-tuned with the help of additional data. For instance, we fine-tune the probabilities corresponding to work activity type with commuting data at the municipality level, that is, the number of individuals that reside in a given municipality and commute for work to a given municipality. We do this by way of IPF. Let $N_w(z_{M_o}, z_{M_d})$ be the desired number of individuals that reside in municipality z_{M_o} and work in municipality z_{M_d}. Also, let $N(z, w, m)$ be the number of agents in zone z who use mode m for activity type w (work) such that it belongs to the short distance class. This is easy to deduce since we know the total number of agents in zone z who use mode m for activity type w as well as the number of agents (in zone z who use mode m for activity type w) for whom the home-work distance belongs to the long distance class (provided by the long distance OD matrices obtained either directly from Sampers or by way of IPF). Let $\mathcal{M}$ be the set of all modes. In order to fine-tune the probabilities corresponding to work activity type using the commuting data at the municipality zonal level, we iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$\forall z_o, z_d, \forall m \in \mathcal{M}:$$

$$N(z_d|z_o, w, m) \leftarrow N(z_o, w, m) \cdot \mathbb{P}(z_d|z_o, w, m) \tag{5.9}$$

$$N(z_d|z_o, w, m) \leftarrow N(z_d|z_o, w, m) \cdot \frac{N_w(z_{M_o}, z_{M_d})}{\sum_{m \in \mathcal{M}} \sum_{z_{c'} \in z_{M_o}} \sum_{z_{b'} \in z_{M_d}} N(z_{b'}|z_{c'}, w, m)}, \tag{5.10}$$

where $z_{M_o} \ni z_o, z_{M_d} \ni z_d$

$$\mathbb{P}(z_d|z_o, w, m) \leftarrow \frac{N(z_d|z_o, w, m)}{\sum_{z_{b'}} N(z_{b'}|z_o, w, m)} \tag{5.11}$$

Eq. 5.9 gives the number of agents whose work location is in zone z_d given that their home is in zone z_o and they use mode m for work trip, for each z_d, z_o, m, by multiplying the corresponding probability with $N(z_o, w, m)$. Letting z_{M_o} and z_{M_d} to be the municipalities containing DeSO zones z_o and z_d respectively, Eq. 5.10 scales the obtained numbers $N(z_d|z_o, w, m), \forall z_o, z_d, \forall m \in \mathcal{M}$ such that they are consistent with the desired number of individuals that reside in municipality z_{M_o} and work in municipality z_{M_d}. Eq. 5.11 transforms the obtained numbers into probabilities.

5.3.2 Travel mode assignment

We assign the modes to each trip occurring between activities in each agent's activity schedule in three-step: mode distribution, potential mode usage, and mode assignment.

Mode distribution. For the cases in Table 5.2 using the OD matrices that are either directly provided by the regional and national models or by way of IPF (the rows coloured green, blue and yellow in the table), we obtain the mode distribution using the matrices from the models directly. On the other hand, for cases using the gravity models, we employ the methodology further described below.

From the travel survey, we obtain the zone-specific mode distributions for each activity type. In order to ensure that we have sufficient number of data points for each zone and activity type, we calculate the mode distribution at the county level. We then make a simplified assumption that the mode distribution for a given activity type for a given DeSO zone is same as for the county the DeSO zone is a part of (Deso zones are subdivision of counties).

Potential mode usage. Once we deduce the number of agents in a given zone that use a given mode for reaching the location of a given activity type, we then determine the corresponding set of agents. For instance, two agents residing in the same zone may have different probabilities of using a car for going to work (depending on their ages, income, etc.). In order to make this distinction, we introduce the concept of *potential mode usage* of an agent, and define it to be the probability distribution over the usage of the different modes.

We use a neural network classifier trained on the national travel survey for deducing the potential mode usage of the agents. Since we consider 5 modes, and each mode could be either used or not, we have a total of $2^5 - 1 = 31$ classes (excluding the class signifying that none of the modes are used). Once we deduce the probabilities of belonging to the different classes for each agent, we obtain the probabilities of using the different modes. If each of the 31 classes represents a set of modes being used, and if $\mathbb{P}_i(c)$ is the probability of an agent i belonging to class c as per the classifier, we obtain the probability of using mode m as $\mathbb{P}_i(m) = \sum_{m \ni c} \frac{\mathbb{P}_i(c)}{|c|}$.

Since the mode usage behavior of individuals would generally depend of their travel times, we consider different classifiers for different travel time classes. Furthermore, we assign an agent

zero probability of using a particular mode if the agent does not qualify to use that mode for traveling, in general, and hence redistribute the probability equally over the modes that the agent is qualified to use. For instance, we assign the probability of an agent using a car as a driver to be zero if the agent does not have access to a car or is less than 18 years old.

Mode assignment. Now that we know the number of agents in a given zone that use a given mode for reaching the location of a given activity type as well as the potential mode usage of each agent, we proceed to deduce the mode that an agent would use for reaching the location of the given activity type.

We first deduce the probability of an agent i using mode m, given that the agent resides in zone z and the activity type under consideration is a; let this be denoted by $\mathbb{P}_i(m|z,a)$. We utilize the IPF technique, with the initialization value being $\mathbb{P}_i(m)$ that is obtainable from the agent's potential mode usage. Let $S(z,a)$ be the set of agents with home location in zone z and involved in activity type a, let $N(z,a,m)$ be the number of agents in zone z who use mode m for the activity type a, and let $\mathcal{M}$ be the set of all modes. We iterate over the following sequence of update rules for a fixed large number of iterations or until a desired level of convergence is reached.

$$\forall z, a, \forall i \in S(z,a):$$

$$\mathbb{P}_i(m|z,a) \leftarrow \frac{N(z,a,m)}{\sum_{j \in S(z,a)} \mathbb{P}_j(m|z,a)} \mathbb{P}_i(m|z,a) \tag{5.12}$$

$$\mathbb{P}_i(m|z,a) \leftarrow \frac{\mathbb{P}_i(m|z,a)}{\sum_{m' \in \mathcal{M}} \mathbb{P}_i(m'|z,a)} \tag{5.13}$$

Eq. 5.12 drives the values $\mathbb{P}_i(m|z,a)$ such that the expected number of agents using a given mode for a given activity type starting from a given origin zone, is approximately equal to the desired number of agents using that mode for that activity type starting from that origin zone. Equation (5.13) is the normalization step ensuring that the obtained values are indeed probabilities, that is, $\sum_{m' \in \mathcal{M}} \mathbb{P}_i(m'|z,a) = 1$. The final step in mode assignment is multinomial sampling of the mode from the deduced values of $\mathbb{P}_i(m|z,a), \forall m \in \mathcal{M}$. Thus, we deduce the mode used for reaching the locations of all of the primary activities.

We assume that an agent uses the same mode for all trips between home departure and the immediate next arrival at home. For example, if an agent's activity sequence contains multiple primary activities in the interval between the departure and arrival to home activity (e.g., $-H$-W-S-$H-$), we want to ensure that a common mode is used for reaching the locations of primary activities between the home activities.

5.3.3 Activity location assignment

We assign activity location for each individual agent's activity, first at the zone level and then at the building level. To deduce zones, given a primary activity, we first have the deduced $N(z,a,m)$ – the number of agents residing in zone z who use mode m corresponding to that activity. Also, we have the deduced $\mathbb{P}(z_D|z,a,m)$ – the probability that an agent residing in zone z travels to zone z_D for activity a given that it uses mode m corresponding to that activity. We can thus deduce the number of agents residing in zone z who use mode m corresponding to a given primary activity a, and who travel to zone z_D for the given activity. Let this quantity be denoted by $N(z,a,m,z_D)$, and it can be deduced by independently drawing $N(z,a,m)$ samples from the multinomial distribution given by $\mathbb{P}(z_D|z,a,m)$. Note that as before, here a refers to activity types work and school and primary 'other'.

In order to assign the destination zone corresponding to a primary activity for each individual agent, we follow a simple rule that, given a set of agents residing in a given zone and using a given mode corresponding to a given primary activity type, agents with a higher travel time per leg are assigned farther destination zones. This rule, by way of ordering, ensures that the correlation between travel times and travel distances is accounted for. For instance, if two agents reside in the same zone and use the same mode 'Car' for travelling to work, and if the travel time per leg for the first agent is higher than that for the second agent, then the first agent travels to a destination zone that is at least as far away as or further from the destination zone of the second agent.

While assigning the location for each primary activity of each agent at the level of zones (which are very fine) would suffice for most applications, certain applications (e.g., routing) may necessitate location information that is more spatially precise. Hence, we assign a building for the location of each activity. Recall that using our buildings' data, we can deduce the set of buildings that correspond to a given activity type in a given zone. In order to assign a building corresponding to each activity of each agent, we employ a simplified approach in our model – given an activity and its location at the zonal level, assign a building uniformly at random from the set of zones corresponding to that given activity type in the given zone.

5.4 Secondary activities

We assume that the trips to secondary activities use the same mode of transportation as the adjacent primary activities. For the location assignment of secondary activities, we employ an adapted form of the gravity model. Since a secondary activity's location depends on primary activities, we assign the secondary activity location using the location of an activity preceding and succeeding in the activity sequence. These two activities (reference points) are not necessarily adjacent to the secondary activity.

The adapted gravity model has two parameters corresponding to the distances of the secondary activity location from the two reference points. We calibrate these parameters using the national travel survey. As each of the two reference points could correspond to one of the activities {home, work, school, other}, we could potentially have $4^2 = 16$ different gravity models for each of the 5 modes for a total of 80 different gravity models. This is an unreasonably large number of models to be calibrated using travel survey which typically presents a very limited number of intermediate 'other' activities. In order to reduce the number of gravity models, we group the intermediate 'other' activities into 3 broad types (see Table 5.4, out of which type HOH captures dedicated 'other' activities) and the modes into 2 broad types (namely, motorized and non-motorized), thus resulting in a total of 6 gravity models.

5.4.1 Reference activities

In our procedure, each type of 'other' activity have a defined level of priority. We assign the locations of the 'other' activity instances based on their priority, that is, we assign the locations of the highest priority instances first and that of the least priority instances last. Table 5.4 shows the classification of 'other' activities in descending order of their priorities. As discussed earlier, a primary 'other' activity assigned locations holds the highest priority among all the 'other' activity types; we denote it by HOH. The next priority is for an 'other' activity that is visited between two activities belonging to set {home, work, school}, where not both the primary activities are 'home'. If one of the primary activity is 'home', we denote it by HOX, otherwise we denote it by XOY. The least priority is for an 'other' activity that is visited between an activity belonging to set {home, work, school} and another instance of 'other' activity; we denote it by XOO. In the column showing the considered types, the second letter represents the 'other' activity under consideration for which we aim to assign the location, while

Table 5.4: An overview of our approach for deducing locations of different types of 'other' activities According to the considered secondary activity, the previous activity type in the sequence (p_1), the previous to previous activity type (p_2), the next activity type (n_1), the next to next activity type (n_2)), and finally the columns (A_1 ref and A_2 ref) determining activities whose locations are used as references to deduce the location of the secondary activity.

type	p_2	p_1	n_1	n_2	A_1 ref	A_2 ref	
HOH	$-$	H	H	$-$	-	-	
	$-$	H	O	H	-	-	
	H	O	O	H	-	-	*
HOX	H	O	O	W/S	p_2	n_2	*
	W/S	O	O	H	n_2	p_2	*
	$-$	H	W/S	$-$	p_1	n_1	
	$-$	W/S	H	$-$	n_1	p_1	
	$-$	H	O	W/S	p_1	n_2	
	W/S	O	H	$-$	n_1	p_2	
XOY	W/S	O	O	W/S	p_2	n_2	*
	$-$	W/S	W/S	$-$	p_1	n_1	
	$-$	W/S	O	W/S	p_1	n_2	
XOO	H	O	H	$-$	n_1	p_1	
	W/S	O	W/S	$-$	n_1	p_1	
	H	O	W/S	$-$	n_1	p_1	
	$-$	W/S	O	H	p_1	n_1	
	$-$	$H/W/S$	O	O	p_1	n_1	
	O	O	$H/W/S$	$-$	n_1	p_1	#

the first and the third letters represent the reference activities (based on whose locations, the location of the considered 'other' activity would be determined).[7]

The motivation to formulate a set of rules for classifying the different 'other' activity instances and for determining the two reference activities is the folllowing. Say we have a sequence $-H$-O-O-$W-$. If we classify both these instances as HOX, the reference points for assigning the locations of both the 'other' activity instances would be that of home and work. So, conditional on these reference points, the locations of the two 'other' activity instances would be assigned independently of each other; this is unreasonable since they are adjacent activities. It is hence important that one of the 'other' activity instances is classified as HOX and the other one as XOO. The instance that is classified as HOX is assigned a location based on locations of home and work (as they are the reference points). Following this, the instance classified as XOO is assigned a location based on the location of the 'other' activity instance that is already assigned a location, and the location of either home or work.

5.4.2 Adapted gravity model

To deduce OD probability matrices for the secondary activities, we consider two reference locations, namely, a preceding activity location and a subsequent activity (say, A_1 and A_2) location in the activity sequence. Let $\mathbb{P}(z_b|z_{A_1}, z_{A_2}, m)$ denote the probability that an agent's secondary activity location is z_b, given that the locations of the two reference activities are z_{A_1} and z_{A_2}, and the mode used is m. As earlier, let $d(z_b, z_c)$ denote the spherical distance

[7]Recall that we consider a maximum of 3 consecutive instances of 'other' activity type in an agent's activity sequence. From Table 5.4, if we have a maximum of 3 consecutive 'other' activity instances, the classification of 'other' activity instances into $\{HOH, HOX, XOY, XOO\}$ is indeed mutually exclusive and exhaustive.

between zones z_b and z_c, where $d(z_b, z_b)$ is defined to be the radius of zone z_b. Let the gravity model parameters corresponding to the distances relative to locations z_{A_1} and z_{A_2} be β_m^1 and β_m^2 respectively. Let $s_{z_b}^o$ denote the *attraction potential* of zone z_b for the 'other' activity type. On similar lines as [14], the gravity model for secondary activities can hence be expressed as:

$$\mathbb{P}(z_b|z_{A_1}, z_{A_2}, m) \propto s_{z_b}^o e^{\beta_m^1 d(z_b, z_{A_1}) + \beta_m^2 d(z_b, z_{A_2})} \tag{5.14}$$

$$= \frac{s_{z_b}^o e^{\beta_m^1 d(z_b, z_{A_1}) + \beta_m^2 d(z_b, z_{A_2})}}{\sum_z s_z^o e^{\beta_m^1 d(z, z_{A_1}) + \beta_m^2 d(z, z_{A_2})}} \tag{5.15}$$

Note that in order to employ the above model, it is necessary to know the locations z_{A_1} and z_{A_2} of the reference activities.

