Transportation Research Part C 123 (2021) 102972

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Ensemble learning activity scheduler for activity based travel demand models

Mohammad Hesam Hafezi [a,*], Naznin Sultana Daisy [a], Hugh Millward [b], Lei Liu [a]

[a] Department of Civil and Resource Engineering, Faculty of Engineering, Dalhousie University, 1360 Barrington Street, Halifax, NS B3H4R2, Canada
[b] Department of Geography and Environmental Studies, School of the Environment, Saint Mary's University, 923 Robie Street, Halifax, NS B3H3C3, Canada

ARTICLE INFO

Keywords:
Activity scheduling
Machine learning
Population cohorts
Activity-based models
Travel demand

ABSTRACT

The aim of this paper is to predict travel behavior for a set of model individuals, who represent cohorts with homogeneous time-use activity patterns. The paper presents a new modeling framework that is able to simulate temporal information associated with the traveler's daily activity schedule, for use in activity-based travel demand modeling. We employed a precise and efficient machine learning technique known as Random-Forest for temporal attribute recognition. In order to capture the uncertainty of start time and activity duration, initially we derived unique clusters of homogeneous daily activity patterns from the activity data. The Random-Forest model is formulated based on the socio-demographic characteristics of travelers and temporal features of their activities. Start time and activity duration for every activity type were allocated to a set of bins. In this study, eight different bin structures, varying in the time interval, are designed as response variables. Using a heuristic decision rule-based algorithm, the predicted activities were inserted into the traveler's skeleton schedule. An algorithm was then applied to schedule travelers' activities based on activity importance level and guide information gained from representative patterns for homogeneous population cohorts.

The model was tested using time-diary data drawn from the Space-Time Activity Research (STAR) survey for Halifax, Nova Scotia. Results show that the proposed model is able to assemble the traveler's schedule with an average 81.62% accuracy in the 24-hour period. The insights gained from this study include important temporal information on activities crucial for the scheduling stage of an activity-based model. Finally, the results of this paper are expected to be implemented within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT).

1. Introduction

Increasingly, transport planners are developing and employing disaggregated approaches to the modeling of transport demands, based on the simulated behavior of individual travelers. Superseding earlier four stage models, second generation models known as disaggregate trip-based models and third generation models known as activity-based travel demand models have been developed (Goran, 2001; Bhat et al., 1894). There are close associations between trips and the activity participation of individuals; that is, trip demands are largely derived from activity demands (Ettema et al., 1993; Garling et al., 1994; Kitamura et al., 1997; Ben-Akiva and

* Corresponding author.
E-mail addresses: hesam@dal.ca (M.H. Hafezi), naznin.daisy@dal.ca (N.S. Daisy), hugh.millward@smu.ca (H. Millward), lei.liu@dal.ca (L. Liu).

https://doi.org/10.1016/j.trc.2021.102972
Received 7 September 2019; Received in revised form 31 October 2020; Accepted 6 January 2021
Available online 3 February 2021

Bowman, 1998). Activity-based models focus on the 24-hour activity schedule, and link travel episodes with activities performed by individuals (Kitamura et al., 2000). The aim of this paper is to predict travel behavior for a set of model individuals, who represent cohorts with homogeneous time-use activity patterns. The paper presents a new modeling framework that is able to simulate temporal information associated with the traveler's daily activity schedule, for use in activity-based travel demand modeling.

Over the last half century, many theoretical and practical activity-based travel demand models have been developed. Broadly, these models can be classified into three major groups based on their structure and purpose (De Palma et al., 2011): constraints-based models, random utility models, and computational process models. In constraints-based models, individual's activity patterns are shaped based on Hagerstrand's time-geography theory (Hagerstrand, 1970). In random utility models the time component is modelled as a discrete component (Bhat et al., 2004), while in computational process models it is modelled as a continuous component (Arentze and Timmermans, 2000). By generating more accurate activity patterns we can improve the timing inputs to the scheduling engine in activity based models (Rasouli and Timmermans, 2012). Researchers employ various techniques for producing temporal information associated with individual's daily activity schedules, such as probability distribution function, hazard function, and decision tree (Timmermans and Zhang, 2009; Auld et al., 2009; Roorda et al., 2008).

