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

July 9, 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.

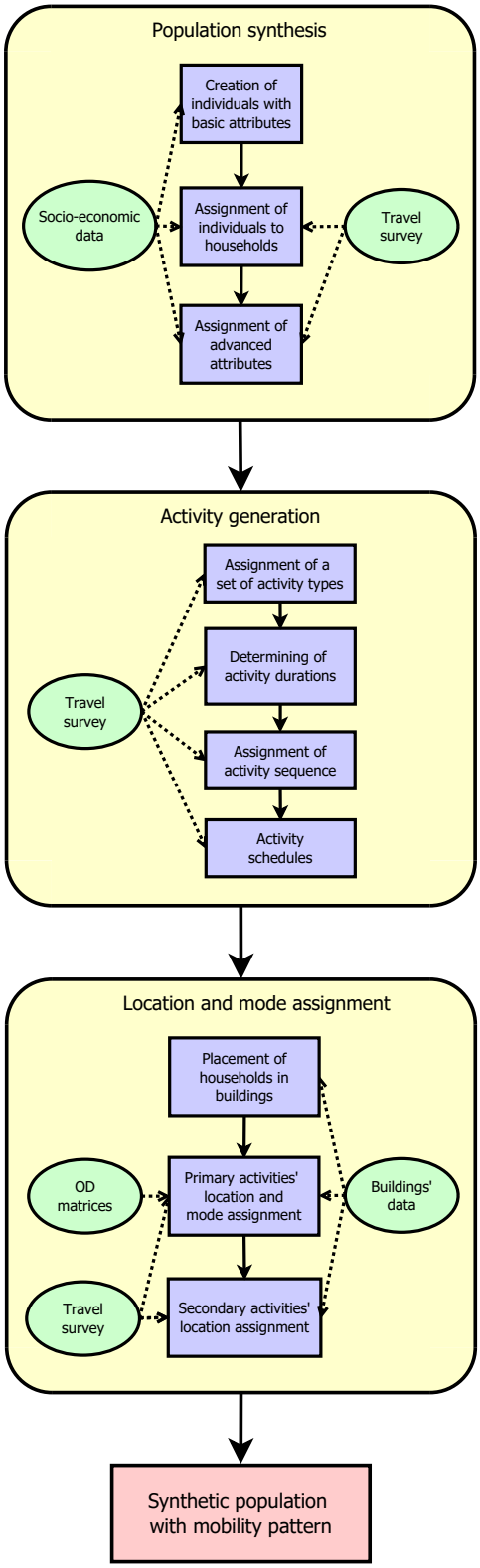


Figure 1.1: Methodology overview of *Synthetic Sweden Mobility (SySMo) Model*. Yellow rectangles: three main components of SySMo model; blue rectangles: procedures of the calculations; green ellipses: input data for modeling and calibration; pink rectangle: the final outputs, a spatially explicit agent-based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

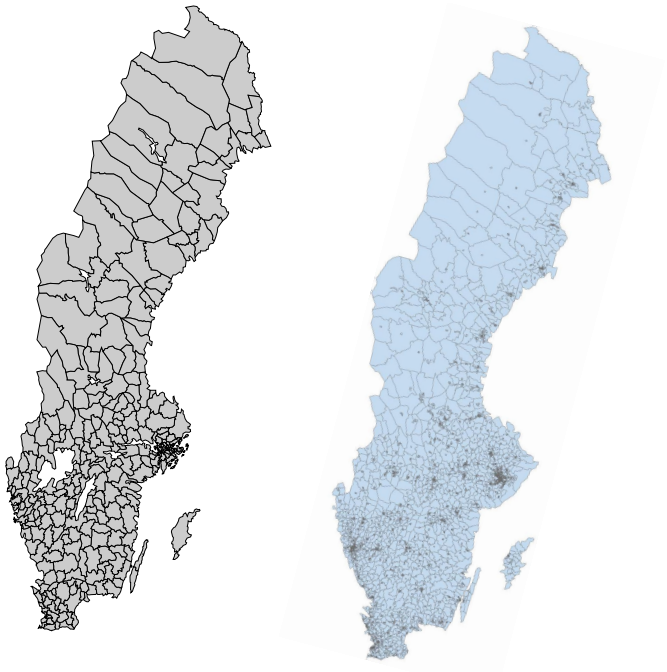
Statistics Sweden (SCB) [2] produces the official statistics at various geographical levels such as municipality or zone system. Fig. 2.1a shows the boundaries of 290 municipalities which act as local government entities. The statistical data at municipal areas are the number of individuals with a given 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.