We calibrate parameters β_m^1 and β_m^2 of the adapted gravity model, using the national travel survey. As discussed earlier, we group the modes into Motorized (Car, CarP, PT) and Non-Motorized (Bike, Walk), in order to not have an exceedingly large number of gravity models.

As data for calibration, we consider all activity subsequences in the travel survey corresponding to types $\{HOX, XOY, XOO\}$ presented in Table 5.4 where the mode used throughout the subsequence is either entirely Motorized or entirely Non-Motorized. For a given subsequence, if a reference activity is adjacent to the 'other' activity instance under consideration, the corresponding distance between the location of the instance and that of the reference activity can be directly obtained from the travel survey. However, if a reference activity is not adjacent to the 'other' activity instance under consideration, this implies the existence of another activity in-between the given instance and the reference activity. In this case, the distance between the location of the instance and that of the reference activity is computed as – the sum of the distances of the locations of the instance and the reference activity, from the location of the in-between activity.

Table 5.5: Gravity model parameters for secondary activity types

Other (intermediate) type	Motorized		Non-Motorized	
	β_m^1	β_m^2	β_m^1	β_m^2
HOX	−0.10	−0.07	−0.38	−0.34
XOY	−0.07	−0.13	−1.22	−1.15
XOO	−0.08	−0.10	−0.46	−0.60

Table 5.5 presents the calibrated values of the parameters. It can be understood from Eq. 5.14 that a more negative parameter value would mean that the probability to travel drops rapidly with an increase in the corresponding distance. One of the most obvious observations from the table is that the parameter values corresponding to the Non-Motorized mode type are more negative than those corresponding to the Motorized mode type. This is natural since when using a Non-Motorized mode, it is likely that the location of the secondary activity is more or less 'on the way' while moving from one reference location to the other; the deviation taken from the shortest path is likely to be much less as compared to the deviation taken when using a Motorized mode. It is also interesting to understand the implications of the parameters' values for the different types of 'other' activities. For type HOX for both mode types, we can see that parameter β_m^1 (corresponding to distance from home location) is more negative than β_m^2 (corresponding to distance from work/school location). This implies that when choosing a secondary activity location between home and work/school locations, its distance from home is

given a higher weight by a typical agent, and it is likely that the location is close to the agent's home.

We now describe how we employ the calibrated gravity models for assigning locations to secondary activities. The number of probability quantities is quadratic in the number of zones for the standard gravity model, whereas the quantities that we need to compute would be cubic for the adapted gravity model [8], since we have two reference locations and one location to be assigned in the adapted model. This would result in a computational complexity that is intractable in terms of both time and space. So, unlike the standard gravity model, we cannot consider all possibilities, and in fact, it is clear that we need not consider all possibilities.

We only consider pairs of reference locations that are visited according to the agents' activity sequences, while applying the adapted gravity model. Note that among all possible pairs of reference locations, only a very small fraction would actually be seen according to the agents' activity sequences. Furthermore, we consider that a secondary activity should be at a certain distance from the reference points. For instance, if the reference locations are very close to each other, it is with almost sure that the location of the secondary activity is also close to them. This assumption thus decreases the number of possible candidate zones for a secondary activity's location. In our model, we employ this by considering only those candidate zones for secondary activity location which satisfy the following: the distance between the first reference zone and the candidate zone is within a certain multiple $\mathcal{M}$ (we consider $\mathcal{M} = 2$) of the distance between the first reference zone and its corresponding furthest second reference zone. Say that an ordered pairs of reference zones (z_{A_1}, z_{A_2}) 'exists' if and only if it is applicable to the activity sequence of at least one agent according to Table 5.4. So, if ρ is the set of all ordered pairs of reference zones, which exist, given the first and second reference zones z_{A_1} and z_{A_2}, we consider a zone z_b as candidate zone only if:

$$d(z_{A_1}, z_b) \leq \mathcal{M} \cdot \max_{z_{A_2}:(z_{A_1}, z_{A_2}) \in \rho} d(z_{A_1}, z_{A_2}) \tag{5.16}$$

These reductions result in the number of probability entries that need to be computed, to be brought well within the tractability limits of modern day computers.

Using Eq. 5.15, $\forall (z_{A_1}, z_{A_2}) \in \rho, \forall m$, and $\forall z_b$ satisfying Equation (5.16), we can now obtain $\mathbb{P}(z_b | z_{A_1}, z_{A_2}, m)$: the probability that an agent's secondary activity location is z_b, given that the locations of the two reference activities are z_{A_1} and z_{A_2}, and the mode type used is m.

5.4.3 Zone and building assignment of secondary activities

For assigning zone corresponding to a secondary activity, we adapt the same rule as the primary activity – agents with a higher travel time per leg are assigned zones whose sum of distances from the given reference zones is larger. Following zone assignment, the building assignment of secondary activities follows exactly the same procedure as that of primary activities.

[8] To give an idea of this in the context of our model, the number of zones is in the order of 10^4 approximately. The number of all possible pairs of reference locations at the granularity of zones is hence in the order of 10^8. For each pair of reference locations, each zone would have a computed probability of being assigned a location for the considered secondary activity. This results in the number of probability entries being in the order of 10^{12}.

Chapter 6

Model Evaluation and Assessment

In this chapter, we present the assessment of model performance and validity of the SySMo model. We first perform in-sample evaluations showing the similarity of the results with the input data used to construct the model. Second, out-of-sample evaluations are performed by comparing the model outputs with data never used in the SySMo model. We also evaluate the performance of the ML technique, neural networks used in various steps in the methodology. These assessments present how well the ML technique performs with data sharing the same structure as the used data in SySMo to make predictions such as activity participation or activity duration. Table 6.1 shows for which steps of SySMo these were used. The comparisons made to validate the model include both independent distribution and dependent(joint) distributions such as agents' attributes and their activity duration. In SySMo, we adopt a sequential modelling approach in which the features regarding the personal characteristics or the activity schedules are deduced in different steps, instead of jointly deducing them. E.g the activity types are determined first, and then activities' duration. In order to understand to what extent the model maintains the correlation between the separately deduced features, the comparison over joint classes is important. In summary, we perform the following evaluations measures:

- Population synthesis
  - Errors in number of individuals with respect to 1) basic attributes (age, gender) in DeSO zones, 2) advanced attributes (employee, car ownership) in DeSO zones, and 3) joint classes in municipalities
  - Disparity between Household size and SCB data
- Activity generation
  - Distribution of activity durations
  - Distributions of activity start and end times
- Mode and location assignment
  - Comparison of total distance travelled (vs. Trafikanalys model)
  - Comparison of daily total travel distance (vs. Sampers model)

6.1 Population Synthesis

In the step of population synthesis (chapter 3), we combine ML, IPF, and sampling to create the static synthetic population. This section presents the model evaluation on this step, where

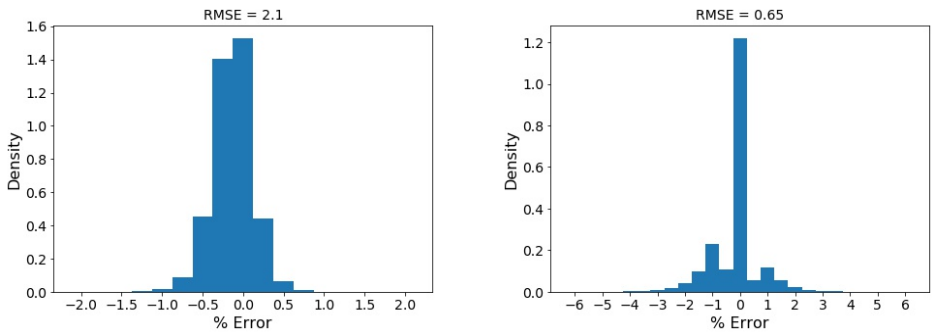
Table 6.1: Performance assessments

Steps	Evaluation types		
	ML performance	In sample	Out of sample
Population synthesis		✓	✓
Activity generation	✓	✓	
Mode and location assignment		✓	✓

the created population is validated against data from Statistic Sweden (SCB, Chapter 2). We calculate the percent difference in the number of individuals with respect to different attributes (age, gender or car ownership) in each DeSO zones and the distribution of the mean-square error (RMSE). To evaluate the performance of the home location assignment (section 5.1), we compare the household sizes in the SCB statistics with the generated synthetic population.

Basic attributes

For gender (Fig. 6.1), the error is between -0.5% and 0.5% in more than 92 percent of the DeSO zones. We find the RMSE = 2.1.



(a) Gender **(b)** Age groups

Figure 6.1: The percent error in the number of individuals by gender(a) and age groups(b).

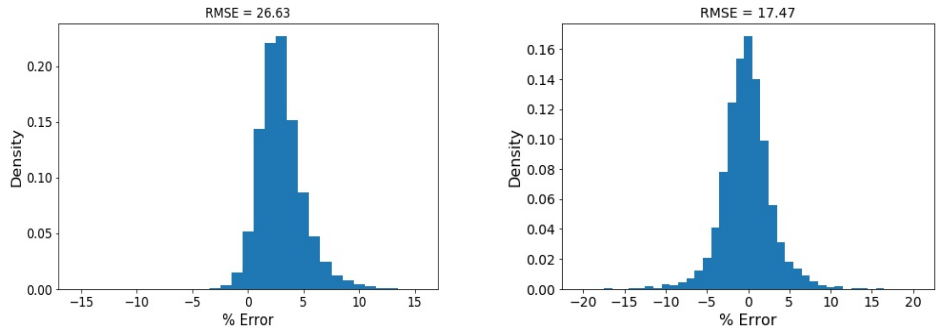
For age (Fig. 6.1), the error is between -1% and 1% in more than 78 percent of the DeSO zones. We found the RMSE to be 0.65. This indicates that 0.65 people in each Deso zone may have been assigned an incorrect age group. [9]

Advanced attributes

The advanced attributes are predicted using the assigned basic attributes (See Section 3). For the percent error in the number of employees (Fig. 6.2 a), the error is between -3% and 3% in more than 55 percent of the DeSO zones. The RMSE is 26.63, indicating that 26.63 people in

[9]The considered age groups in SySMo: 0, 1-6, 7-15, 16-18, 19-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-75, 75-84, 85+

each DeSO zone (populating an average of 1.706 people in each zone) may have been assigned an incorrect work status. Since it is a secondary attribute, i.e. derived based on the basic attributes, the error is expected to be higher.



(a) Employees **(b)** Cars

Figure 6.2: The percent error in the number of employees in each DeSO zones(a) and the percent error in the number of cars in each DeSO zones(b).

For the percent error in the number of cars in each DeSO zones (Fig. 6.2 b), the error is between -3% and 3% in more than 76 percent of the DeSO zones. We find an RMSE of 17.47, indicating that our estimated number of vehicles in each DeSo zone can deviate roughly by 17.5 vehicles.

Attributes over joint classes

For this part of evaluation, we calculate the percent error in the number of individuals by gender and age in each municipality. It is observed that the error is between -8% and 8% in more than 60 percent of the municipalities and RMSE is 140.31. The error is expected to be higher in this case, since it is calculated over joint classes and at the municipal level.

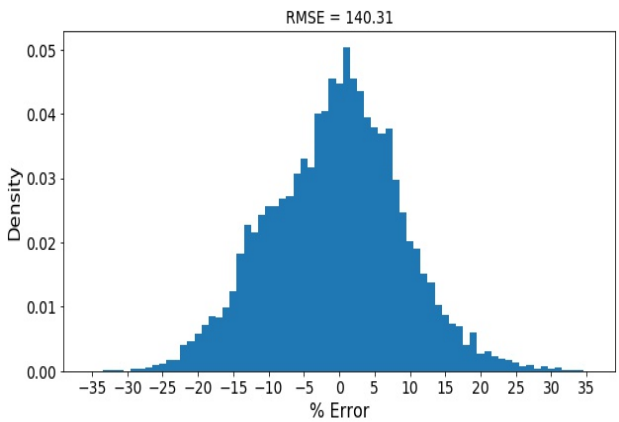


Figure 6.3: The percent error in the number of individuals by gender and age.

Household size

The home location assignment is the first step of the location and mode assignment (Chapter 5) where all activities are assigned to locations. In this step, we assign each household a specific residential building with a building type (e.g., detached house or apartment building). Since the home location assignment is correlated with household size and households are generated in population synthesis step, we place the household size evaluation here. In order to evaluate the performance of this step, we compare the household sizes of the synthetic population against national-level SCB statistics[11]. In SySMo, household size is an important parameter as it maintains the correlation between households and types of residence such as detached house or apartment building. The comparison suggests that our model produces similar household sizes to the official statistics (Table 6.2).

Table 6.2: Household size by dwelling types for Sweden

Dwelling Type	Synthetic Population	SCB Data
Overall average	2.2	2.2
Detached houses	2.7	2.7
Apartment buildings	1.8	1.9

Table 6.2 depicts a comparison of household size by different dwelling types from the synthetic population developed in the frame of SySMo to SCB statistic. The overall average household size is calculated as 2.2 persons per household in the synthetic population and the figure is the same as the statistics. The average household size living in a detached house in Sweden is 2.7 people per household, and we also capture the same number in the synthetic population. The average household size living in an apartment is slightly lower than that of a detached house, with 1.9 persons per household, while the average of 1.8 persons is found in the synthetic population.

6.2 Activity Generation

In this section, we focus on the evaluation of the activity generation step. First we evaluate the performance of the ML models used to generate the activity schedules. We employ a stratified cross-validation method through the Brier skill score. Following this, we compare the outcomes of the activity generation step regarding activity features with the travel survey. This assessment step can be categorised as in-sample evaluation. We calculate the Hellinger distance and Jensen–Shannon(JS) distances to assess the similarity between the distributions of activity duration and start-end time of the two datasets.

