Heterogeneous Activity Generation for a Synthetic Population: Synthetic Sweden Mobility (SySMo) Model

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

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

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

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

SHORT SUMMARY

The heterogeneous diversity of activity patterns within population groups has often been disregarded in most literature, resulting an unrealistic representation of travel behavior. We develop a stochastic approach combined with machine learning (ML) to generate heterogeneous activity in a synthetic population. We implement the novel methodology to model the mobility pattern of the synthetic population of Sweden. Comparing the simulated activity schedules with survey data, our results show that the methodology generates realistic mobility pattern of individuals. The proposed model with a realistic representation of activities can make an unique contribution to capture the complexity of travel behaviors, thus better inform policies to improve future transportation systems.

Keywords: Heterogeneous activity generation; Agent-based modeling; Activity-based modeling; Big data analytics; Machine learning.

1. INTRODUCTION

Urbanization, increasing population, and unsustainable development of the current transportation system make innovations necessary. Micro-mobility, electric cars, and autonomous cars are some examples of innovations that can bring transformative changes to the system and potentially change people's travel behaviors. While assessing these potential changes, decision-makers should be supported by models that are capable of reflecting realistic dynamics of mobility and interactions with key factors that affect travel decisions.

Activity-based modeling is a commonly used travel demand modeling approach that has gained interests in the last decade (Hafezi et al., 2018). The emerging big data sources and the growth of computer processing power have enabled the faster development of activity-based models toward integrating sub-models, capturing the dependencies between the trip chains, higher temporal and spatial resolution, and behavioral realism (Rasouli & Timmermans, 2014). Agent-based model (ABM) of travel demand and activity-based models are often combined (Castiglione et al., 2015). The workflow of most agent-based modeling approaches comprises population synthesis, activity generation, and execution of activities, as shown in Figure 1.

Çağlar Tozluoğlu and Swapnil Dhamal contributed equally to this research.

Population synthesis
Activity generation
Execution of activities

Figure 1: An overview of a typical agent-based modeling workflow. The box with the solid line is the focus of this paper. Other boxes are described in more detail in (Dhamal, Tozluoğlu, et al., 2022).

The activity generation step, the second component in the workflow, assigns daily activity plans to the agents. Previous studies have mainly homogeneous daily activity patterns for each sub-population or group (for instance, a particular income and age range), resulting in the same activity pattern for all individuals within the same group e.g., Miller & Roorda (2003); Arentze & Timmermans (2000); Allahviranloo & Recker (2013); Hafezi et al. (2018). They are insufficient in reflecting the heterogeneity of the population. Although a few studies have focused on specific regions or population groups, there is no study in the current literature that addresses all of Sweden's population and their mobility patterns.

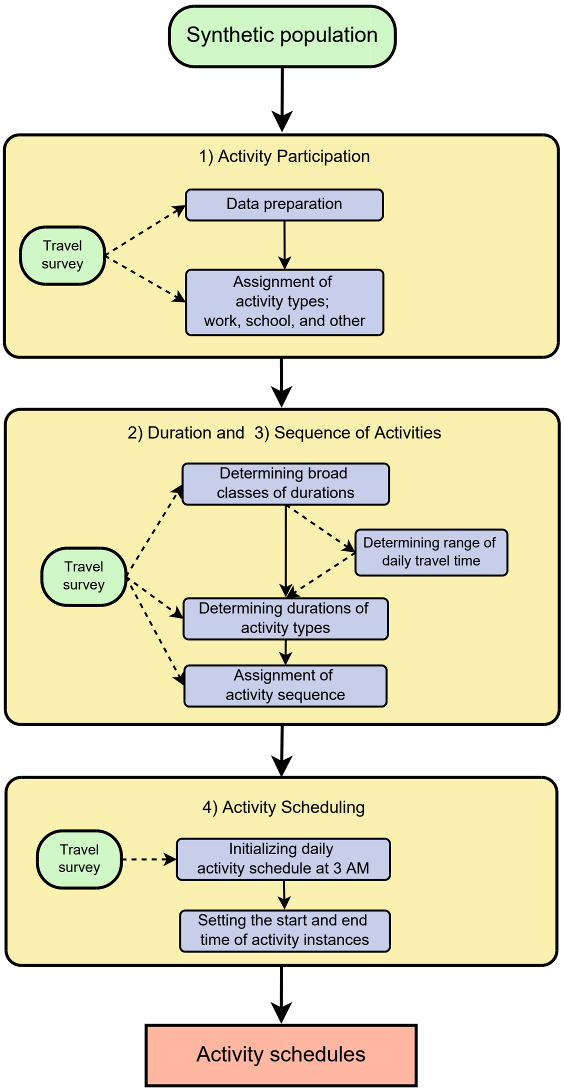
In this paper, we propose a novel methodology for generating daily schedules of agents in a synthetic population. Using ML in conjunction with probability models, we are able to maintain heterogeneity by sampling from the derived probability distributions. The proposed methodology generates more realistic activity schedules as it captures the heterogeneity in activity generation among individuals. This methodology can enable more flexible and precise studies of people's mobility. Neural networks have been chosen among various ML techniques because they have high predictive capabilities on complex data sets (Gunning & Aha, 2019). This paper is part of a large-scale project, called Synthetic Sweden Mobility (SySMo) Model (Dhamal, Tozluoğlu, et al., 2022) that models the mobility patterns of the population in Sweden. We apply our methodology on a synthetic population of Sweden, and validate the obtained activity schedules.