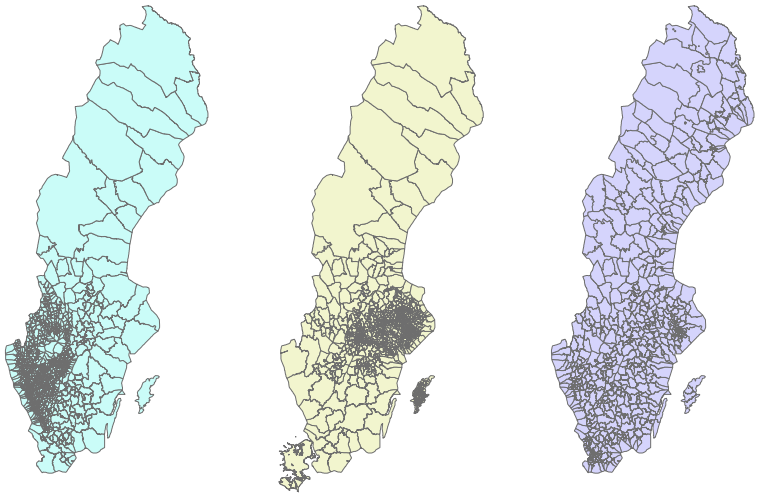
Each participant has one weight V_k according to their socio-demographics and another weight V_d based on the day the participant conducted the survey. These weights directly indicate the representative power of the respondent regarding socio-demographics or travel patterns. The total population can 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 $A = \{0, 1\text{-}6, 7\text{-}15, 16\text{-}18, 19\text{-}24, 25\text{-}29, 30\text{-}34, 35\text{-}44, 45\text{-}54, 55\text{-}64, 65\text{-}75, 75\text{-}84, 85+\}$; $G = \{\text{'male'}, \text{'female'}\}$; $C = \{\text{'single'}, \text{'couple'}, \text{'child'}\}$. 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.

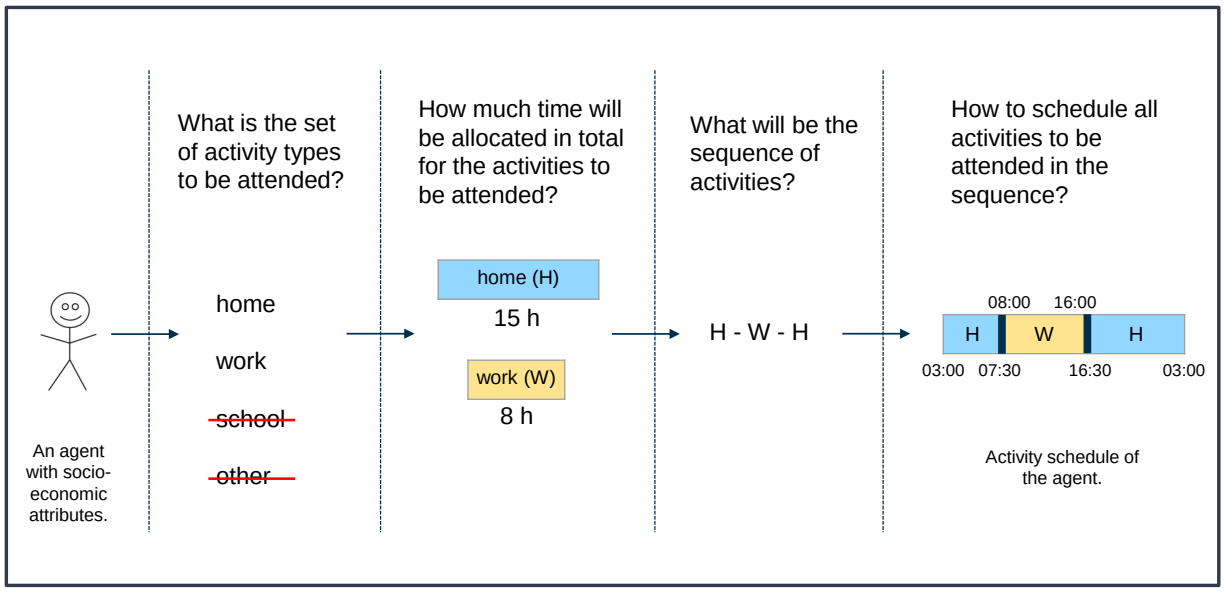


Figure 4.1: The main steps of the activity generation component. Each step in the activity generation component is represented divisions drawn by vertical dashed lines. Activity schedules are generated for agents in the synthetic population.

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.