6.2.1 ML models evaluation

ML models in SySMO refer to a series of probabilistic machine learning methods applied in the step of activity generation (Section 4). They give probability distributions of class memberships instead of assigning a particular class label. To evaluate their performance, we first compare the output from the probabilistic ML models against the travel survey. Given the produced probability distributions are about class memberships, complex measures are needed to interpret and evaluate predicted probabilities. Brier Score (BS) is one of the metrics frequently used to measure the accuracy of probabilistic predictions [17]. The original definition of BS is applicable to multi-class problems by the formula set out as:

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{it} - o_{it})^2 \tag{6.1}$$

where f_{it} denotes predicted probability, while o_{it} is the actual outcome at the instance it. R denotes the number of possible classes, N is total number of samples in all classes. BS always takes on values in the range [0,1], where 0 means a perfect score. The results produced by the Brier Score can be very difficult to interpret when the classes are imbalanced. Brier skill scores (BSS) are calculated to validate ML models used in the activity generation step. BSS gives a score by comparing the BS with a reference measure. The most common formulation:

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{6.2}$$

BSS gives a value between $-\infty$ and 1 by comparing the Brier score with a reference measure BS_{ref} such as a naive model having a constant probability distribution, that shows densities of classes in the dataset for each sample in the dataset. [10] A score of 0 means the model results are identical to a naive model, whereas 1 is the best possible score meaning that predictions are identical to the data compared. A score below 0 means the results are worse than the scores calculated from the naive model. We do the evaluations for all ML models that are used to generate the activity pattern.

We employ the k-fold cross-validation method to evaluate our ML models with Brier scores. In the machine learning field, K-fold cross-validation is a widely used resampling method which divides all samples to fit a model and to measure the performance of the fitted model [18]. It works with the principle of dividing the data into a certain equal number of parts and using 1 part of it for scoring the model each time. In our evaluation step, we use the stratified cross-validation variation that maintains the distribution of the labels in each fold.

Probability of participating in work, school and other activities

Four ML models are created by status (employment = 0/1 and student = 0/1) (Section 4.1). For each model, we calculate BSS the predicted probability for joining work (W), school (S), and other (O) activities using the evaluation data and the predicted data. Table 6.3 presents the BSS scores from these four models. All BSS scores are above 0 except the model including only student status as positive (E = 0, S = 1) which has slightly lower accuracy than the naive model. This may be due to the definition of students being very broad and that these people could have very flexible schedules which are more difficult to model. The average BSS is 0.3067, and the weighted average BSS by people in each group is 0.1320.

Duration of work, school and other activities

SySMo has seven separate ML models by the participation sets of W, S, and O activity. A set of activity participation is denoted S, where $W, S, O \in \{0, 1\}$ and $S \setminus \{0, 0, 0\}$. For each

[10] E.g. let consider a multi-class dataset of 100 samples with 3 different labels. if the labels distribution is 20, 10, 70, respectively the 1st 2nd, and 3rd label, the naive model will be such that it preserves the labels' distribution by sampling. That is, the classes values of the naive model will be 0.2, 0.1, 0,7, respectively and it repeats the given number of samples.

Table 6.3: Brier skill scores for probability of participating in work, school, and other activities by employment (E) and student (S) status. A scores 0 means being identical to the naive model, whereas 1 is the best possible score. A score below 0 means worse scores than the scores calculated from the naive model.

Status	*Percentage of pop. (%)*	*BSS*	*Standard dev.*
E = 0, S = 0	21	0.2770	0.0307
E = 0, S = 1	21	-0.0516	0.1764
E = 1, S = 0	55	0.1020	0.0138
E = 1, S = 1	3	0.8995	0.0041

model, BSS measures the similarity of the predicted duration (in broad categories, see below) for W, S, and O between the evaluation data and the predicted data. The scores are reported in Table 6.3. All BSS scores are above 0, and some models scores such as (W = 0, S = 1, O = 1) are close to 1, the best possible score. The average BSS is 0.5528, and the average BSS weighted by people in each group 0.2682.

The broad duration classes for the activities are: Home = 0-12h, 12-18h, 18-24h; Work = 0-6h, 6-10h, 10-24h; School = 0-6h, 6-8h, 8-24h; and; Others = 0-2h, 2-5h, 5-24h (See more in Section 4).

Table 6.4: Brier skill scores for assessing the model performance on estimating the broad duration classes in work (W), school (S) and other (O) activity

Activity participation	*Percentage of pop. (%)*	*BSS*	*Standard dev.*
W = 1, S = 0, O = 0	38.1	0.1848	0.0156
W = 0, S = 1, O = 0	10.7	0.5585	0.0234
W = 1, S = 1, O = 0	7.2	0.6645	0.0219
W = 0, S = 0, O = 1	22.7	0.3933	0.0515
W = 1, S = 0, O = 1	21.0	0.0899	0.3003
W = 0, S = 1, O = 1	0.2	0.9953	0.0015
W = 1, S = 1, O = 1	0.1	0.9831	0.0031

6.2.2 Activity duration and start-end time distributions

One of the main outcomes of the activity generation step is the activity duration and the start-end time (See Section 4). We evaluate these outcomes against the travel survey by measuring the distance between the probability distributions of the model and the survey. Many different measurement methods can be seen in the literature, but the Kullback-Leibler divergence and squared Hellinger distance are one of the most prominent of these [19]. Therefore, we choose the Hellinger distance and a variation of Kullback-Leibler divergence that is Jensen–Shannon (JS) distance to perform the evaluations.

The probability distributions that we want to compare are p and q. We define the Hellinger distance as the Euclidean norm of the difference of the square root of p and q ($\sqrt{p}$ and $\sqrt{q}$ respectively) divided by the square root of two (Equation (6.3)). The Hellinger distance always takes on values in the range [0,1], where 1 is the maximum distance.

$$H(p,q) = \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2 \tag{6.3}$$

We utilise JS distance to evaluate the model's results. Kullback-Leibler divergence is a statistical distance but it does not qualify as a metric. Since it lacks properties of being a metric such as symmetry between each pair of points ($D(p,q) == D(q,p)$). JS is a symmetrized and smoothed variation of Kullback-Leibler divergence [20]. To calculate the JS distance, we deduce the Kullback-Leibler divergence at first. From the KL divergence the JS distance can be calculated with Equation (6.4). The distances have values in the range [0,1], where 1 means the maximum distance.

$$\begin{aligned} KL(p,q) &= \begin{cases} p\log(p/q) - p + q & p > 0, q > 0 \\ q & p = 0, q \geq 0 \\ \infty & otherwise \end{cases} \\ JS(p,q) &= \sqrt{\frac{KL(p,M) + KL(q,M)}{2}} \end{aligned} \tag{6.4}$$

For this example, M is the mean of p and q and $KL(p,q)$ is the Kullback-Leibler divergence. We use the scipy library implementation of the distance $KL(p,q)$ ([21]) in the evaluations.

Activity duration distribution by activity type

In order to evaluate the model performance, we compare the distributions of activity duration by activity type derived from the model output and the travel survey (e.g., Fig. 6.4).

The shorter the distance (close to 0), the closer the two distributions are to one another. We calculate the Hellinger and JS distance between the two distributions of work activity duration to 0.1054 and 0.1260 respectively. Although the calculated values for school are slightly higher than for work (0.1378 and 0.1645 respectively) they are is still quite close to zero.

Distribution of activity duration by activity type and personal attributes

Next we evaluate the activity duration distributions over the joint classes of activity type and personal attributes (Fig. 6.5 and Fig. 6.6). Besides measuring similarities between the distributions, we also evaluate to what extent the model maintains the correlations between outputs from the different steps. First, we compare the activity duration distributions by activity type and gender, one from SySMo and the other from the travel survey. Fig. 6.5 shows these two distributions. The Hellinger, and JS distances between work activity duration distributions are 0.1058 and 0.1260 respectively for males, and 0.1245 and 0.149 for females .

Fig. 6.6 illustrates activity duration distributions by home activity type and income levels. The population is divided into five income groups: no, low, lower middle, upper middle, and high. While the Hellinger distance between work activity duration distributions of the low-income group is 0.1396, and JS distance is 0.167, the distances between work activity duration

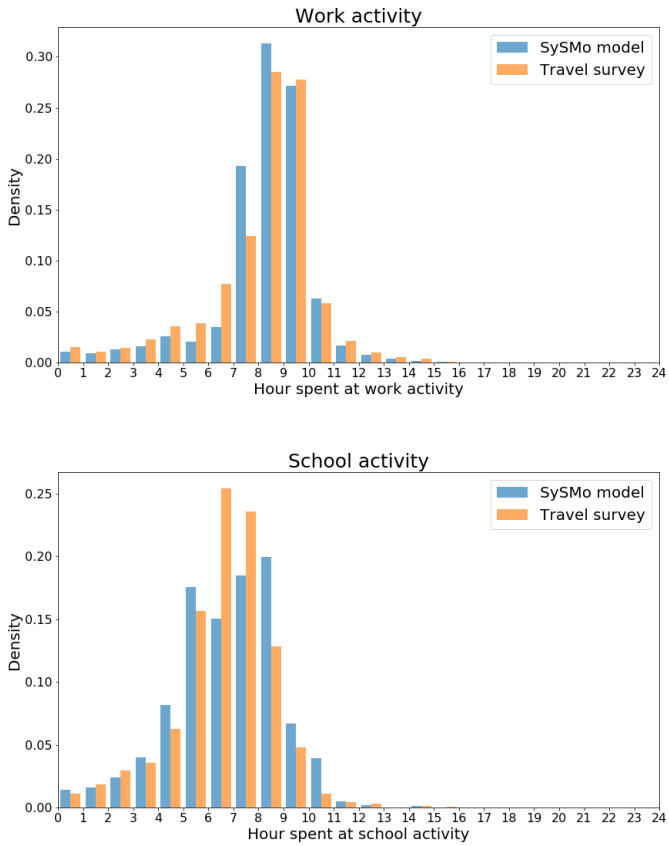


Figure 6.4: Comparison of activity duration by activity type.

distributions of the upper-middle-income group are 0.1353 and 0.1131, respectively. It is worth noting that the survey population contains mostly individuals having some activities during a day and has much fewer persons with no activity and staying in their homes all day. In contrast, SySMo also models these individuals having very high home activity duration to cover the entire population. To make two data with different numbers of samples comparable, we use densities instead of exact values in the y-axis for each bin in the histograms. A very high density for the bin corresponding to the population who spent 24 hours at home in the synthetic population results in lower densities being calculated for all the other bins. The small differences in the density values corresponding to the bins showing less than 24-hours spent at home can be explained by the used density-based representation.

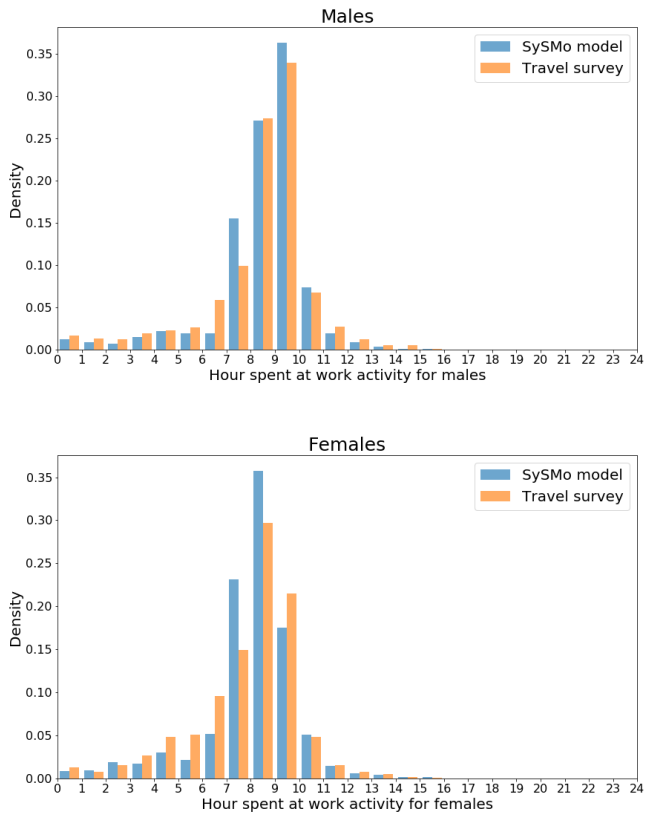


Figure 6.5: Comparison of activity duration by activity type and gender.

Distribution of activity duration by activity type and willingness to participate

We also evaluate the model performance on the distributions of activity durations over the joint classes activity type and activity participation of agents. Fig. 6.7 shows the duration of 'other' activity by whether or not participating in work activity. Since more than 99 percent of the sub-populations have less than 12 hours of other activity duration, we limit the x-axis to 12

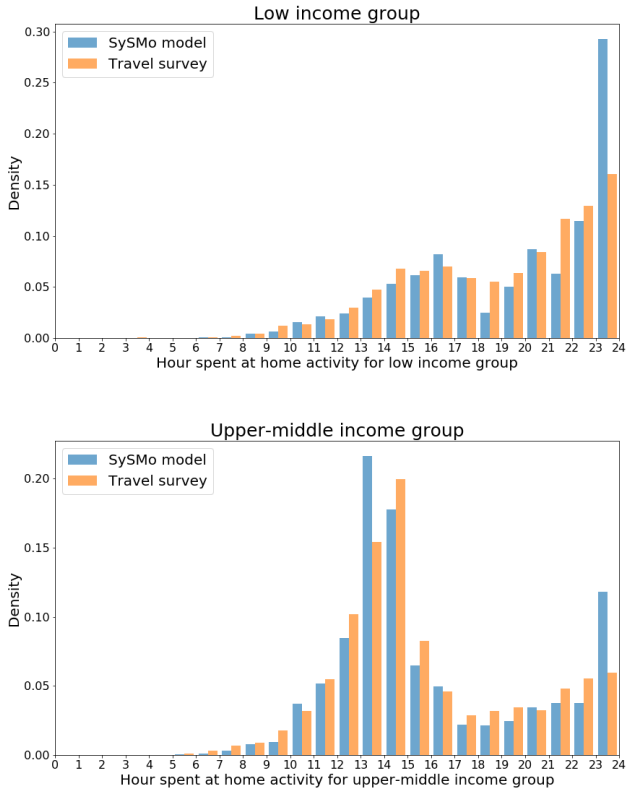


Figure 6.6: Comparison of activity duration by activity type and income group.

hours in the illustration. The Hellinger distance between other activity duration distributions for those participating in a work activity is 0.1990 and the JS distance is 0.1679. For those not participating in a work activity during the day, the Hellinger distance for other activity duration distributions is 0.2351 and the JS distance is 0.1986.