2. METHODOLOGY

The activity generation framework comprises four major steps: (1) assignment of a set of activity types, (2) determination of the duration of each activity type, (3) assignment of the sequence of activities, and (4) creation of activity schedules for each individual (Figure 2). This article contains a brief summary of the methodology of the activity generation module of SySMo model; interested readers can refer to the SySMo model documentation (Dhamal, Tozluoğlu, et al., 2022) for more details. Since the travel patterns on weekdays and weekends are significantly different, we model daily travel patterns corresponding to an average weekday and an average weekend.

Activity Participation

For each agent in the synthetic population, we assign a set of activity types that these agents have the potential to be involved in. Four types of activities are considered: *home* (H), *work* (W), *school* (S), and *other* (O) like visiting shops, restaurants, etc. We use a neural network classifier (NNC) to model an individual's willingness to participate in each activity type given the agent's socio-economic attributes. Our model assumes that each individual visits the home location at least once a day. The Swedish national travel survey (*The Swedish national travel survey*, 2021) is used for training the classifier. A total of $2^3 = 8$ classes are considered since each of activity type could be either 0 or 1 except home activity. We develop four models depending on the employment status



Synthetic population
1) Activity Participation
Travel survey
Data preparation
Assignment of activity types; work, school, and other
2) Duration and 3) Sequence of Activities
Determining broad classes of durations
Determining range of daily travel time
Travel survey
Determining durations of activity types
Assignment of activity sequence
4) Activity Scheduling
Travel survey
Initializing daily activity schedule at 3 AM
Setting the start and end time of activity instances
Activity schedules

Figure 2: Methodology overview of the activity generation module of *Synthetic Sweden Mobility (SySMo) model.* *Yellow rectangles: major steps of the activity generation; purple rectangles: steps of the calculations; green ellipses: input data; pink rectangle: model outputs of activity schedules for each individual.*

(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.

Let the variable activity type be A, where $A \in \{H,W,S,O\}$. The willingness for activity type A is denoted by θ_A. The model outputs $\mathbb{P}_i(\theta_W = x, \theta_S = y, \theta_O = z)$ denote 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. Thus, every agent is assigned its willingness for work, study, and 'other' activities.

Activity Duration

The method proposed here replicates people's heterogeneity in the population by allowing agents with similar attributes to have different activity durations. The durations of different activity types are determined using a two-step method. In the first step, we jointly deduce broad duration classes for the different activity types; this enables us to capture the correlation between the durations of the different activity types. Using these broad classes and attributes of individuals, we deduce the overall travel time in a day. In the second step, using the deduced broad classes of durations of all the activity types and the range of daily travel time, we derive precise durations of all the activity types for each agent. The rationale for the two-step method is that if we deduce hourly duration classes for the 4 different activity types while preserving correlations directly (without the two-step method), the number of potential joint classes would be $\binom{24}{4} = 10{,}626$. This is an exceedingly high number of classes.

In the first step, broad duration classes classify an individual's total activity time for different activities as low, moderate or high. Since we have 3 broad classes for each of the 4 activity types, the total number of joint classes is $3^4 = 81$. Classifiers are trained using socio-economic attributes and employment/studenthood statuses to deduce the joint broad duration class for an agent. We consider different classifiers for different sets of activity participation. 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. The broad duration classes of the activity types are determined using multinomial sampling. To deduce the precise activity durations for an agent, we estimate the range of daily total travel time for that agent using NNC. 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 $\overline{t_{TT}}$.