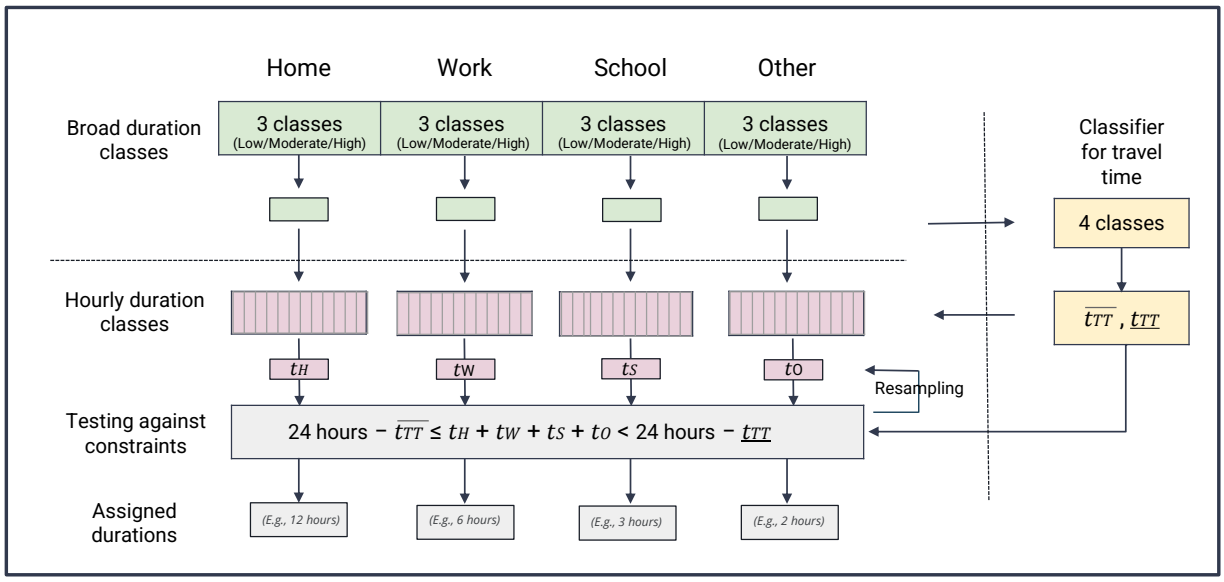


Figure 4.2: The flow chart of activity duration assignment methodology in SySMo. Green rectangles: joint model for broad activity duration, yellow rectangles: model for travel time, pink rectangles: model for hourly activity duration, and gray rectangles: final activity duration satisfying the constraint.

- 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, $-W\text{-}W-$ 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.

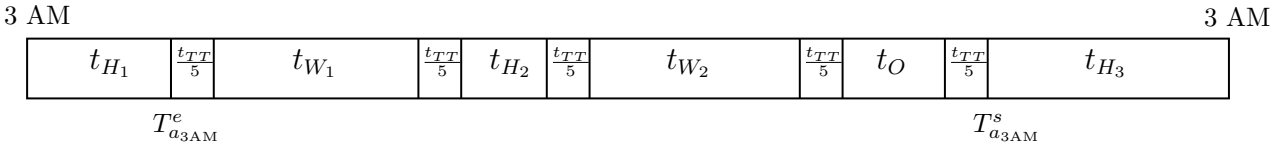


Figure 4.3: Activity schedule of an agent with activity sequence is *H*-*W*-*H*-*W*-*O*-*H*. The daily travel $t_{TT} = 24 - (t_H + t_W + t_S + t_O)$ **.**

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.

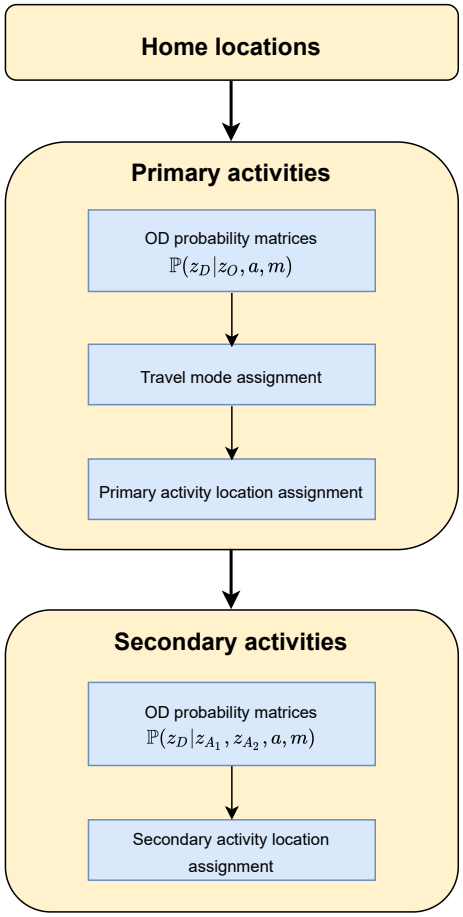


Figure 5.1: A flow chart of activity, mode and location assignment Yellow rectangles: major steps of the activity location assignment methodology; blue rectangles: sub-steps within the main steps.

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.

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

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.

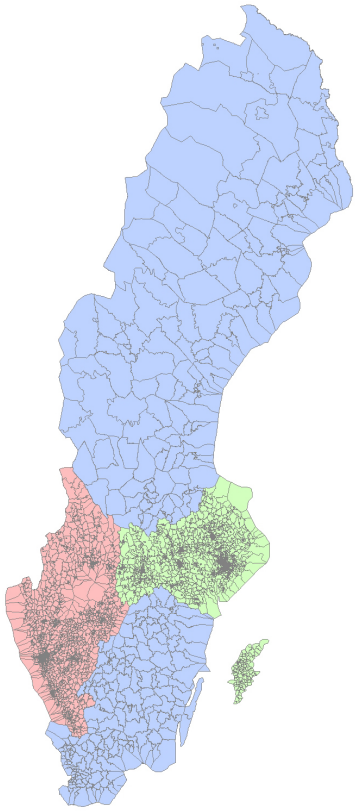


Figure 5.2: The zone system used in SySMo. Pink: zones according to Väst regional model, green: zones according to Samm regional model, and blue: zones according to the national model.