Start-end time distribution by activity type

Fig. 6.8 shows the end time distribution of the home activity instances, which take place at midnight (03:00). The Hellinger distance and JS distance are 0.0732 and 0.0876, respectively.

Start-end time distribution by activity type and activity participation

In this part, we evaluate the model's performance of the distribution of the start or end time of an activity over the joint classes of activity type and activity participation. Fig. 6.9 contains two panels. In the top panel: the distribution of the end time of the home activity, taking place at midnight (03:00), for the population participating in a work activity. For these distributions,

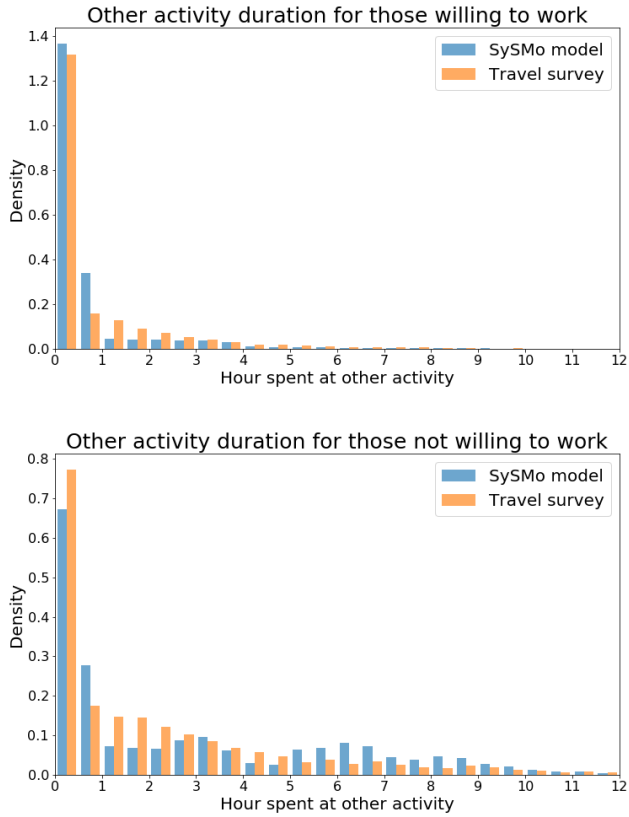


Figure 6.7: Comparison of activity duration by activity type and activity participation.

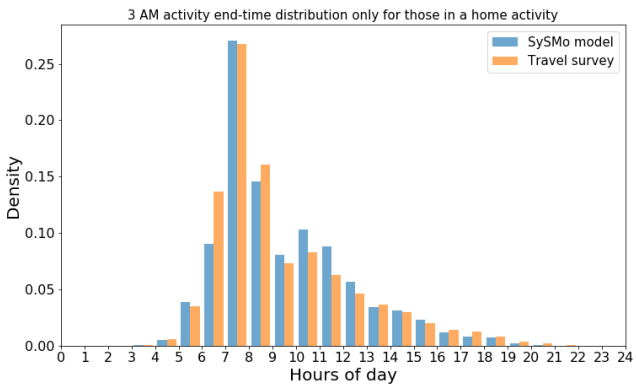


Figure 6.8: Comparison of activity end time distribution by activity type.

the Hellinger distance is 0.0993 and the JS distance is 0.1188. The bottom panel presents the

distribution of the end time of the home activity, which takes place at midnight (03:00), for the population who only participate in an other activity. The Hellinger distance is 0.0847 and the JS distance is 0.1012 for these distributions.

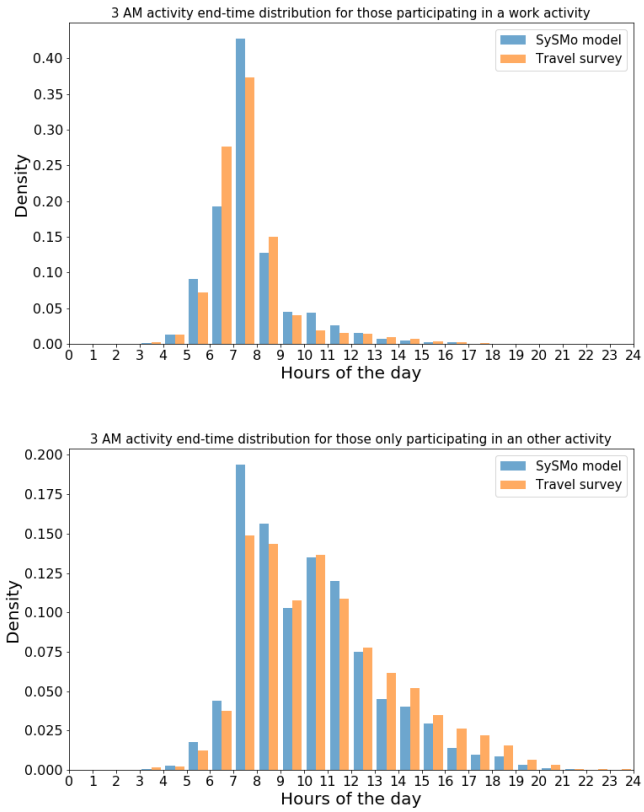


Figure 6.9: Comparison of activity end time distribution by activity type and activity participation.

6.3 Mode and Location Assignment

This section presents the evaluation of mode and activity location assignment. It is very difficult to find data showing departure and arrival points of trips by mode. Although new datasets emerge with the development of technology such as mobile phone call data [22], access to these data is not very easy and its reliability is questionable. One of the most common methods for evaluation is to compare the results with other model outputs. We perform out-of-sample evaluation by comparing results produced by SySMo with the Trafikanalys model and in-sample evaluation by comparing with the Sampers model.

Comparison of total distance travelled (vs. Trafikanalys model)

In this part of the evaluation, we use passenger and goods transport statistics describing the activity of the transport system (see Chapter 2 for more details). The statistics from Trafik-

analys shows the total distances travelled annually by modes from 2000 to 2020. After 2016, two figures are published for cars, bicycles, and walking trips since the agency adopted a new calculation technique. Since the SySMo model is developed based on the year 2018, we use the data corresponding to this year in the comparisons.

The SySMo model is developed to produce daily travel patterns corresponding to an average weekday or an average weekend day. However, the statistics from Trafikanalys are in the form of annual totals. In order to compare the outputs of the SySMo model with the data of Transport Analysis, we calculate the annual total by weighting the SySMo model outputs on weekdays and weekends. The Euclidean distances of the trips in SySMo are calculated by using the starting and ending locations. We multiply the Euclidean distances by $\sqrt{2}$ to find the actual road (network) distances [23]. We have applied this conversion only to Car Driver and Passenger modes. Data from the travel survey were also used for comparison. We scale up the distance per respondent by the weights given in the survey data and compute the total distance travelled by mode.

Table 6.5: Annual total passenger kilometres by mode in 2018 (in billions km) In the Trafikanalys column, the numbers calculated using the old technique are on the left side, and on the right side are from the new technique.

Mode	SySMo weighted by weekdays and weekends	Trafikanalys	Survey
Car Driver+Passenger	98	95 - 116	113
Public Transport	24	26	30
Bike	3	2.8 - 3.1	3.3
Walking	4	2.0 - 3.7	3.8

The comparison of annual total passenger kilometre suggests that our model results are very close to the Survey and Trafikanalys data (Table 6.5). While the passenger kilometre of car driver+passenger is calculated as 98 billions km in the SySMo model, it is 95 and 116 billions km in the Trafikanalys model according to the new and old techniques, respectively. Passenger kilometre by car driver+passenger is deduced 113 billions km from the Survey.

Comparison of daily total travel distance (vs. Sampers model)

The OD matrices from the Sampers models show the number of trips between the origins and destinations by different purposes such as work, other, business and private. In SySMo, we have 3 trip purposes namely work, school and other but only work trips are comparable with Sampers as the definitions of the trip purpose are the same in the both model (See more in Chapter 5). From Väst regional matrices, we calculate the daily total travel distance between activity locations using corresponding zone centres. On the other hand, in SySMo we have the exact activity locations to calculate travel distances. Since the regional models provide data with a lower spatial resolution out of their core area, we make comparisons within the Väst regional model's core areas. Even though these differences in the calculation of the daily trip distances lead to slightly different distributions, the overall patterns are captured. We use the spherical distance to calculate travel distances between activity locations in both datasets. Since there is no mode indicating car passenger in Sampers OD matrices, we calculate it using the official occupancy rates obtained from Trafikverket [24]. We show the daily travel distance of individuals between home and work trips by car, car passenger, public transport, bike, walk on

Fig. 6.10. We also calculate the Hellinger and JS distances between distributions and show on Table 6.6. Besides the illustrations, we report the statistical comparisons containing, median, mean, 90th percentile, and maximum values on Table 6.7.

Table 6.6: The Hellinger and JS distances between daily total travel distance distributions by the travel modes

Modes	H_dist	JS_dist
Car	0.046	0.117
CarP	0.070	0.171
PT	0.071	0.176
Bike	0.030	0.080
Walking	0.016	0.042

All the distributions are quite similar and the calculated distances also show the similarity. We deduce the Hellinger distance of 0.0479 for distributions showing the daily distance travelled by car. The car mode has the shortest Hellinger distances compared to other modes. It is followed by public transport with a 0.0631 distance score. The shortest JS distance among distributions corresponds to bike mode with 0.121. It is followed by the car mode with a 0.122 distance score.

Table 6.7: Comparison of daily total travel distance(km) by the travel modes

Modes	Median		Mean		Percentile_90		Max	
	SySMo	Sampers	SySMo	Sampers	SySMo	Sampers	SySMo	Sampers
Car	7.5	6.5	12.3	9.6	29.0	22.2	393.3	360.2
CarP	7.7	6.2	14.4	9.1	32.6	20.8	344.3	354.9
PT	6.8	4.6	13.2	7.7	29.7	18.6	376.0	92.7
Bike	1.9	1.9	2.9	2.5	6.3	5.1	399.1	298.0
Walking	1.1	1.4	1.8	2.1	3.9	4.8	366.3	114.8

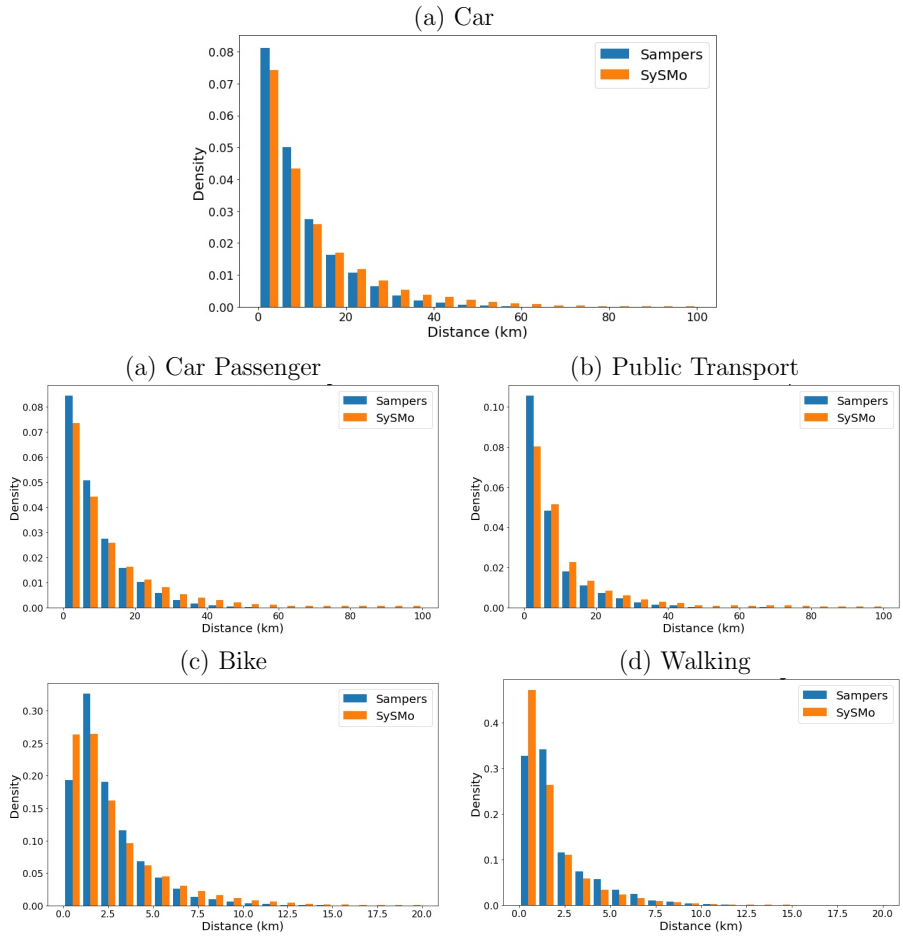


Figure 6.10: Comparison of daily travel distance of individuals between home and work by travel modes.

Bibliography

[1] J. Castiglione, M. Bradley, and J. Gliebe, *Activity-based travel demand models: A primer*, ser. SHRP 2 Report. Transportation Research Board, 2015, no. S2-C46-RR-1.

[2] "Statistics Sweden," https://www.statistikdatabasen.scb.se/pxweb/en/ssd/, 2020.

[3] "Demographic Statistical Areas (DeSO)," https://www.scb.se/en/services/open-data-api/open-geodata/deso--demographic-statistical-areas/, 2020.

[4] "The Swedish National Travel survey," https://www.trafa.se/en/travel-survey/travel-survey/, 2021.