In the second step, we determine the precise durations of the different types of activities. The sum of the precise durations 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,

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

We model the hourly duration of a given activity type for an agent using NNC. Then, we sample the precise durations of all types of activities such that they collectively satisfy the constraint in Equation (1). If the constraint is not satisfied, the activity durations are resampled.

Activity Sequencing

We assume that individuals with similar socio-economic attributes and activity type durations, would have similar activity sequences. We choose a set of candidate individuals from the travel survey, considering individuals having the same set of willingness for the activity types and having as many similar socio-economic attributes as possible. We use the approach of having daily activity durations as proxy parameters to find the most similar individual among candidates. For a synthetic agent, we choose the individual using the Euclidean distance between their activity durations' tuples (t_H, t_W, t_S, t_O). Then, the sequence is directly copied from the individual chosen to the agent.

Activity Scheduling

To generate the activity schedule for each agent in the synthetic population, we assume the day starts and ends at 3 AM. Modeling the start and end times of the 3 AM activity will facilitate to arrange remaining activities during a day using activity sequences and durations, as the head and tail of the sequence would be defined. We first deduce hourly distributions of start and end times of the 3 AM activities, using NNC trained using the travel survey. For the sampling process, we impose a certain constraint:

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

where $t_{A_{\mathrm{3AM}}}$. denote the total duration of the 3 AM activity type $t_{a_{\mathrm{3AM}}}$ is 3 AM activity instance. f_{3AM} deduced using neural network regression, represents that the fraction of the total duration of the 3 AM activity type to the duration of the 3 AM activity instance ($\frac{t_{a_{\mathrm{3AM}}}}{t_{A_{\mathrm{3AM}}}}$). We sample the start and end times of the 3 AM activity instance from their corresponding hourly distributions that we deduced earlier such that they satisfy the constraint in Equation (2).

To find start and end times of activity instances that are not the 3 AM activity (remaining activities in the sequence), we distribute activities total duration equally among its instances in the sequence,i.e. if an individual goes to work, than other, and then back to work, the two work durations will be of equal length. After computing the daily activity schedules of all the agent, we then distribute the travel times between adjacent activity instances.

3. RESULTS AND DISCUSSION

The proposed activity generation modeling framework is applied to the synthetic agents created within the SySMo model. We aim to reproduce heterogeneous daily activity schedules including activity type, start-end time, duration, and sequence for the synthetic population. Below we first show the validation against the Swedish national travel survey, then we illustrate selected results of aggregated activity pattern of agents by activity type.

Validation

The validation consists of two types: replication validation and in-sample validation. Replication validation reflects how well the ML techniques can perform if it is used again with a new data set. In in-sample validation, studies are conducted showing the similarity of the results with the input data used to construct the model.

We first conduct replication validation by validating the probabilistic ML models using the travel survey as ground truth data. Brier Score (BS) is one of the metrics frequently used to measure the accuracy of probabilistic predictions (Brier et al., 1950). However, the results produced by the Brier Score can be very difficult to interpret when the classes are imbalanced. We use stratified cross validation method with the Brier skill score (*BSS*) (For an in-depth explanation, see (Dhamal, Tozluoğlu, et al., 2022)). BSS gives a score by comparing the BS of the model with BS of a reference measure. The most common formulation of BSS is

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

BSS gives a value between $-\infty$ and 1 by comparing the Brier score with a reference measure such as a naive model having the constant probability distribution, that shows densities of classes in the data set, in each instance in the data set. 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 repeated the calculation for all four models that are used to generate the activity pattern.

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

Table 1: 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

We also compare the results produced by the SySMo model with the travel survey, i.e. in-sample validation. The comparisons of subgroups of the population by agent attributes and activity features are also performed. Below we plot the density histograms of activity durations (Figure 3) by the joint classes (e.g. by activity type and gender), and start-end time for a selected activity (Figure 4) by a single class (e.g. by activity type). For each plot, we calculate the Hellinger distance and Jensen–Shannon (JS) distances to quantify how similar the distributions are. The distances have values in the range [0,1], where 1 means the maximum distance (see (Dhamal, Tozluoğlu, et al., 2022) for more details). For each instance, when a positive probability is assigned to each class with a probability equal to 0 and vice versa, the score takes the maximum distance 1. E.g., $\mathbb{P}_i(\theta_1 = x, \theta_2 = y, \theta_3 = z)$ denote the probability that an instance i's being the θ_1 is x, being the θ_2 is y, and being the θ_3 is z, where $x, y, z \in \{0, 1\}$. If $\mathbb{P}_j = (1, 0, 0)$ and $\mathbb{P}_k = (0, 0.6, 0.4)$, the distance between j and k instances will be 1.