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:

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

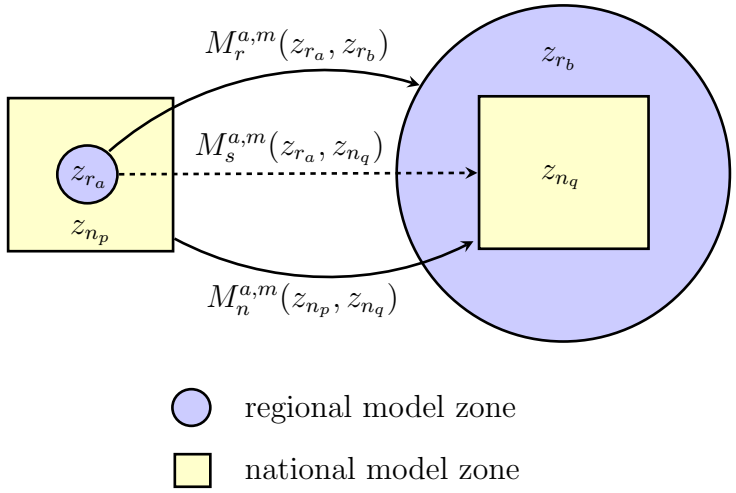


Figure 5.3: An abstract illustration of regional and national model zones, and OD matrices' values to be used for IPF (arrows point from origin to destination; solid arrow means that the value is available from Sampers OD matrices; dotted arrow means that the value is to be deduced)

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^a_m. 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 $s^a_{z_d}$ denote the *attraction potential* of zone z_d for activity type a. $s^a_{z_d}$ 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:

$$\mathbb{P}(z_d|z_o, a, m) \propto s^a_{z_d} e^{\beta^a_m d(z_o, z_d)} \tag{5.7}$$

$$= \frac{s^a_{z_d} e^{\beta^a_m d(z_o, z_d)}}{\sum_z s^a_z e^{\beta^a_m d(z_o, z_d)}} \tag{5.8}$$

Table 5.3 presents the calibrated values of the parameters. A more negative value of parameter β^a_m means that the probability drops rapidly with an increase in distance. We see that across all modes, the values of parameter β^a_m 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^a_m 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

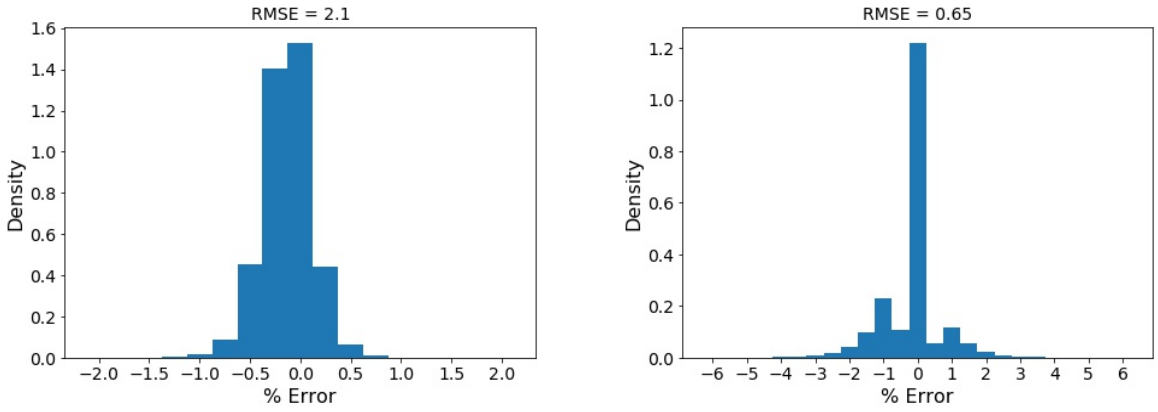
Table 6.1: Performance assessments

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).

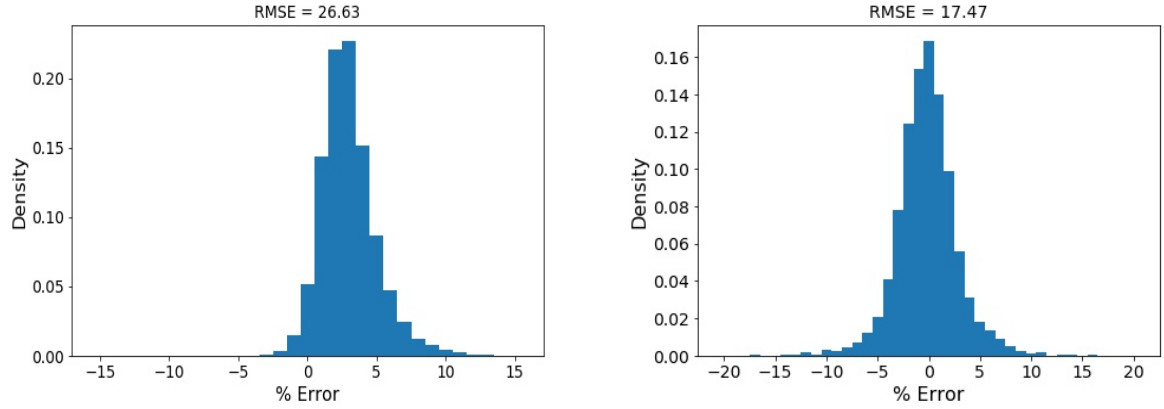
For age (Fig. 6.1), the error is between -1% and 1% in more than 78 percent of the DeSO zones. We found the RMSE to be 0.65. This indicates that 0.65 people in each Deso zone may have been assigned an incorrect age group. [9]

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.