[5] M. Beser and S. Algers, "SAMPERS: The new Swedish national travel demand forecasting tool," in *National Transport Models*. Springer, 2002, pp. 101–118.

[6] "GSD Property Map by Lantmäteriet," https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/geodataprodukter/produktlista/fastighetskartan/, 2020.

[7] Passenger and goods transport report. [Online]. Available: https://www.trafa.se/ovrig/transportarbete/

[8] M. H. Hafezi, L. Liu, and H. Millward, "Learning daily activity sequences of population groups using random forest theory," *Transportation research record*, vol. 2672, no. 47, pp. 194–207, 2018.

[9] K. Lum, Y. Chungbaek, S. Eubank, and M. Marathe, "A two-stage, fitted values approach to activity matching," *International Journal of Transportation*, vol. 4, no. 1, pp. 41–56, 2016.

[10] S. Dhamal, Ç. Tozluoğlu, S. Yeh, F. Sprei, M. Marathe, C. Barrett, and D. Dubhashi, "Synthetic Sweden: A spatially explicit agent-based mobility model with an advanced synthetic population," 2021.

[11] "Statistics Sweden: Households' housing," https://www.scb.se/en/finding-statistics/statistics-by-subject-area/household-finances/income-and-income-distribution/households-housing/pong/statistical-news/households-housing/, 2018.

[12] T. A. Arentze and H. J. P. Timmermans, *ALBATROSS: A learning based transportation oriented simulation system*. EIRASS, 2000.

[13] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transportation Research Part A: Policy and Practice*, vol. 35, no. 1, pp. 1–28, 2001. [Online]. Available: www.elsevier.com/locate/tra

[14] A. Adiga, A. Agashe, S. Arifuzzaman, C. L. Barrett, R. Beckman, K. Bisset, J. Chen, Y. Chungbaek, S. Eubank, S. Gupta *et al.*, "Generating a synthetic population of the United States," *Technical Report*, 2015.

[15] "Transportation and spatial modelling: Mode choice," https://ocw.tudelft.nl/wp-content/uploads/Lecture4.pdf, 2013.

[16] M. Lenormand, A. Bassolas, and J. J. Ramasco, "Systematic comparison of trip distribution laws and models," *Journal of Transport Geography*, vol. 51, pp. 158–169, 2016.

[17] G. W. Brier *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[18] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, vol. 26.

[19] M. Budka, B. Gabrys, and K. Musial, "On accuracy of pdf divergence estimators and their applicability to representative data sampling," *Entropy*, vol. 13, no. 7, pp. 1229–1266, 2011.

[20] F. Österreicher and I. Vajda, "A new class of metric divergences on probability spaces and its applicability in statistics," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 3, pp. 639–653, 2003.

[21] Kullback-leibler divergence scipy v1.7.1 manual. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.kl_div.html

[22] 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.

[23] C. L. Barrett, R. J. Beckman, K. Maleq, V. A. Kumar, M. V. Marathe, P. E. Stretz, T. Dutta, and B. Lewus, "Generation and analysis of large synthetic social contact networks," in *Proceedings of the 2009 Winter Simulation Conference M.* Winter Simulation Conference, 2009.

[24] "Analysmetod och samhällsekonomiska kalkylvärden för transportsektorn: ASEK 7.0," https://www.trafikverket.se/contentassets/4b1c1005597d47bda386d81dd3444b24/asek-7-hela-rapporten_210129.pdf, p. 10, 2020.

Paper B

The Heterogeneous Travel Activity of a Synthetic Population

The Heterogeneous Travel Activity of a Synthetic Population

Çağlar Tozluoğlu*[1], Swapnil Dhamal[1], Sonia Yeh[1], Frances Sprei[1], Yuan Liao[1], Madhav Marathe[2], Christopher Barrett[2], and Devdatt Dubhashi[3]

[1]Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden

[2]Department of Computer Science, University of Virginia, Charlottesville, United States

[3]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abstract

The Synthetic Swedish Mobility (SySMo) model is a large-scale agent-based model (ABM) that provides a scaffold on which to build decision support tools to model and analyze future mobility scenarios. It replicates a statistically accurate representation of the real population of Sweden, but is completely synthetic so that (a) it does not violate any privacy issues; and (b) the behaviours of the agents can be modified easily to create alternative scenarios. It is the latter feature that makes the model an ideal tool for modeling and analyzing future scenarios in which behavioral change constitutes the largest uncertainty. The current literature on synthetic population is limited to homogeneous activity patterns within sub-populations, and the scope tend to be local/regional. We develop a stochastic approach combined with machine learning (ML) to generate heterogeneous activity patterns for all agents in the *(SySMo)* model. For evaluation, we compare our generated activity schedules with the Swedish national travel survey. Our method offers an improved modeling tool to asses policy options for future sustainable transportation systems. The modeling tool can be a valuable planning and visualization tool for public and private stakeholders in Sweden. In additional, the methodology can be broadly applied to other regions with new data and carefully calibrated parameters.

Keywords: Activity generation; Agent-based modeling; Activity-based modeling; Machine learning; Daily activity pattern.

1. INTRODUCTION

Urbanization, increasing population, and unsustainable development of the current transportation system make it necessary to transform the transport sector Shukla et al. (2022). Autonomous cars, electric cars, shared mobility and new forms of micro-mobility such as electric scooters, electric bikes are examples of innovations that have changed people's travel behaviors Fulton (2018); Matyas & Kamargianni (2019), and have the potentials to transform the future transport systems. Activity-based modeling is a travel demand modeling approach that has grown in popularity over the last decade (M. Hafezi et al., 2018). The emerging big data sources and the growing computer processing power have enabled a faster development of activity-based models that capture the dependencies between trip chains, higher temporal and spatial resolution, and behavioral realism (Rasouli & Timmermans, 2014).

1.1. Literature review

The most widely employed methods for activity-based models can be grouped into three main categories: constraint-based models (Lenntorp, 1977; Jones et al., 1983; Dijst & Vidakovic, 1997), econometric models (Bhat et al., 2004; Bowman & Ben-Akiva, 2001; Vovsha & Chiao, 2006), and computational process models (Gärling et al., 1994; Miller & Roorda, 2003; Pendyala et al., 1997). The constraint-based models evaluate whether a given activity schedule is doable in a certain space-time context instead of predicting activity-pattern Lenntorp (1977). Econometric models deduce the activity schedule that provides the maximum utility to the individuals from their activity-travel choices based on the theory that individuals constantly desire to maximize their utilities from their choices Rasouli & Timmermans (2014). These approaches requires heavy mathematics and statistics to create activity schedules that accurately reflect human travel behavior. Studies using machine learning (ML) have achieved at least as good or better results than conventional techniques (Koushik et al., 2020). Additionally, the adoption of machine learning techniques simplifies the design of the models generating the activity schedules while yielding more accurate results.

Computational process modeling—also known as rule-based models since travel behaviors are modeled based on heuristic rules—is the latest to emerge (Rasouli & Timmermans, 2014). Some of the most prominent studies include ALBATROSS (Arentze & Timmermans, 2000), one of the first implementations of this approach by using decision trees. The model assumes that individuals make plans based on their activity priorities, therefore it groups activities into fixed (e.g. work) and flexible (e.g. shopping) activities (Ettema et al., 1993; Doherty, 2000). Allahviranloo & Recker (2013) model individuals' daily activity schedules consisting of activity type and sequence using a ML method called support vector machines (SVM). M. H. Hafezi et al. (2019) propose a new modeling framework to explore and understand activity patterns. Twelve clusters of daily activity patterns were defined using a fuzzy C-means (FCM) clustering algorithm. They extended their work to deduce the dependencies between activity type, activity sequence, and socio-demographic characteristics of individuals (M. Hafezi et al., 2018). AgentPolis is an open-source simulation framework using neural networks, where individuals can dynamically replan their activities at any point in time (Čertickỳ et al., 2015). Recently, a data-driven activity scheduler (DDAS) using supervised machine learning methods was introduced by Drchal et al. (2019). DDAS sequentially generates the activity schedule that consists of activity type, start-end time, location, and mode choice via four separate models.

Agent-based models of travel demand and activity-based models are often combined (Castiglione et al., 2015). Activity-based models are used to generate travel demand for each agent in ABMs. Characterising agents within an ABM typically comprises the following components: population synthesis, activity generation, and execution of activities. The synthetic population can then be used as agents in ABMs. The activity generation step assigns daily activity plans to the agents by using activity-based approach. Given their relative advantages, ABMs have gained popularity in the last decade and have become an important modeling tool in transport (González, Hidalgo, and Barabasi 2008; Chee et al. 2020), disease transmission (Wesolowski et al. 2012), terrorism (Waldrop 2018), electricity models (Ringler, Keles, and Fichtner 2016), etc. The transport ABMs have the advantage of being forward-looking (as opposed to static) and technically sophisticated (combing both supply – vehicle traffic and public transportation, and demand

– location, mode and activity).

1.2. Research gaps

Previous computational process models often assign homogeneous daily activity patterns for each sub-population or group (for instance, a particular income and age range), resulting in the same activity pattern for all individuals within the subgroup ((Arentze & Timmermans, 2000; Allahviranloo & Recker, 2013; M. Hafezi et al., 2018). While generating homogeneous activity schedules within subpopulation groups maybe a reasonable simplification for many applications, it may not always be sufficient to estimate the impacts of policy efforts that depend on, or can result in, significant behavioral changes. For example, D. Yang et al. (2013); Hutchinson (2018) explore the activity travel behaviours of senior people and show that heterogeneity provide better understanding of travel behavior. Thus, decision-makers can enact more informed policies that can increase the accessibility of these groups.

Additionally, a complete synthetic population with the activity pattern for Sweden does not exist. Several ABMs in Sweden simulate the future travel demand to provide inputs for transport planning. These ABM applications in Sweden, however, are so far limited to small regions (i.e., Stockholm (Canella et al. (2016)) or focusing on a single mode (e.g., electromobility for long-distance travel (Márquez-Fernández et al. (2021))).

1.3. Our contributions

We propose a novel approach to generating heterogeneous daily activity schedules (activity type, start-end time, duration, and sequence) for a synthetic population in Sweden. The methodology captures the heterogeneity in activity generation between individuals and creates realistic daily plans of the individual mobility. Using ML in conjunction with probability models, we maintain the heterogeneity by sampling from the probability distributions of the attributes such as activity types or activity duration constituting the daily schedules. Neural networks are selected as a machine learning technique in order to produce accurate results. Gunning & Aha (2019) shows in his study that neural networks have the high prediction ability of neural networks over the complex data set.The focus of this paper is to apply machine learning techniques rather than comparing machine learning techniques with other statistical methods. This paper is part of a large-scale project, called Synthetic Sweden Mobility (SySMo) Model (Tozluoğlu et al., 2022), that models the mobility patterns of the population in Sweden. To the authors' best knowledge there are no other studies that have previously done this at this scale.

The paper is organized as follows: Section 2 describes the major data sources used in the model. In Section 3 we describe the methodology of representing heterogeneous activity generation for a synthetic population. Section 4 describes model evaluation and assessment. In Section 5, we present the results and Section 6 discusses the limitations of the methodology, suggestions for future work, and conclusions.

2. Data Description

Our model relies on two main sources of data: Swedish static synthetic population from the SySMo model, and the Swedish national travel survey. We present a brief description

of the data in the sections below. Other data are explained in the paper where suitable.

2.1. Static Swedish Population

The first module of the SySMo model synthesizes a static population of Sweden. The population consists of over 10 million agents to statistically represent the entire Swedish population. Each synthetic agent in the population has certain attributes related to the agent's personal characteristics and the household to which they belong. The attributes consist of age, gender, civil status, residential zone, personal income, household income, car ownership, household size, and number of children ≤ 6 years old. The attributes also contain employment and student statuses of agents.

To create agents in the population, statistical data from Demographic Statistical Areas (Demografiska statistikområden - DeSO) is used as input. The population is then created by combining machine learning and IPF techniques.This ensures that the number of individuals by primary attributes such as gender or age groups in the created population in DeSO zones is almost identical to the official statistics for 2018. DeSO zones are published by Statistics Sweden (SCB) to provide high spatial resolution data *Demographic Statistical Areas (DeSO)* (2020). There are a total of 5,984 DeSO zones, which are drawn based on population size and governmental or physical boundaries. The zones are very small in densely populated areas like city centers, while relatively larger in rural areas.

2.2. Swedish National Travel Survey

The Swedish national travel survey (2021) provides data about the travel behaviour of anonymized individuals with their socio-economic and geographical characteristics. The survey period is between 2011 and 2016, and consists of around 40000 participants aged 6-84 years. The travel survey was conducted with individuals, not households. However, the survey respondents provide information regarding the household and number of people in the household. In the travel survey, the activity location information of individuals is deduced from the start and end point of travel. To use in the proposed methodology, we classify activity types into four broad classes: home, work, school, and other.

Each participant has one weight according to their socio-demographics and another weight based on the day the survey was conducted. These weights directly indicate the representative power of the respondent regarding socio-demographics or travel patterns (Liao et al., 2022). The total population can be generated using these weights.

In our model, we use the travel survey to train our ML algorithms and obtain characteristics of the synthetic population and their activities such as employment status, activity sequence, and trip modes.

3. METHODOLOGY

In this section, we describe the methodology developed for heterogeneous activity generation for a synthetic population. The activity generation framework comprises four major steps (Figure 1). The first main step is the assignment of activity types namely home, work, school and other activities to the individuals. The second main step includes the calculation of the daily total activity duration for each activity type. In the third main step,

an activity sequence is assigned to each individual through matching with an individual from the travel survey. The last main step calculates the duration of activity instances in the schedules and the creation of activity schedules containing activity type, activity sequence, and start and end times of activity instances for each individual. The proposed method directly or indirectly ensures heterogeneity in the population using sampling techniques from probability distribution, while at the same time mantaining correlations with the different attributes. Agents with similar attributes can have different activity participation sets, activity sequences, activity durations, and activity start-end times. Table 1 shows a summary of the variables used to explain the methodology.

Table 1: Summary of variables used in the activity generation module.