Over the past few years, machine learning techniques have become more efficient at storing and processing large amounts of data, and learning complicated correlations between variables in the transportation arena. Despite much progress in predicting daily activity patterns through econometric or rule-based techniques, there have been few efforts to employ the capability of machine learning to learn daily activity engagement patterns. In this study, we develop a new model that is able to learn and predict temporal information associated with travelers' activities in their daily agenda. Employing a robust machine learning method known as Random Forest (RF), decision trees were grown using the Classification and Regression Tree (CART) technique. The resulting models were trained from observed activity durations. Previous studies reveal that ensemble methods performed better when they were applied to clustered data (Allahviranloo and Recker, 2013; Jiang et al., 2012; Liu et al., 2015). Therefore, in this study, we applied the models to twelve unique population clusters derived using a novel pattern recognition model. The results of this paper are expected to be implemented within the activity-based travel demand modeling framework Scheduler for Activities, Locations, and Travel (SALT). The SALT framework comprises five main components: population synthesizer, time-use activity pattern recognition, tour mode choice, activity destination choice, and activity/trip scheduling. The modeling framework adopts a pattern recognition approach which identifies population clusters with homogeneous time-use activity patterns. A series of behaviorally realistic econometric models and rule-based models are then developed for modeling time-use activity patterns in each identified cluster. Finally, this paper contributes by providing additional insights to the scheduling modules in the overall activity-based modeling framework. The remainder of the paper is structured as follows: following the literature review, a discussion of the data used in the modeling is presented. The methods for modeling activity-travel scheduling behaviors are described in the next section, followed by a discussion of model results. The paper concludes with a summary of the main contributions, and an outline of future research directions.

2. Literature review

Activity-based models focus on building the 24-hour activity schedule, and link travel episodes with activities performed by individuals (Kitamura et al., 2000; Hafezi et al., 2018; Daisy et al., 2018c). These models include a series of universal components such as activity generator and scheduler, tour mode choice, tour and trip time of day, tour and trip destination, and network assignment. Since activity demands generate trip demands, more accurate modeling of activity sequences will improve the accuracy of modeling trip departure time and trip duration (Rasouli and Timmermans, 2012; Oberkampf et al., 2002).

A wide array of theory and methods have been developed to produce information for the scheduling module in activity-based travel demand models. For instance, in CARLA (Combinatorial Algorithm for Rescheduling Lists and Activities), activities are generated and added to the individual's schedule using four rules: logical rules that refer to the presumption of one unique activity at a time at one location, environmental rules that refer to authority constraints (e.g. access time restrictions to different places, and travel times between locations), inter-personal rules that refer to coupling constraints (e.g joint activities with other household members), and personal rules that refer to personal preferences (Jones et al., 1983). In STARCHILD (Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions), activities are generated in three steps. First, all possible alternatives to participating in different activities with respect to all constraints are explored. Next, through a series of statistical tests, similar alternatives are clustered in three to ten groups. Finally, a representative activity travel pattern is chosen for each group. The ultimate activity choices are estimated by the multinomial logit model and activities are scheduled through employing a series of rules (Recker et al., 1986; Recker et al., 1986).

In the cognitive model, alternative decisions for shaping individual's agenda at various levels of abstraction are generated through the application of a series of rules for activity planning processes. The cognitive model of planning can be accounted as the first rule-based simulation activity-based model (Hayes-Roth and Hayes-Roth, 1979). In AMOS (Activity Mobility Simulator), the choice of different potential activities for the individual is generated from alternative activity travel patterns. Activities are ordered in the agenda through a rule-based scheduling engine, and an activity adjuster is used for conflict resolution. Activity purposes, frequencies, time budget, duration, location, and priority list are inter-connected in AMOS (Kitamura et al., 1996). In SMASH (Simulation Model of Activity Scheduling Heuristics), a set of alternative activity travel patterns, along with type, timing, travel mode, travel time, and location for each activity, are generated at the first step (Ettema et al., 1993). Next, the searching process considers ties in activity time and adds the activities that have been prioritised as high in the schedule. Decisions for adding or rescheduling activities in SMASH are made based on the choice of activities, sequencing, travel mode, travel time, location, and choice of joint activity.