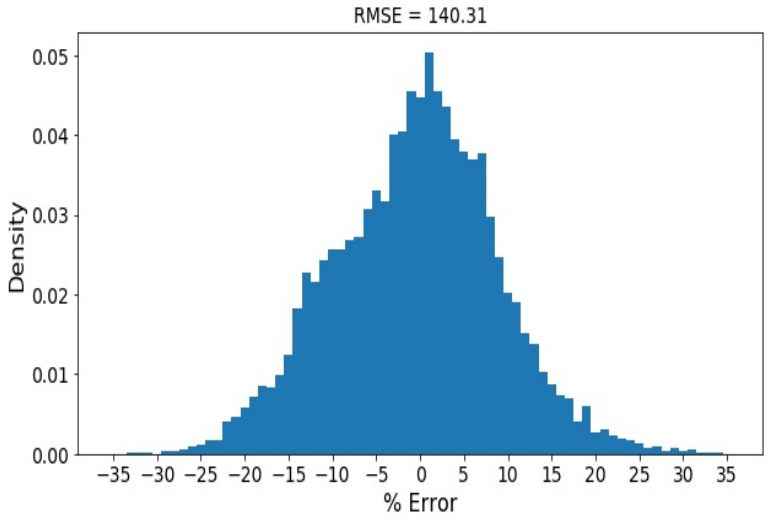


Figure 6.3: The percent error in the number of individuals by gender and age.

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

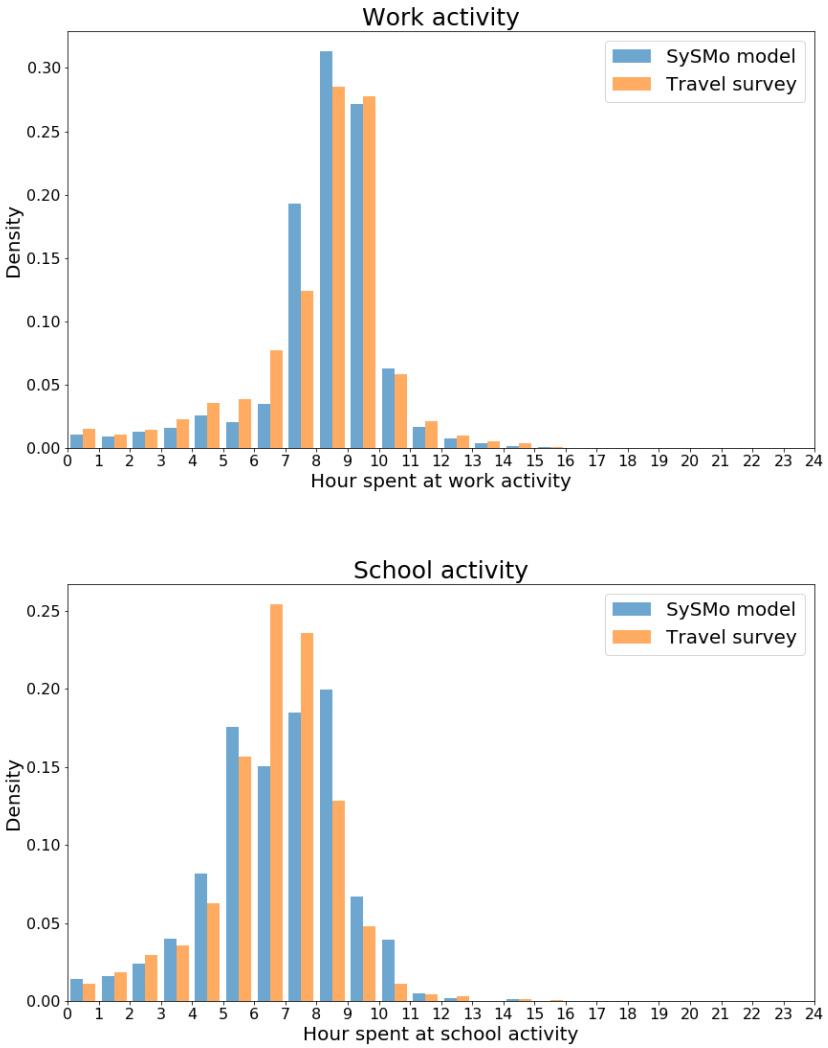


Figure 6.4: Comparison of activity duration by activity type.

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.