Symbol	Description
H	home activity
W	work activity
S	school activity
O	other activity
t_A	duration of activity type A
θ_A	willingness for activity type A
ψ_W	employment status
ψ_S	student status

In the proposed methodology, we model daily travel patterns corresponding to an average weekday and an average weekend separately. The travel patterns on weekdays and weekends are significantly different (Rutherford et al., 1997; Quade, 2000). While people have more commuting trips or school trips during the weekdays, more recreational trips on the weekends. In the paper, we report results illustrating an average weekdays.

This article contains a brief summary of the methodology of the activity generation module of SySMo model; interested readers can refer to the SySMo model documentation (Dhamal, Tozluoğlu, et al., 2022) for more details. It can be noted that we utilize neural networks in various steps with the following hyperparameters; the stochastic gradient descent solver, 1 hidden layer, and 100 nodes in the network.

3.1. Activity Participation

For each agent in the synthetic population, we assign a set of activity types that the agents have the potential to be involved in. Four types of activities are considered: *home* (H), *work* (W), *school* (S), and *other* (O) like visiting shops, restaurants, etc. It is assumed that each individual visits the home at least once a day and each individual is willing to join the home activity Schläpfer et al. (2021); Barbosa et al. (2018). Therefore, our model does not include a separate step to determine an individual's willingness for home activity.

Let the variable capturing the daily duration of an activity type A be t_A, where $A \in \{H, W, S, O\}$. The willingness to participate in activity type A is denoted by θ_A, where $A \in \{W, S, O\}$ since H is always = 1. We model jointly an individual's willingness to work (θ_W), study (θ_S), and 'other' activities (θ_O) given its socio-economic attributes. The considered attributes are age, gender, civil status, residential zone, personal income, household income, car ownership, household size, and number of children ≤ 6 years old (see more in Section 2). Modeling over joint classes preserves the correlation between the participation of the different activity types. We develop four models depending on

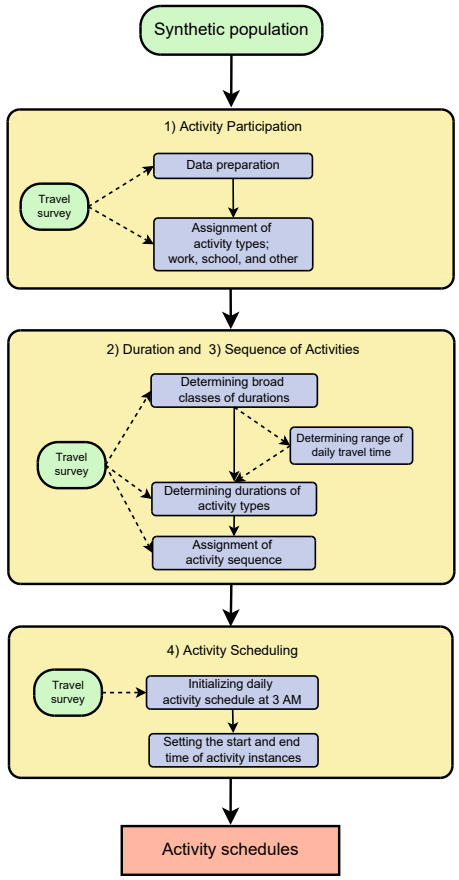


Figure 1: Methodology overview of the activity generation module of *Synthetic Sweden Mobility (SySMo) model.* *Yellow rectangles: major steps of the activity generation; purple rectangles: steps of the calculations; green ellipses: input data; pink rectangle: model outputs of activity schedules for each individual.*

the employment status (0/1) and student status (0/1). The status considered are: neither employee nor student (0, 0), only employee (1, 0), only student (0, 1), and both employee and student (1, 1). Developing four separate models ensures that non-employees do not participate in work activities and non-students do not participate in school activities.

The Swedish national travel survey is used for training the classifier; the features considered are age, gender, civil status, coordinates of the municipality's center, household size, number of vehicles owned, income level, and number of children ≤ 6 years old in the household. The model's output $\mathbb{P}_i(\theta_W = x, \theta_S = y, \theta_O = z)$ denotes the probability that a synthetic agent i's willingness to work is x, willingness to study is y, and and willingness for 'other' activities is z, where $x, y, z \in \{0, 1\}$. A class is hence assigned for every synthetic agent using multinomial sampling corresponding to the deduced probabilities. The used multinomial sampling technique allows agents to have different sets of activities, even if they have very similar attributes. Thus, every agent is assigned its willingness to work, school, and other activities.

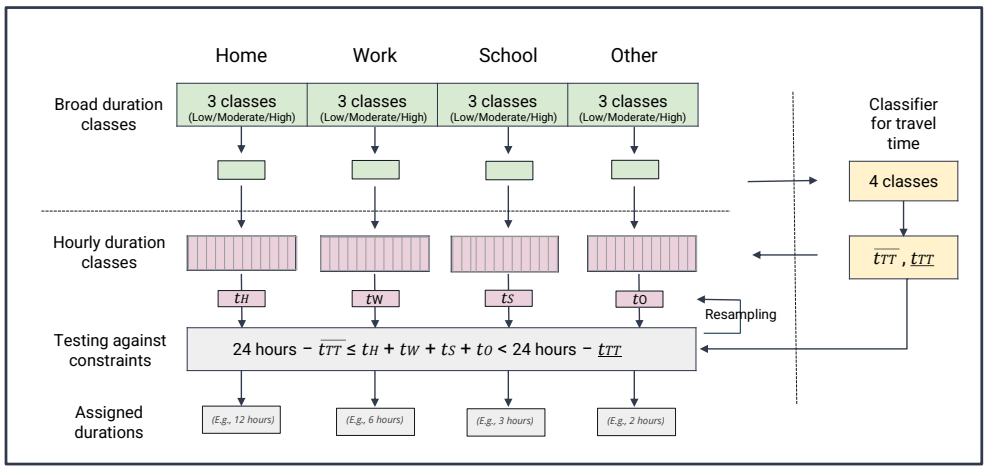


Figure 2: The flow chart of activity duration assignment methodology in SySMo. Green rectangles: joint model for broad activity duration, yellow rectangles: model for travel time, pink rectangles: model for hourly activity duration, and gray rectangles: final activity duration satisfying the constraint.

3.2. Activity Duration

The duration of different activity types are determined using a two-step method applying neural network classifiers and sampling techniques. Fig. 2 shows the flow chart of activity duration assignment methodology. In the first step, we jointly deduce broad duration classes for the different activity types; this enables us to capture the correlation between the duration of the different activity types. Using these broad classes and attributes of individuals, we deduce the overall travel time in a day. In the second step, using the deduced broad classes of duration of all the activity types and the range of daily travel time, we derive duration of all the activity types for each agent.[1] The method proposed here replicates people's heterogeneity in the population by allowing agents with similar attributes to have different activity duration.

Broad Activity Duration Classes

In the first step, we deduce broad duration classes by classifying an individual's total activity time for different activities as low, moderate or high. The definitions of low, moderate, and high depend on the activity type. The broad duration classes' boundaries are determined by dividing the generated hourly activity duration distribution from the travel survey into classes of approximately equal size. We consider for the different activity types are (in hours):

- Home: (0,12], (12,18], (18,24]
- Work: (0,6], (6,10], (10,24]

[1] if we directly deduce hourly duration classes for the 4 different activity types instead of the two-step method while preserving correlations, the number of potential joint classes would be $\binom{24}{4} = 10{,}626$. This is an exceedingly high number of classes and almost impossible to model with the existing data set.

- School: (0,6], (6,8], (8,24]
- Other: (0,2], (2,5], (5,24]

Since we have 3 broad classes for each of the 4 activity types, the total number of joint classes is $3^4 = 81$. Neural network classifiers are trained using the same socio-economic attributes as in the previous step and employment/studenthood statuses to deduce joint broad duration classes. The classifier produces the probabilities of an agent belonging to the joint classes. We develop different classifiers for different sets of activity participation, to increase the robustness of the classifiers. Since all individuals are assumed to be involved in home activity, there are 7 different activity participation sets in total, excluding not participating in any activity (i.e., a set of activity types is of the form $\{H\} \cup S$, where $S \in 2^{\{W,S,O\}} \setminus \{\}$). After assigning the joint classes' possibilities to all agents using the classifiers, the broad classes of duration of activity types are hence assigned using multinomial sampling.

Daily Travel Time Range

To deduce more specific durations of the different activity types , we estimate the total daily travel time for each agent. The sum of the duration of the activities is then set equal to 24 hours minus the total travel time. We consider 4 classes for estimating daily travel times, namely (the numbers are hours): (0,0.5], (0.5,1], (1,2], (2, 6]. These classes are approximately based on the four quartiles for daily travel time in the travel survey. The class assigned to an agent is $\left(\underline{t_{TT}}, \overline{t_{TT}}\right]$, the lower limit of the range of its daily travel time is $\underline{t_{TT}}$ and the upper limit is $\overline{t_{TT}}$.

For all agents, a neural network classifier is trained using the travel survey, the features being the socio-economic attributes, the employment and studenthood statuses, the set of activity types, and the broad classes of duration deduced above. The classifier outputs the probability distribution over the 4 classes of total daily travel time for each agent. From the probability distribution, each agent is assigned total travel time class through multinomial sampling.

Duration of Activity Types

After the broad classes of activity duration and the range of daily travel time are calculated, an hourly duration for each activity type is assigned. First, we deduce the probability distribution over hourly duration of each activity type, by considering 24 hourly classes per activity type. Then, we sample the duration of all types of activities such that they collectively satisfy Constraint (1). This constrain implies that, the sum of the duration of the activity types should be within 24 hours minus the range of the day's total travel time.

$$24 \text{ hours} - \overline{t_{TT}} \;\leq\; t_H + t_W + t_S + t_O \;<\; 24 \text{ hours} - \underline{t_{TT}} \tag{1}$$

To calculate the probability distributions over the 24 hourly duration classes, we use neural network classifiers. An hourly duration class is of the form $[T, T+1)$ hours, where $T \in \{0, 1, \ldots, 23\}$. The features considered are socio-economic attributes, employment

and studenthood statuses, willingness for the activity types, broad classes of duration of the activity type, and the class corresponding to daily travel time. We develop a classifier for each of the 3 broad duration classes of each activity type, thus, 12 classifiers in total.

We now deduce the duration of all activity types such that their sum satisfies Constraint (1). There are fundamentally two ways to achieve this, namely, mathematically [2] and simulation-based. In our model, we employ a simulation-based approach since the implementation of the method is easier and can be implemented in one step without having to first create a truncated joint distribution. For an agent, we first sample the hourly duration of the four activity types from the aforementioned probability distributions. Then, numbers sampled uniformly at random in [0,1) are added to each of the sampled hourly activity duration to introduce idiosyncratic variances and generate the final duration even in minutes. If Constraint (1) is satisfied for an agent, the four activity types are assigned the sampled duration. On the other hand, if the constraint is not satisfied, we repeat the sampling for the hourly duration and the idiosyncratic variances procedure. After 30 iterations, about 99% of the agents are assigned duration while satisfying Constraint (1). In order to ensure that all agents satisfy the constraint, we employ a simple heuristic procedure that trims or adds sampled times.

3.3. Activity Sequencing

For each agent in the synthetic population, we assign an activity sequence providing information about the frequency, order and type of activities participated in a day. We assume that individuals with similar socio-economic attributes and activity durations, would have similar activity sequences. While similarity between two sets of activity duration (t_H, t_W, t_S, t_O) could be quantified since they have the same unit (i.e. time), it is not clear how similarity between two sets of socio-economic attributes (e.g., age, gender, etc.) can be quantified, since they do not have the same unit and are not directly comparable. To utilize both socio-economic attributes and activity durations in similarity measurements, we employed a two-step method to assign the daily activity patterns to the agents in the synthetic population.

For each agent, we first choose a set of candidate individuals sharing similar socio-economic attributes from the travel survey. We consider only individuals having the same set of willingness for the activity types. To have as many similar socio-economic attributes as possible between synthetic and survey populations, we gradually filter candidate individuals based on their attributes. The gradual process ensures that a certain number of candidates remain in the set after each filtering.

We use daily activity durations as proxy parameters to determine the most similar individual (we use a method similar to Lum et al. (2016)). The similarity is measured using the Euclidean distance in the 4-dimensional space, between activity duration' tuples (t_H, t_W, t_S, t_O). For a synthetic agent, we choose the individual from the previously identified candidates using the Euclidean distance. Then, the sequence is directly copied from the individual chosen to the agent.

[2] In the mathematical approach, one would need to create a truncated joint distribution of the hourly duration of the four activity types, which can be obtained by combining the distributions of the activity types' duration and truncating to satisfy Constraint (1). The hourly duration can then be sampled from this truncated joint distribution, followed by adding a few minutes to the hourly duration so as to introduce a natural idiosyncratic variance, while ensuring that Constraint (1) is not violated.

3.4. Activity Scheduling

With the duration of the different activity types and the activity sequence ready, we generate the activity schedule for each agent in the synthetic population. We begin by assuming that the day starts and ends at 3 AM, since this is the time of day with the least number of individuals travelling according to the travel survey. We then deduce the start and end times of the activity that takes place at 3 AM. Modeling the start and end times of the 3 AM activity will help arrange the remaining activities during a day using the activity sequence and duration, since the head and tail of the sequence is defined. Thereafter, we distribute the total duration of an activity type among its individual instances in the activity sequence, and hence generate an activity schedule.

The start and end times of the 3 AM activity

The 3 AM activity type is directly obtainable from the first/last activity type in the deduced activity sequence. We thus start by determining its start and end time. Let $a_{3\text{AM}}$ denote the 3 AM activity instance and $t_{a_{3\text{AM}}}$ be its duration. Let $T^s_{a_{3\text{AM}}}$ and $T^e_{a_{3\text{AM}}}$ denote the start and end times of the 3 AM activity. In order to deduce $T^s_{a_{3\text{AM}}}$ and $T^e_{a_{3\text{AM}}}$, we first deduce the hourly distributions, using neural network classifiers (with 24 classes each) trained using the travel survey. As we did when determining activity duration, we develop different classifiers for each activity type.