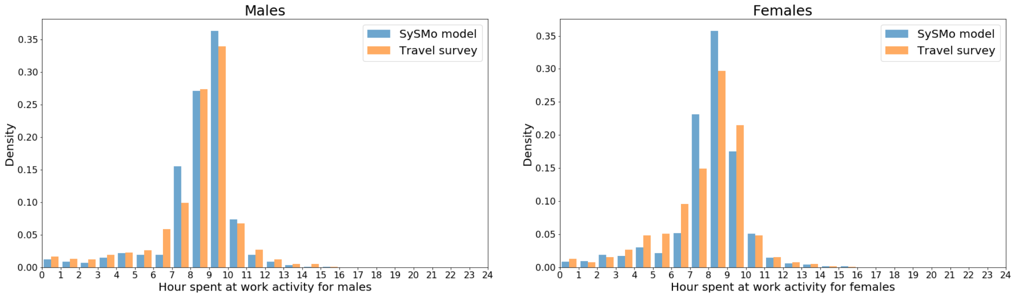


Figure 3: Comparison of work activity duration by gender. The left panel shows hours spent at work activity for males and the right panel shows hours spent at work activity for females.

Figure 3 shows that the Hellinger distance between work activity duration distributions of males is 0.1058, and JS distance is 0.1260. We found the Hellinger distance between work activity duration distributions of females is 0.1245, and JS distance is 0.149.

Figure 4 shows the end time distribution of 3 AM activity only for those with home activity type. The Hellinger distance is 0.0732, and JS distance is 0.0876 for these distributions. These distance values show that the model generated distributions of both work activity duration and activity start-end time have quite similar shapes as the distributions from the travel survey data. Even in subgroups by agent attributes or activity features, the distributions show similar character to distributions derived from the surveyed population. The comparisons made according to subgroups reveal that the correlation between attributes and activity schedules of individuals is maintained.

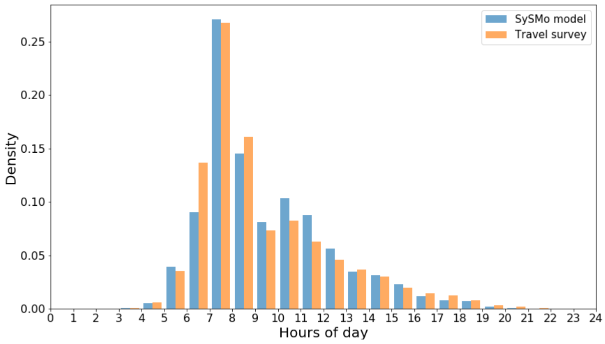
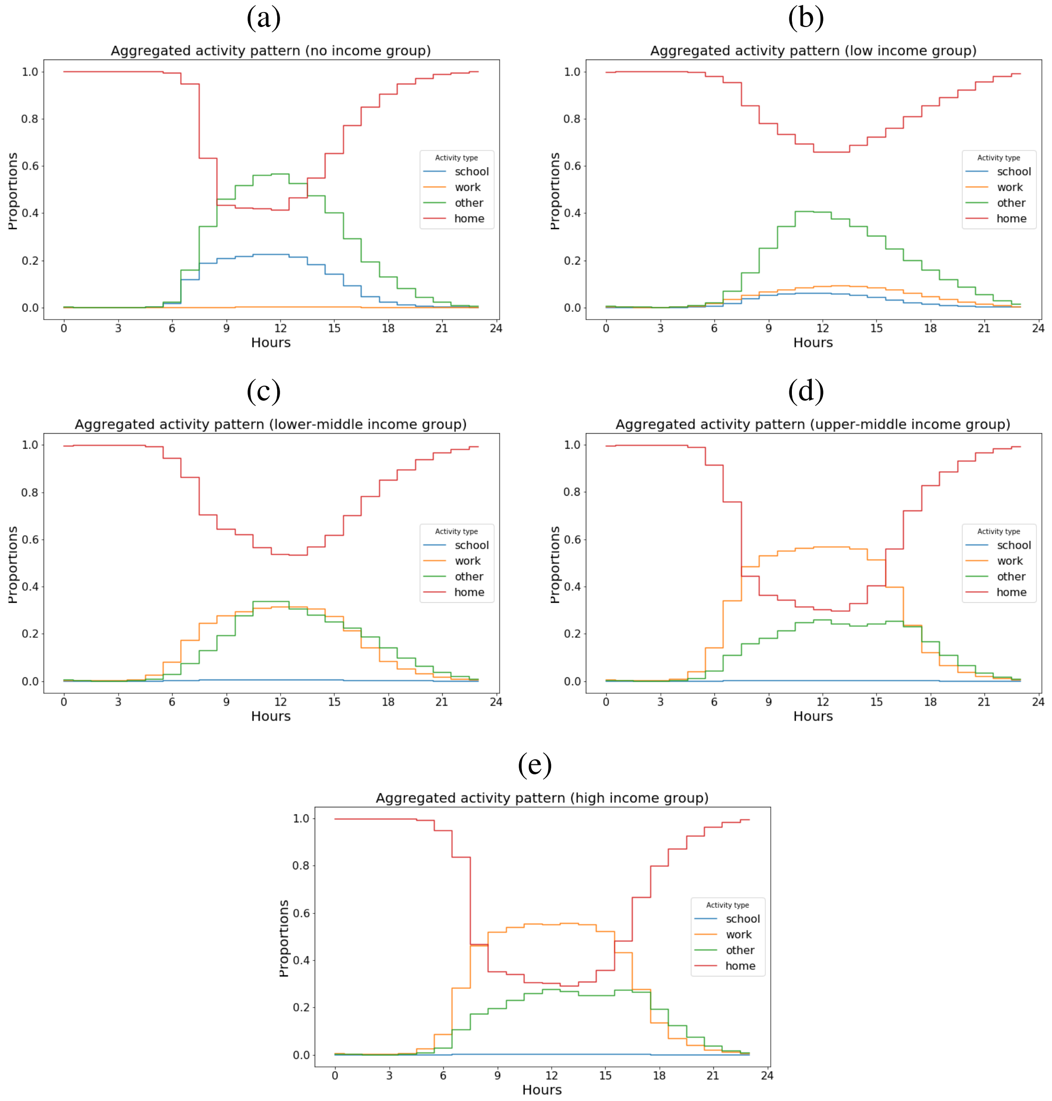