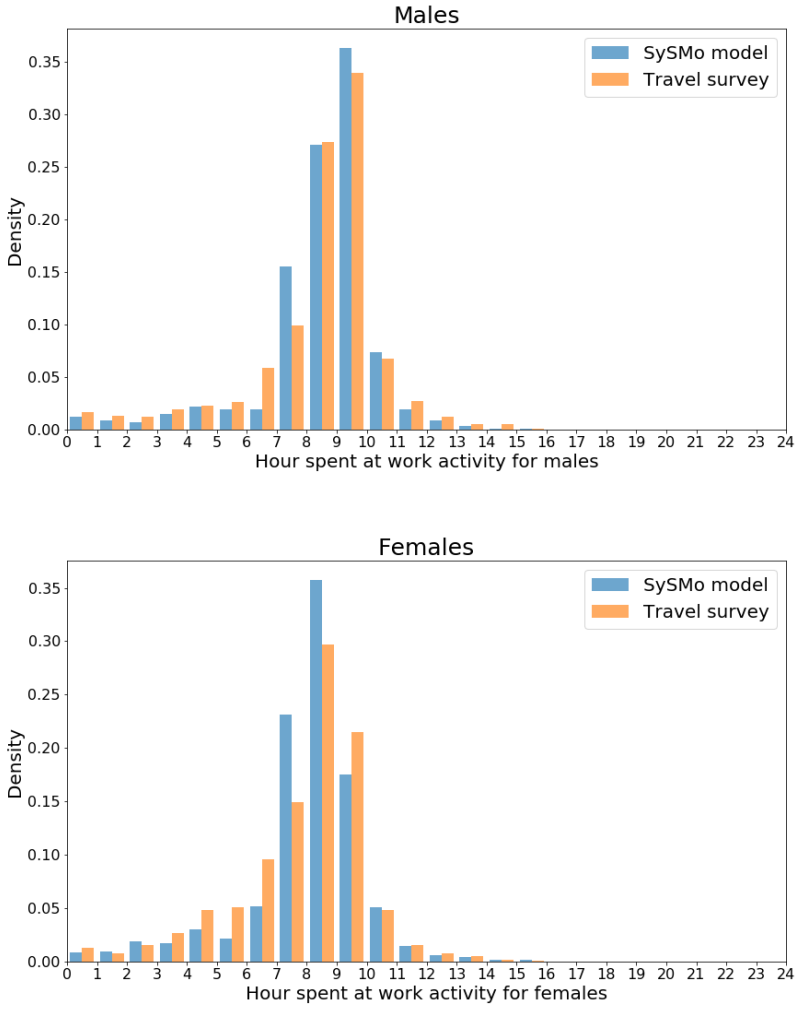


Figure 6.5: Comparison of activity duration by activity type and gender.

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

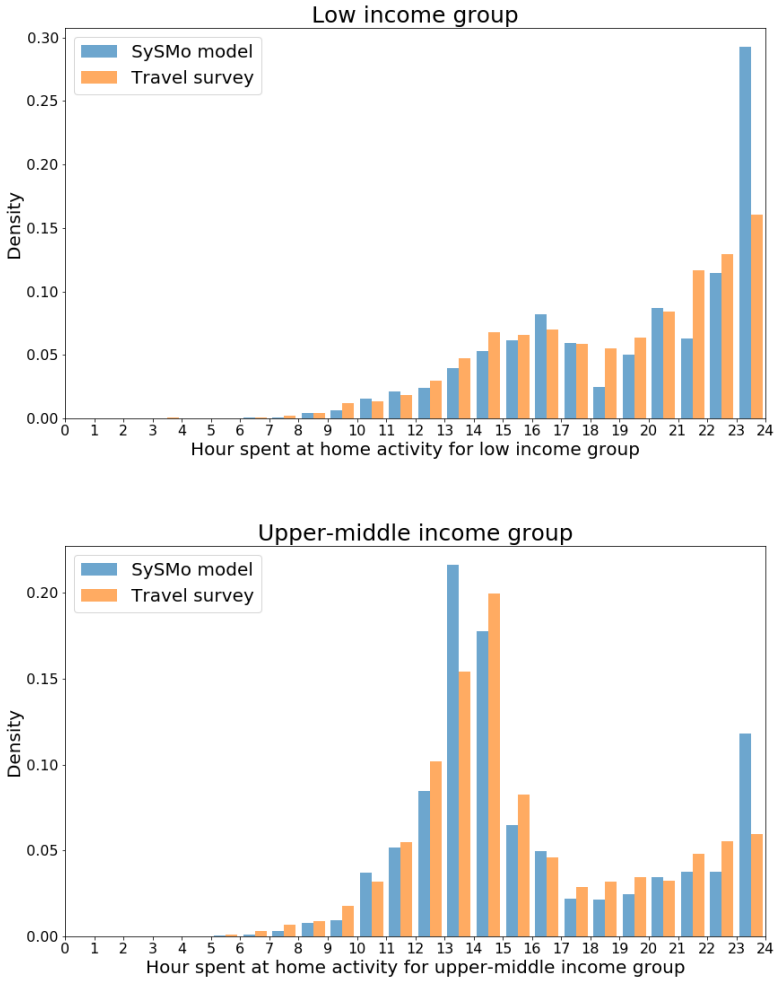


Figure 6.6: Comparison of activity duration by activity type and income group.