For the sampling process, we impose a certain constraint with regard to the amount of time spent for the 3 AM activity: it should not exceed the total duration of the activity type corresponding to the 3 AM activity. We impose a lower bound such that the mean of the upper and lower bounds equals the deduced time of the 3 AM activity instance. Let $D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}})$ denote the deduced amount of time spent for the 3 AM activity instance. Since we have already calculated the total duration of the activity type $A_{3\text{AM}}$, the fraction of the total this total time that is allotted to the 3 AM activity instance can be denoted $f_{3\text{AM}} = \frac{t_{a_{3\text{AM}}}}{t_{A_{3\text{AM}}}}$. We deduce $f_{3\text{AM}}$ by way of regression using neural network trained using the travel survey. To have a lower bound such that the mean of the upper and lower bounds equals the deduced spent time for the 3 AM activity instance, we formulate the lower bound as $\big(1-2(1-\hat{f}_{3\text{AM}})\big)$. We hence obtain the following constraint:

$$\big(1-2(1-\hat{f}_{3\text{AM}})\big)t_{A_{3\text{AM}}} \;<\; D(T^s_{a_{3\text{AM}}}, T^e_{a_{3\text{AM}}}) \;<\; t_{A_{3\text{AM}}} \tag{2}$$

We sample the start and end times of the 3 AM activity instance from their corresponding hourly distributions deduced earlier, and add idiosyncratic variances to them to obtain times that satisfy Constraint (2). We employ a similar simulation based approach as the one used for sampling activity duration. While the sampled start and end times of the 3 AM activity instance satisfying the constraint are assigned to agents, the start and end times that do not satisfy are iteratively re-sampled. For the small fraction of agents whose start and end times of the 3 AM activity instance do not satisfy Constraint (2), we employ a simple heuristic procedure to ensure that they do.

Start and end times of activity instances

To find the start and end times of remaining activity instances in the sequence, we distribute the activities total duration equally among those of the same type in the sequence, i.e. if an individual goes to work, than other activity, and then back to work, the two work duration will be of equal length. After computing duration of all activity instances in the daily activity schedules of all the agent, we assign the travel times between adjacent activity instances. To calculate the total travel time for the day, the sum of the duration of the different activity types is subtracted from 24 hours (i.e., $t_{TT} = 24 - (t_H + t_W + t_S + t_O)$). We then distribute the total daily travel time equally across the different trips between the activities in the sequence. Note that in this trip we deduce a total daily travel time, while earlier (see section x), we had only deduced a range.

Now that we have a temporal arrangement of all activity instances within a day for every agent (that is, the activity sequence along with the start and end times of each activity instance), the daily activity schedules of all the agents in the synthetic population are ready.

4. Model Evaluation and Assessment

In this section we present the evaluations of the methodology using the Swedish national travel survey. We first evaluate the performance of the neural networks used in several steps in the methodology. Then, we perform in-sample evaluations showing the similarity of the model results with the input data used to construct the model.

4.1. ML models evaluation

We evaluate the probabilistic NNCs using the travel survey as ground truth data. Brier Score (BS) is one of the metrics frequently used to measure the accuracy of probabilistic predictions (Brier et al., 1950). However, the results produced by the Brier Score can be very difficult to interpret when the classes are imbalanced. We thus use the k-fold cross-validation method with the Brier skill score (*BSS*). To have a similar label distribution to the data in each fold, we run stratified cross-validation (For an in-depth explanation, see (Dhamal, Tozluoğlu, et al., 2022)). BSS gives a score by comparing the BS of the model with BS of a reference measure. The most common formulation of BSS is

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{3}$$

BSS gives a value between $-\infty$ and 1 by comparing the Brier score with a reference measure. A naive model having a constant probability distribution that shows densities of classes in the data set for each sample can be used as a reference measure. A score of 0 means the model results are identical to a naive model, whereas 1 is the best possible score meaning that predictions are identical to the data compared. A score below 0 means the results are worse than those from the naive model. We perform the calculation for all ML models used to generate the activity schedules.

The models of participating in activities, and the models of broad activity duration are reported here (Table 2). Four ML models are created by status (employment = 0/1 and student = 0/1). For each model, BSS scores are calculated for participating in work (W), School (S), and other (O)) activities between the validation data and the prediction data sets. Table 2 presents BSS scores. All BSS scores are above 0 except the model including only studenthood status as positive (E = 0, S = 1) which has slightly lower accuracy than the naive model. This may be due to the definition of students being very broad and that these people could have very flexible schedules that are more difficult to model. The average BSS = 0.3067, and the weighted average BSS by people in each group = 0.1320.

Table 2: Brier skill scores for probability of participating in work, school, and other activities by employment (E) and student (S) status. A scores 0 means being identical to the naive model, whereas 1 is the best possible score. A score below 0 means worse scores than the scores calculated from the naive model.

Status	*Percentage of pop. (%)*	*BSS*	*Standard dev.*
E = 0, S = 0	21	0.2770	0.0307
E = 0, S = 1	21	-0.0516	0.1764
E = 1, S = 0	55	0.1020	0.0138
E = 1, S = 1	3	0.8995	0.0041

ML models are employed to predict the duration of home (H), work (W), school (S), and other (O) activities. For the broad duration classes, we developed seven separate ML models by W, S, and O activity participation sets (See more in Section 3.2). A set of activity participation is denoted S, where $W,S,O \in \{0,1\}$ and $S \setminus \{0,0,0\}$. For each model, BSS calculates the match of the predicted duration (in broad categories, see below) for W, S, and O between the validation data and the prediction. The scores are reported in Table 2. All BSS scores are above 0, and some models scores such as (W = 0, S = 1, O = 1) is close to 1, the best possible score. We found the average BSS = 0.5528, and the weighted average BSS by people in each group BSS = 0.2682.

Table 3: Brier skill scores for matching the broad duration classes in work (W), school (S) and other (O) activity

Activity participation	*Percentage of pop. (%)*	*BSS*	*Standard dev.*
W = 1, S = 0, O = 0	38.1	0.1848	0.0156
W = 0, S = 1, O = 0	10.7	0.5585	0.0234
W = 1, S = 1, O = 0	7.2	0.6645	0.0219
W = 0, S = 0, O = 1	22.7	0.3933	0.0515
W = 1, S = 0, O = 1	21.0	0.0899	0.3003
W = 0, S = 1, O = 1	0.2	0.9953	0.0015
W = 1, S = 1, O = 1	0.1	0.9831	0.0031

4.2. Activity duration and start-end time distributions

We also compare our results regarding the activity duration and the start-end time distributions with the travel survey. The comparisons are performed for different subgroups of the population based on agent attributes and activity features. Below we plot the density histograms of activity durations, both from our model and the travel survey (Figure 3) by joint classes (i.e. by activity type and agent attribute), and start-end time for a selected activity (Figure 5) by a single class (e.g. activity type). For each plot, we calculate the

Hellinger distance and Jensen–Shannon (JS) distances to quantify how similar the distributions are (see Tozluoğlu et al. (2022) for more details). These distances have values in the range [0,1], where 1 denotes the maximum distance between the distributions. If each class with a value of 0 in one distribution gets a positive value in the other distribution or each class with a positive value in one distribution gets a value of 0 in the other distribution, the distance between these two distributions will be the maximum distance, 1. E.g., $\mathbb{P}_i(\theta_1 = x, \theta_2 = y, \theta_3 = z)$ denotes the probability that an instance i being θ_1 is x, being θ_2 is y, and being θ_3 is z, where $x, y, z \in \{0, 1\}$. If $\mathbb{P}_j = (1, 0, 0)$ and $\mathbb{P}_k = (0, 0.6, 0.4)$, the distance between j and k instances will be 1, which indicates that the distributions are far from each other.

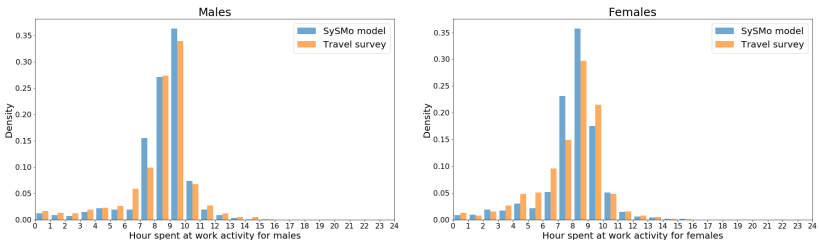


Figure 3: Comparison of work activity duration by gender. The left panel shows hours spent at work activity for males and the right panel shows hours spent at work activity for females.

Figure 3 shows that the Hellinger distance for distribution of work activity duration of males is 0.11, and the JS distance is 0.13. The mean duration of work activity in the survey and synthetic populations are 8.3 and 8.5 hours, respectively. We found the Hellinger distance for the distribution of work activity duration of females is 0.1245, and the JS distance is 0.149. The mean duration of work activity in the survey and synthetic populations are 8 and 7.8 hours, respectively. The work activity duration distributions by gender obtained from the model show similar work activity duration distributions as the survey data. For the school, home, and other activity duration distributions by gender, we calculate the Hellinger distance in the range of [0.1071, 0.2113], and the JS distance in the range of [0.1282, 0.2518]. The largest distance is obtained in the comparison of hours spent at other activity for females. Apart from this, the distances are less than 0.18 in all other comparisons.

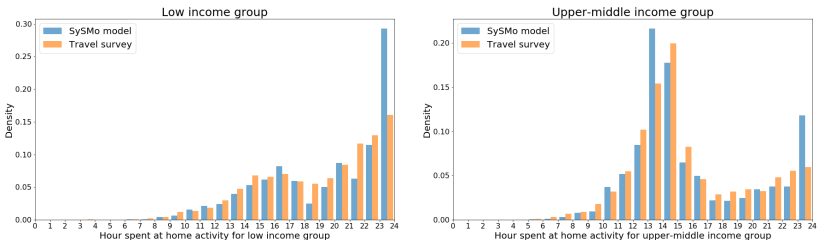


Figure 4: Comparison of home activity duration by income group. The left panel shows hours spent at home activity for individuals in low income group and the right panel shows hours spent at home activity for individuals in upper-middle income group.

We also compare the distributions for the five different income classed in out model: no income, low, lower middle, upper middle and high. For illustration, we report the

distribution of hours spent for home activity for individuals in low income and upper-middle income group in Figure 4. The Hellinger distance for the distribution of home activity duration for individuals in the low income group is 0.14, and the JS distance is 0.17. While the mean value is calculated as 18.8 hours in the survey, it is calculated as 19.6 hours in our model. We find the Hellinger distance for the distribution of home activity duration for individuals in upper-middle income group to be 0.11, and the JS distance to be 0.14. The mean value calculated from the survey is 15.6 hours, in our model it is 16.0 hours. These results indicate that the distributions of home activity duration by income groups in the survey and our model have show similar patterns.

Since our output has a much larger sample size than the travel survey, we use densities on the y-axis to make the histograms comparable. Another difference is that the survey population contains mostly individuals having some activities during a day and fewer people with no activity staying home all day. In contrast, our model includes this group as well to cover the entire population. The implication of this is that the bin corresponding to the population who spends 24 hours at home in the synthetic population is much larger and this big bin leads to all other bins having lower densities. This is part of the explanation for the difference in density values for all other bins.

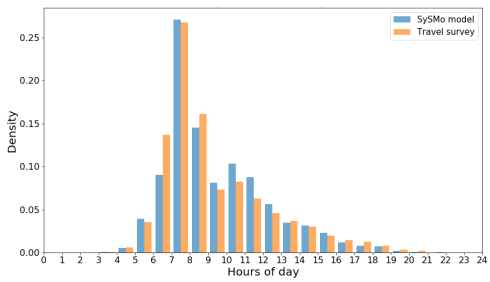


Figure 5: Comparison of 3 AM activity end-time distribution only for those with home activity type.

The distribution of the end time of the 3 AM activity is important since it will help to capture the morning peak in the average weekday travel pattern. The vast majority of the population (about 99 percent) is at home at this time and very few agents attend an out-of-home activity such as work activity. Figure 5 shows the end time distribution of the home activity instances, which take place at 3 AM. The Hellinger distance is 0.07, and the JS distance is 0.09 for these distributions.

These distance values show that the model generates distributions of both activity duration and activity start-end time similar to the distributions from the travel survey data. Even in subgroups by agent attributes or activity features, the distributions show similar characteristics to distributions derived from the surveyed population. The comparisons based on subgroups shows that the correlation between the attributes and activity schedules of individuals is maintained.

5. Results and discussion

The simulated temporal activity pattern for each agent is one of the main outcomes of the proposed methodology. Figure 6 shows the aggregated activity schedules of agents by

type of activity and age group over 24 hours from 00:00 (midnight). The y-axis shows the share of individuals participation in each activity type. Most people are home from 12 AM (00:00) to 6 AM in all income groups. A significant proportion of the population engages in out-of-home activities after 6 AM and we can assume that travel demand increases because of this.