Some researchers have used explanatory data analysis and statistical methods to generate activities. For instance, in TASHA (Travel

and Activity Scheduler for Household Agents), activities with similar socio-demographic and temporal characteristics (e.g. start time) are grouped into classes through a series of empirical data analysis (Miller and Roorda, 2003). Start time and activity duration are generated through a probability distribution function. A series of heuristic rules (e.g. add, delete, shift, or truncate activities) are used to schedule the individual's agenda. In ADAPTS (Agent-Based Dynamic Activity Planning and Travel Scheduling Model), activities with similar characteristics are identified through a hazard function, and essential information for the scheduling engine, such as activity start times and durations, are generated. Activities are added to the individual's agenda using a set of heuristic rules based on the TASHA model (Auld and Mohammadian, 2009). In ALBATROSS (A Learning-Based Transportation Oriented Simulation System), activities are generated and added to the individual's agenda based on their flexibility level and spatial-temporal constraints (Arentze and Timmermans, 2004). These constraints include location, travel mode, and time budget availability. Start time and activity duration are predicted using the CHAID decision tree. The scheduling engine in ALBATROSS uses these constraints to order activities. Fixed activities are added to the individual's agenda first, and flexible activities are then scheduled with respect to prior fixed activities.

In recent years there has been growing interest in incorporating machine learning techniques in the modeling of activity generation and activity scheduling steps in activity-based travel demand models (Liao et al., 2007; Allahviranloo et al., 2017; Li and Lee, 2017). For instance, Allahviranloo used a k-mean clustering algorithm to identify unique clusters and utilized the AdaBoost algorithm to predict start time and activity duration based on the socio-demographic characteristics of individuals. In another study, a probabilistic context-free grammars technique is adopted to produce a set of activities in the individual's daily agenda (Li and Lee, 2017). Also, hierarchical markov models have been used for predicting individual's activity choices (Liao et al., 2007). In another study, a sampling method using activity-travel pattern type clustering and travel distance is developed to derive multi-day activity-travel data through sampling from readily existing single-day household travel survey data (Zhang et al., 2018). The interpersonal changeability detected in cross-sectional single-day data of a cluster of individuals is used to produce the day-to-day intrapersonal variability. Drchal et al. introduced a data driven activity scheduler approach for activity scheduling component in the agent-based mobility models (Drchal et al., 2019). The new technique swaps several expert-designed mechanisms and their complicatedly engineered interactions with a collection of machine learning models.

Despite much progress made in generating temporal information for the scheduling engine in activity-based travel demand models, there is undeniably considerable room for improving model performance in terms of estimation accuracy, computational efficiency, and practical application. In this paper, we propose an algorithm employing the Random Forest approach to model temporal dimensions associated with the traveler's daily activity patterns. Although the RF method is used in other transportation fields such as traffic incident detection and transport mode recognition (Shafique and Hato, 2015; You et al., 2017), to the best of our knowledge RF models using the Classification and Regression Tree (CART) classifier, combined with variable importance measurement methods, have not previously been employed for predicting start time and activity duration in travel behavior analysis. The proposed theoretical model in this study is able to learn and predict temporal information of individual's daily activity patterns with regard to heterogeneity features. The application of this framework to activity-based modeling not only reveals the efficiency of machine learning to model temporal attributes of daily traveler activity patterns, but also contributes further insights to the linkage between activity scheduling and trip assignment modules.

3. Data

This study uses time-diary and GPS geo-coordinate data, from the Space-Time Activity Research (STAR) household survey undertaken in Halifax, Canada. STAR was a combined time-use and travel survey, and included the world's first large-scale employment of global positioning system (GPS) technology for tracking out-of-home activity. The STAR project produced a wide variety of data, including the household roster data, main file (questionnaire information data file), vehicle data, time diary (episode and summary data file), activity diary (episode data file), land use database, business hours survey data, places and locations (PAL) directory data, and global positioning systems (GPS) data. Full descriptions of the survey design and the socio-demographic features of respondents can be found in (Daisy et al., 2018a, 2018b) and TURP (TURP, 2008).

STAR survey data were collected from 1971 randomly designated households in Halifax Regional Municipality (HRM), in the period April 2007 to May 2008. In each household, a primary respondent over age 15 was randomly selected, and completed a 2-day time diary, supplemented and verified through GPS tracking. In the research reported here, nine aggregated classes of activities were included in the proposed model: home chores, home leisure, night sleep, workplace, shopping & services, school/college, organizational/hobbies, entertainment, and sports activities. The final data set after data cleaning comprised 2778 weekday person-days (1389 individuals, twodayseach).