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,

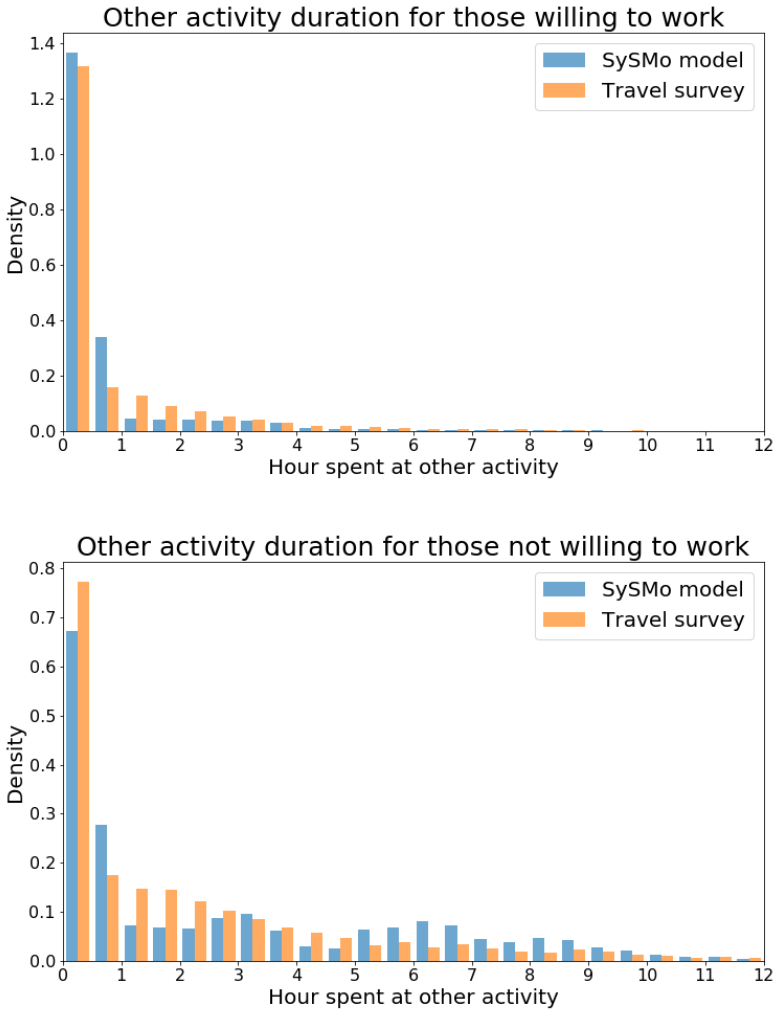


Figure 6.7: Comparison of activity duration by activity type and activity participation.

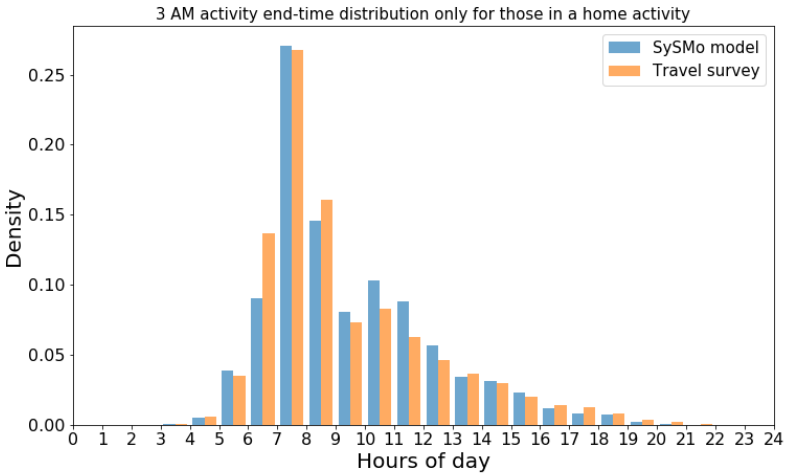


Figure 6.8: Comparison of activity end time distribution by activity type.

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.