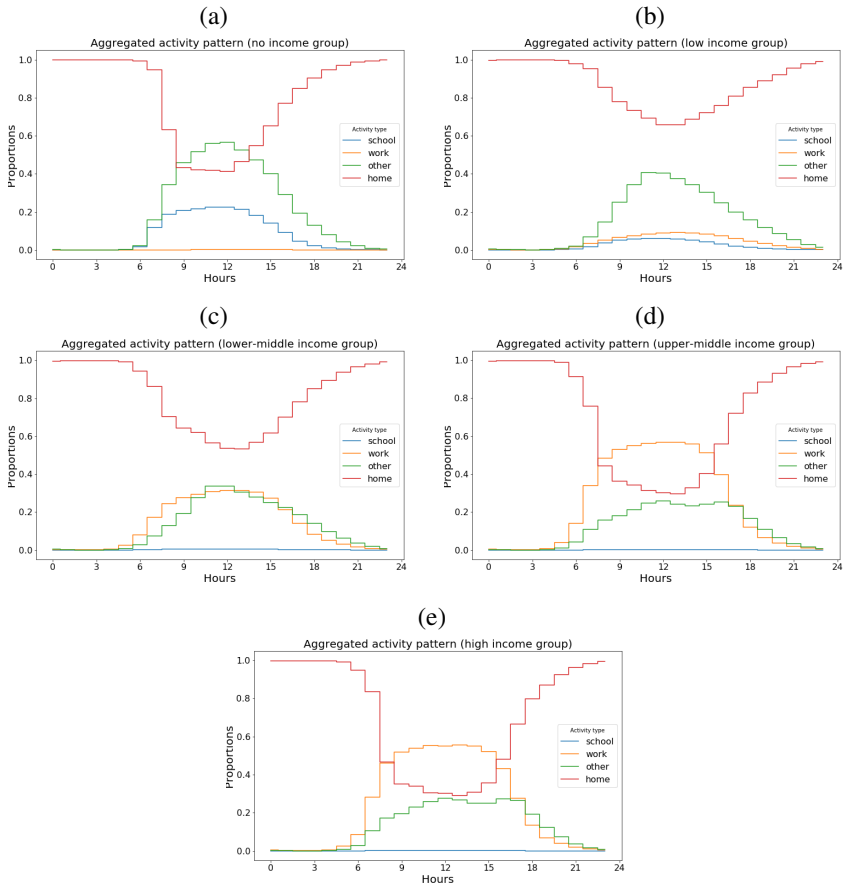


Figure 6: Aggregated activity pattern of the synthetic agents by activity type and income group. (a): no income group (23 percent of the population), (b) low income group (19 percent of the population), (c) lower-middle income group (20 percent of the population), (d) upper-middle income group (19 percent of the population), (e) high income group (19 percent of the population).

Individuals with no-income mostly engage in other and school activities during the day. The largest participation in school activities is observed in this group. Individuals in the low-income group mostly participate in other activities, participation in school or work activities are low. The share of agents participating in the work activity increase with income level and the highest participation is seen in the upper-middle and high-income groups. Since it is possible to participate in more than one activity during an hour, the total number of people engaged in activities may be more than the entire population. In all income groups except no income, the proportion of people participating in the work and other activity types increases at around 6 in the morning, and a peak is generally observed

around noon.

The activity schedule of an agents will depend on its socio-economic characteristics however, due to the heterogeneity introduces in our methodology two agents with the same set of characteristics will not have identical activity schedules. Figure 7 depicts this heterogeneous activity pattern by plotting the activities of a specific set of the population. The set considered contains individuals with an age range of 40-45, male, married, employed, in high-income class, with no children ≤ 6 years old in household, no car in household, and residing in Stockholm. Panel a in the figure shows aggregated activity pattern of the population set with the share of participation in different activity types, showing that at a given hour different activity types are present. Panel b depicts the frequency of the 10 most frequent daily activity sequences in the population. While there is a dominant activity pattern (H-W-H),this is still below half of the agents and other activity sequences are present within the group.

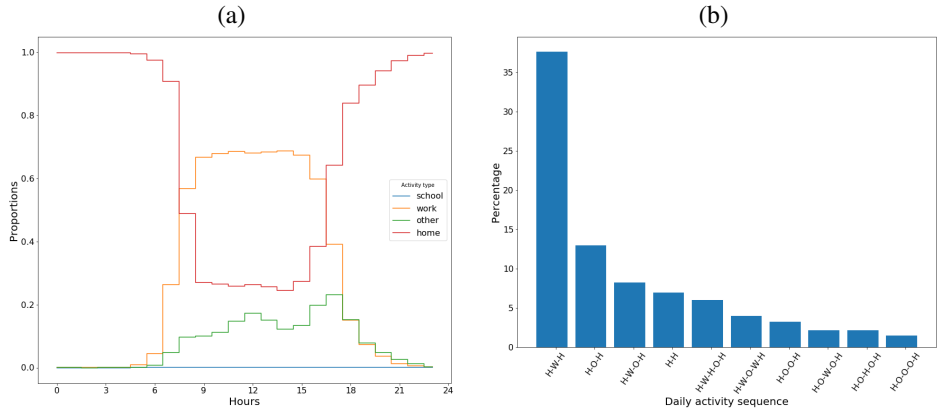


Figure 7: Activity pattern of the synthetic agents; aged 40-45, male, married, employee, in high-income class, no children ≤ 6 years old in household, and no car in household, residing in Stockholm. (a): Aggregated activity pattern of the sub-population by activity type, (b): Percentage of 10 most frequent daily activity sequences in the the sub-population (26 thousand agents in total).

6. DISCUSSION

The daily activity generation module plays a crucial role toward creating a realistic mobility pattern. Most previous studies have created homogeneous activity patterns within sub-populations Arentze & Timmermans (2000); Allahviranloo & Recker (2013); M. Hafezi et al. (2018), meaning that individuals belonging to the same group (i.e., a particular income and age range) have the same activity pattern which is not a very realistic assumption. The models generating homogeneous activity patterns within the population groups fail to capture the behavioral differences that occur within the group, even though these accurately capture many indicators of travel behavior at an aggregated level, such as total distance traveled in the population.

Here we propose a model generating population with a heterogeneous activity pattern. The developed model also maintains the correlation between attributes (e.g. gender, and

income group) and activity schedules of individuals. We think that these contributions to the literature will facilitate more sensitive analysis, and more targeted policy interventions. For instance, if the socio-economic characteristics of population groups with a certain travel pattern are predicted such as groups performing long commuting trips by private car, the effectiveness of policies designed to increase adaptation to new technologies on these groups can be more accurately measured.

We assess the performance and validity of the proposed methodology by performing in-sample evaluations against the travel survey. The results of the comparisons show that the activity schedules generated from the model simulate those of the travel survey reasonably well. In order to see how well the heterogeneity is captured in the proposed model, comparisons are also made in specific sets of the population by joint classes (e.g. by agent attributes like age and gender). We also evaluate the ML techniques used to generate activity schedules in SySMo. The result also shows that ML is an useful tool to predict features regarding the activity schedules of individuals. Ideally one would want to validate the activity schedules at a micro-level with data not used to develop the model, however there is a lack of such data and at this stage we rely only on a comparison with the travel survey. Future work could include comparing with emerging data sources, however even these have their challenges when it comes to validity and representativeness Yuan et al. (2018, 2020).

In the field of transportation modeling, machine learning techniques are increasingly applied and substitute conventional techniques. For example, these are some of the examples of using machine learning to model mode choice (Zhang & Xie (2008); Zhu et al. (2018); Moons et al. (2007)), activity pattern predicting (M. Yang et al. (2014); M. H. Hafezi et al. (2019)), and route choice (Sun & Park (2017)). We employ artificial neural network approach in ML to model the complexity of the activity patterns within a synthetic population since the approach has high predictive capabilities. However, we haven't examined other ML approaches in this research. Convolutional neural networks (CNN) have shown high performance to forecast human travel behavior (Liang & Wang, 2017; Liu et al., 2017) as well as other fields such as image classification or object detection tasks (Ciregan et al., 2012; Erhan et al., 2014). Using more advanced ML methods such as convolutional neural networks to the current methodology could be a future research topic to further improve the result with different approaches.

REFERENCES

Allahviranloo, M., & Recker, W. (2013). Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological*, *58*, 16–43.

Arentze, T., & Timmermans, H. (2000). *ALBATROSS: A learning based transportation oriented simulation system*. EIRASS.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1–74.

Bhat, C., Guo, J., Srinivasan, S., & Sivakumar, A. (2004). Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record*, *1894*(1), 57–66.

Bowman, J., & Ben-Akiva, M. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, *35*(1), 1–28.

Brier, G. W., et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Canella, O., Flötteröd, G., Johnsson, D., Kristoffersson, I., Larek, P., & Thelin, J. (2016). *Flexible coupling of disaggregate travel demand models and network simulation packages (ihop2)* (Tech. Rep.). Technical report, KTH, Sweco, WSP.

Castiglione, J., Bradley, M., & Gliebe, J. (2015). *Activity-based travel demand models: A primer* (No. S2-C46-RR-1). Transportation Research Board.

Čertickỳ, M., Drchal, J., Cuchỳ, M., & Jakob, M. (2015). Fully agent-based simulation model of multimodal mobility in European cities. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (pp. 229–236).

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 ieee conference on computer vision and pattern recognition* (pp. 3642–3649).

Demographic Statistical Areas (DeSO). (2020). https://www.scb.se/en/services/open-data-api/open-geodata/deso--demographic-statistical-areas/.

Dijst, M., & Vidakovic, V. (1997). Individual action space in the city. *Activity-based approaches to travel analysis*.

Doherty, S. T. (2000). An activity scheduling process approach to understanding travel behavior. In *79th Annual Meeting of the Transportation Research Board, Washington, DC*.

Drchal, J., Čertickỳ, M., & Jakob, M. (2019). Data-driven activity scheduler for agent-based mobility models. *Transportation Research Part C: Emerging Technologies*, *98*, 370–390.

Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. (2014). Scalable object detection using deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2147–2154).

Ettema, D., Borgers, A., & Timmermans, H. (1993). Simulation model of activity scheduling behavior. *Transportation Research Record*, 1–1.

Fulton, L. M. (2018). Three revolutions in urban passenger travel. *Joule*, *2*(4), 575–578.

Gärling, T., Kwan, M.-p., & Golledge, R. (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B: Methodological*, *28*(5), 355–364.

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, *40*(2), 44–58.

Hafezi, M., Liu, L., & Millward, H. (2018). Learning daily activity sequences of population groups using random forest theory. *Transportation Research Record*, *2672*(47), 194–207.

Hafezi, M. H., Liu, L., & Millward, H. (2019). A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation*, *46*(4), 1369–1394.

Hutchinson, J. W. (2018). *Exploring patterns of heterogeneity in activity–travel behaviors of older people* (Unpublished doctoral dissertation).

Jones, P., Dix, M., Clarke, M., & Heggie, I. (1983). *Understanding travel behaviour*. Gower Publishing.

Koushik, A., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behaviour research: a review. *Transport reviews*, *40*(3), 288–311.

Lenntorp, B. (1977). Paths in space-time environments: A time-geographic study of movement possibilities of individuals. *Environment and Planning A*, *9*(8), 961–972.

Liang, X., & Wang, G. (2017). A convolutional neural network for transportation mode detection based on smartphone platform. In *2017 ieee 14th international conference on mobile ad hoc and sensor systems (mass)* (pp. 338–342).

Liao, Y., Yeh, S., & Gil, J. (2022). Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, *49*(1), 137–161.

Liu, G., Yin, Z., Jia, Y., & Xie, Y. (2017). Passenger flow estimation based on convolutional neural network in public transportation system. *Knowledge-Based Systems*, *123*, 102–115.

Lum, K., Chungbaek, Y., Eubank, S., & Marathe, M. (2016). A two-stage, fitted values approach to activity matching. *International Journal of Transportation*, *4*(1), 41–56.

Márquez-Fernández, F. J., Bischoff, J., Domingues-Olavarría, G., & Alaküla, M. (2021). Assessment of future ev charging infrastructure scenarios for long-distance transport in sweden. *IEEE Transactions on Transportation Electrification*.

Matyas, M., & Kamargianni, M. (2019). The potential of mobility as a service bundles as a mobility management tool. *Transportation*, *46*(5), 1951–1968.

Miller, E., & Roorda, M. (2003). Prototype model of household activity-travel scheduling. *Transportation Research Record*, *1831*(1), 114–121.

Moons, E., Wets, G., & Aerts, M. (2007). Nonlinear models for determining mode choice. In *Portuguese conference on artificial intelligence* (pp. 183–194).

Pendyala, R., Kitamura, R., Chen, C., & Pas, E. (1997). An activity-based microsimulation analysis of transportation control measures. *Transport Policy*, *4*(3), 183–192.

Quade, P. B. (2000). Douglas, inc.(2000). *Comparative analysis weekday and weekend travel with NPTS integration for the RT-HIS: regional travel-household interview survey. Prepared for the New York Metropolitan Council and the North Jersey Transportation Planning Authority, February*.

Rasouli, S., & Timmermans, H. (2014). Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), 31–60.

Rutherford, G. S., McCormack, E., & Wilkinson, M. (1997). *Travel impacts of urban form: Implications from an analysis of two seattle area travel diaries* (Tech. Rep.).

Schläpfer, M., Dong, L., O'Keeffe, K., Santi, P., Szell, M., Salat, H., ... West, G. B. (2021). The universal visitation law of human mobility. *Nature*, *593*(7860), 522–527.

Shukla, P., Skea, J., Slade, R., Khourdajie, A. A., R. van Diemen, D. M., Pathak, M., ... Malley, J. (2022). *Ipcc, 2022: Climate change 2022: Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change* (Tech. Rep.).

Sun, B., & Park, B. B. (2017). Route choice modeling with support vector machine. *Transportation research procedia*, *25*, 1806–1814.

The swedish national travel survey. (2021). https://www.trafa.se/en/travel-survey/travel-survey/.

Tozluoğlu, Ç., Dhamal, S., Yeh, S., Sprei, F., Marathe, M., Barrett, C., & Dubhashi, D. (2022). Synthetic Sweden Mobility (SySMo) model documentation.

Vovsha, P., & Chiao, K.-A. (2006). Development of New York metropolitan transportation council tour-based model. *Innovations in Travel Demand Modeling*, 21.

Yang, D., Timmermans, H., & Grigolon, A. (2013). Exploring heterogeneity in travel time expenditure of aging populations in the netherlands: results of a chaid analysis. *Journal of Transport Geography*, *33*, 170–179.

Yang, M., Tang, D., Ding, H., Wang, W., Luo, T., & Luo, S. (2014). Evaluating staggered working hours using a multi-agent-based q-learning model. *Transport*, *29*(3), 296–306.

Yuan, Y., Lu, Y., Chow, T. E., Ye, C., Alyaqout, A., & Liu, Y. (2020). The missing parts from social media–enabled smart cities: Who, where, when, and what? *Annals of the American Association of Geographers*, *110*(2), 462–475.

Yuan, Y., Wei, G., & Lu, Y. (2018). Evaluating gender representativeness of location-based social media: A case study of weibo. *Annals of GIS*, *24*(3), 163–176.

Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record*, *2076*(1), 141–150.

Zhu, Z., Chen, X., Xiong, C., & Zhang, L. (2018). A mixed bayesian network for two-dimensional decision modeling of departure time and mode choice. *Transportation*, *45*(5), 1499–1522.