In this study, we use data on twelve clusters of respondents previously-established through application of a novel pattern recognition modeling framework to the STAR time-use activity data. Pattern complexity of weekday activity sequences in the dataset was recognized using the fuzzy C-means (FCM) algorithm, and respondents with similar patterns were grouped into homogeneous cohorts (Hafezi et al., 2017, 2018). The cluster memberships selection in the FCM is comparable to other alternative clustering approaches such as k-means (Jiang et al., 2012; Liu et al., 2015; Oberkampf et al., 2002; Jones et al., 1983; Recker et al., 1986; Recker et al., 1986; Hayes-Roth and Hayes-Roth, 1979; Kitamura et al., 1996; Miller and Roorda, 2003; Auld and Mohammadian, 2009; Arentze and Timmermans, 2004; Liao et al., 2007; Allahviranloo et al., 2017). In the FCM each data point has the probability of belonging to numerous clusters, and this results in creating more homogeneous activity patterns in each cluster. For example, we identified two different worker clusters that, regardless of their similarity in activity sequences, are distinguished by start time and end time at the workplace. Furthermore, the FCM clustering algorithm is more straightforward and easy to implement in practical activity-based

travel demand models compared to previous studies that used complex methods to capture frequent and infrequent activities in a dataset (Liu et al., 2015).

Table 1 and Table 2 present an analysis of the cluster data, showing socio-demographic characteristics of cluster members, and summary data on their daily time-use activity patterns. Overall, the pattern recognition model recognized twelve clusters as follows: cluster#1: extended worker, cluster#2: non-worker midday activities, cluster#3: 8–4 worker, cluster#4: non-worker evening activities, cluster#5: stay-at-home, cluster#6: shorter worker, cluster#7: 7–3 worker, cluster#8: non-worker morning shopping, cluster#9: non-worker afternoon shopping, cluster#10: evening worker, cluster#11: 9–5 worker, and, cluster#12: students. Individuals in each cluster differ considerably from those in other clusters with regards to daily activity type, activity sequencing, probability of start time, and activity duration, and also with regard to their socio-demographic characteristics.

4. Methods

The proposed modeling framework for scheduling travelers' activities in this paper consists of two steps, as follows. First, temporal information including start time and activity duration for the set of activities in the agenda is predicted using the Random Forest (RF) model. Second, predicted activities are inserted into a skeleton schedule through a heuristic decision rule-based technique, and are scheduled with respect to two-tier constraints. The algorithm is started by predicting cluster membership for the selected individual. This process is done through a Classification and Regression Tree (CART) developed based on the socio-demographic characteristics of individuals in each cluster. Based on the random generated number and cumulative probability functions, the CART algorithm can find specific clusters for particular leaf nodes based on the high probability that an individual belongs to it (Hafezi et al., 2017, 2018; Daisy et al., 2018). Next, activity agendas and activity sequences of the traveler (the relationships between activities) are predicted using the advanced RF model. For each activity type, start time and activity duration are generated from a uniform distribution within their interval time range. Random Forest theory is based on the use of numerous decision trees that have been grown using the bagging technique (Breiman, 2001; Suthaharan, 2015). Each tree acts as a weak learner in the algorithm and the aggregation of these weak inputs provides a powerful ensemble learning model. Outcomes achieved from the RF model are based on the majority votes in the ensemble models.

4.1. The Random Forest (RF) model

The Random Forest (RF) structure for predicting start time and activity duration is shown in Fig. 1. The RF method is based on an ensemble of many decision trees (Breiman, 2001; Suthaharan, 2015). Each tree acts as a weak learner and makes a prediction $\left\{\widehat{P}_1, \widehat{P}_2, \widehat{P}_3, \cdots, \widehat{P}_M\right\}$ and the grouping of these weak inputs constructs a robust ensemble learning model. The eventual prediction outcome gained is based on the majority votes for $\widehat{P}$. The predictor variables X_n are socio-demographic characteristics of travelers and are selected from Table 3. The response variablesY_i are defined as one of the activity start time /activity duration bin numbers in the model. The start time and activity duration for each activity type, based on their duration intervals, are transformed to a set of bins. We use start time bins spaced 10,15,30 and 180 min apart, and duration bins of 15,30,60 and 360 min duration. The modeling period starts at 4 a.m., and continues for 24 h.

Generally, the RF model takes the following steps. First, test dataset (T_e) and training dataset $D_r = ((X_i, Y_1), \cdots, (X_i, Y_N))$ are drawn from the primary dataset using a cross-validation partition process. Prior to the growing of each tree, $E_n(C)$ observations are randomly drawn from the sampled data points $E_n \in \{1, \cdots, n\}$. Next, at each node in the decision tree a number of predictor variables are randomly

Table 1
Share of Various Socio-Demographic Variables for Twelve Respondent Clusters.