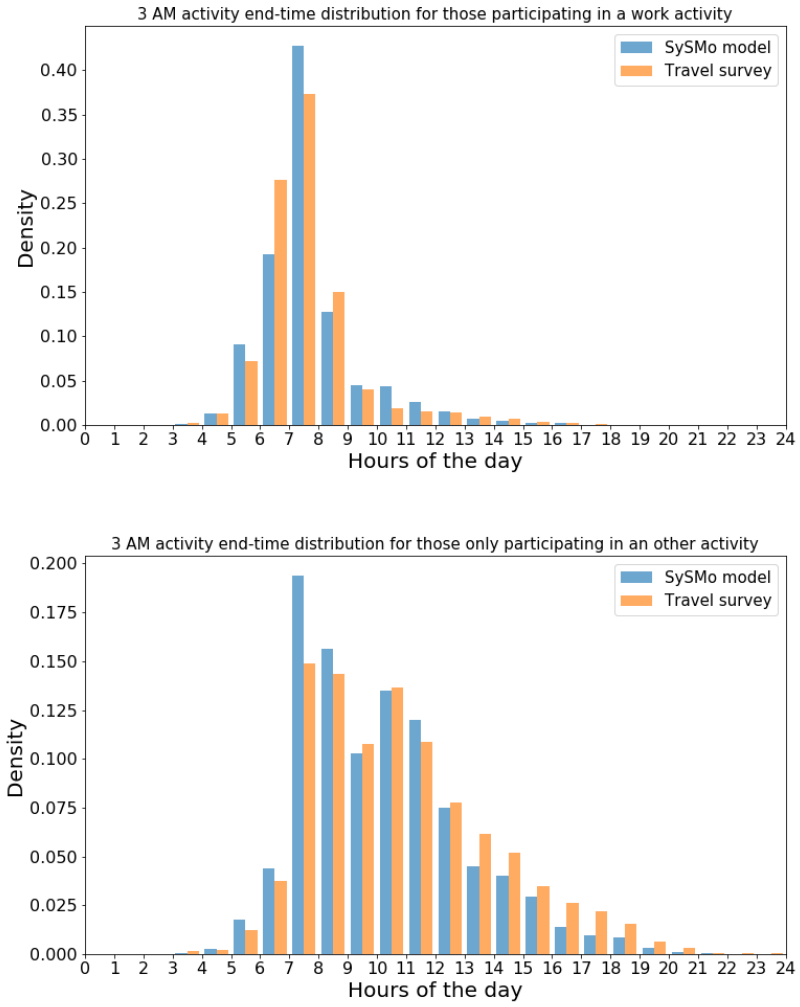


Figure 6.9: Comparison of activity end time distribution by activity type and activity participation.

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

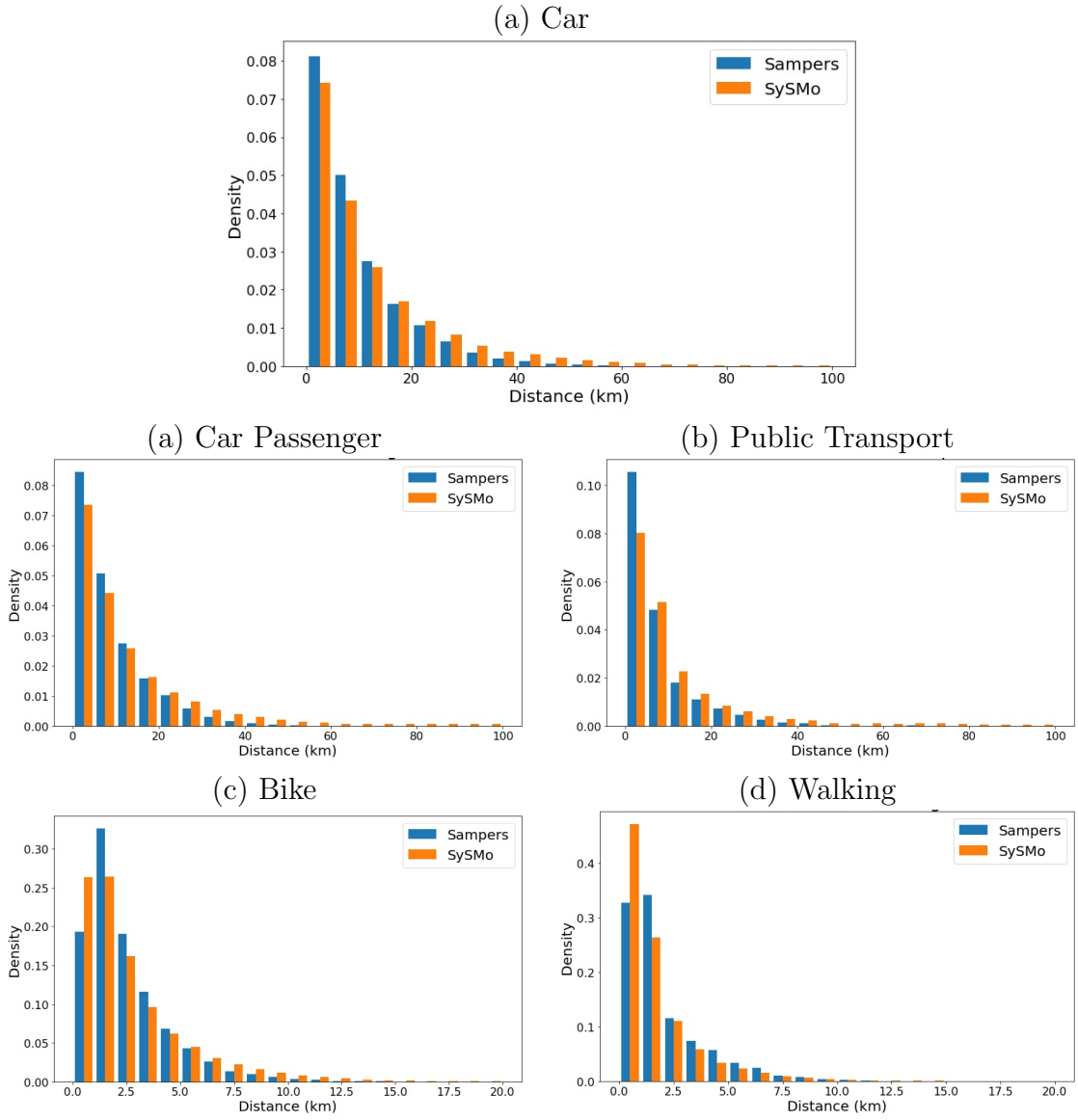


Figure 6.10: Comparison of daily travel distance of individuals between home and work by travel modes.

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.