Figure 4: Comparison of 3 AM activity end-time distribution only for those with home activity type.

Results

The simulated temporal pattern of activities for each agent is one of the main outcomes of the SySMo model. Figure 5 shows the aggregated activity schedules of agents by type of activity and age group over 24 hours from 00:00 (midnight) to the following day. The y-axis shows the proportion of individuals' participation in each activity type by time of day. By assumption most people are at home from 12 AM (00:00) to 6 AM in all income groups. A significant proportion of the population engages in out-of-home activities after 6 AM and we can assume that travel demand increases in parallel with this.

Individuals in the no-income group mostly prefer to engage in the other activity type and school activities during day times. The most participation in school activities is observed in this group. Individuals in the low-income group mostly participate in the other activity type, participation in school or work activities remain at a low level. The proportion of participating in the work activity type increase with income level and the highest participation is seen in the upper-middle and high-income groups. Since it is possible to do more than one activity in an hour, the total number of people engaged in activities may be more than the total of the population.


(a)

(b)

(c)

(d)

(e)

Figure 5: Aggregated activity pattern of the synthetic agents by activity type and income group. (a): no income group (23 percent of the population), (b) low income group (19 percent of the population), (c) lower-middle income group (20 percent of the population), (d) lower-middle income group (19 percent of the population), (e) high income group (19 percent of the population).

4. CONCLUSIONS

The daily activity generation module plays a crucial role toward creating a realistic mobility pattern. Most previous studies have created homogeneous activity patterns for each sub-population. Here we propose a model that preserves the complex patterns of generating daily activity schedules, while maintaining the correlation between attributes (e.g. gender, and income group) and activity schedules of individuals.

We conduct two types of validation: replication validation and in-sample validation. The results of the comparisons show that the activity schedules generated from the model replicate the Swedish population reasonably well. The result shows that ML is an useful tool to predict travel behavior of individuals.

In this paper, artificial neural network approach in ML is used to model the complexity of the activity patterns within a synthetic population. We haven't examined other ML approaches in this research. Using more advanced ML methods such as convolutional neural networks to the current methodology could be a future research topic to further improve the result with different approaches.

REFERENCES

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

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

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

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

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

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

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

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

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

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