Social demographic variables		Sample mean (%)	Mean of cluster (%)											
			#1	*#2*	*#3*	*#4*	*#5*	*#6*	*#7*	*#8*	*#9*	*#10*	*#11*	*#12*
Gender	Female	0.53	0.53	0.53	0.44	0.59	0.56	0.52	0.47	0.54	0.59	0.51	0.53	0.60
Age	Young adults (ages 15–35 years)	0.10	0.12	0.05	0.10	0.10	0.11	0.10	0.05	0.09	0.07	0.15	0.09	0.60
	Middle-aged adults (ages 36–55 years)	0.49	0.67	0.29	0.66	0.38	0.32	0.71	0.72	0.29	0.32	0.66	0.70	0.31
	Older adults (aged older than 55 years)	0.41	0.20	0.66	0.24	0.53	0.57	0.19	0.23	0.63	0.61	0.19	0.22	0.08
Education	High School Diploma or higher	0.67	0.76	0.58	0.76	0.62	0.66	0.85	0.53	0.57	0.65	0.64	0.80	0.38
Occupation	Regular shift	0.53	0.73	0.22	0.93	0.26	0.24	0.87	0.93	0.19	0.24	0.43	0.89	0.13
	Irregular schedule	0.10	0.22	0.10	0.03	0.10	0.11	0.09	0.07	0.07	0.07	0.47	0.08	0.02
	Student	0.03	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.03	0.02	0.04	0.01	0.67
	Retired	0.23	0.02	0.52	0.02	0.39	0.41	0.01	0.00	0.53	0.41	0.00	0.00	0.08
	Work at home	0.15	0.23	0.10	0.13	0.15	0.16	0.30	0.06	0.11	0.09	0.09	0.26	0.02
Flexible schedule	Have no flexibility in a work schedule	0.50	0.55	0.48	0.54	0.46	0.43	0.44	0.63	0.48	0.51	0.75	0.40	0.43
Job number	Have more than one job	0.07	0.09	0.04	0.04	0.16	0.08	0.05	0.07	0.11	0.05	0.08	0.08	0.00
Income	Low-income (<= $ 40,000)	0.39	0.28	0.44	0.22	0.48	0.49	0.32	0.29	0.53	0.48	0.47	0.26	0.78
	Middle-income ($ 40,000 - $ 100,000)	0.53	0.60	0.46	0.68	0.45	0.45	0.55	0.64	0.42	0.46	0.49	0.59	0.19
	High-income (> $ 100,000)	0.09	0.12	0.10	0.10	0.07	0.07	0.13	0.08	0.05	0.06	0.04	0.15	0.03

Table 2
Summary Statistics for Twelve Respondent Clusters: Membership Analysis.

	Cluster number											
	#1	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***
Cluster membership (person-days)	137	225	401	238	419	171	229	247	262	53	348	48
Cluster membership (%)	4.93	8.10	14.43	8.57	15.08	6.16	8.24	8.89	9.43	1.91	12.53	1.73
Home chores (%)	25.09	34.73	27.81	41.39	43.10	33.78	30.14	40.54	40.12	35.01	29.47	35.52
Home leisure (% of in-home time)	12.04	20.26	16.57	14.42	19.17	16.22	17.73	19.96	19.18	13.61	15.32	13.81
Night sleep (%)	62.88	45.01	55.62	44.19	37.73	50.00	52.13	39.50	40.70	51.38	55.21	50.67
Total in-home (%)	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*
Workplace (% of out-of-home time)	**89.26**	4.76	**82.61**	6.41	8.16	**79.53**	**89.31**	10.53	13.12	**90.90**	**86.56**	5.77
Shopping & services (%)	1.61	27.27	3.16	15.31	30.19	5.56	2.41	**32.26**	**38.27**	2.90	2.98	3.38
School/college (%)	0.00	1.00	0.14	1.20	0.42	0.40	0.29	3.16	1.28	0.66	0.15	**69.29**
Organizational/hobbies (%)	2.34	**29.57**	3.24	**31.28**	19.90	5.11	1.76	21.11	21.94	1.57	2.62	11.14
Entertainment (%)	4.67	17.49	6.92	**33.73**	10.35	6.48	3.73	8.75	11.66	1.93	4.30	4.53
Sports (%)	2.12	19.90	3.93	12.06	**30.98**	2.93	2.50	24.20	13.74	2.04	3.40	5.89
Total out-of-home (%)	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*

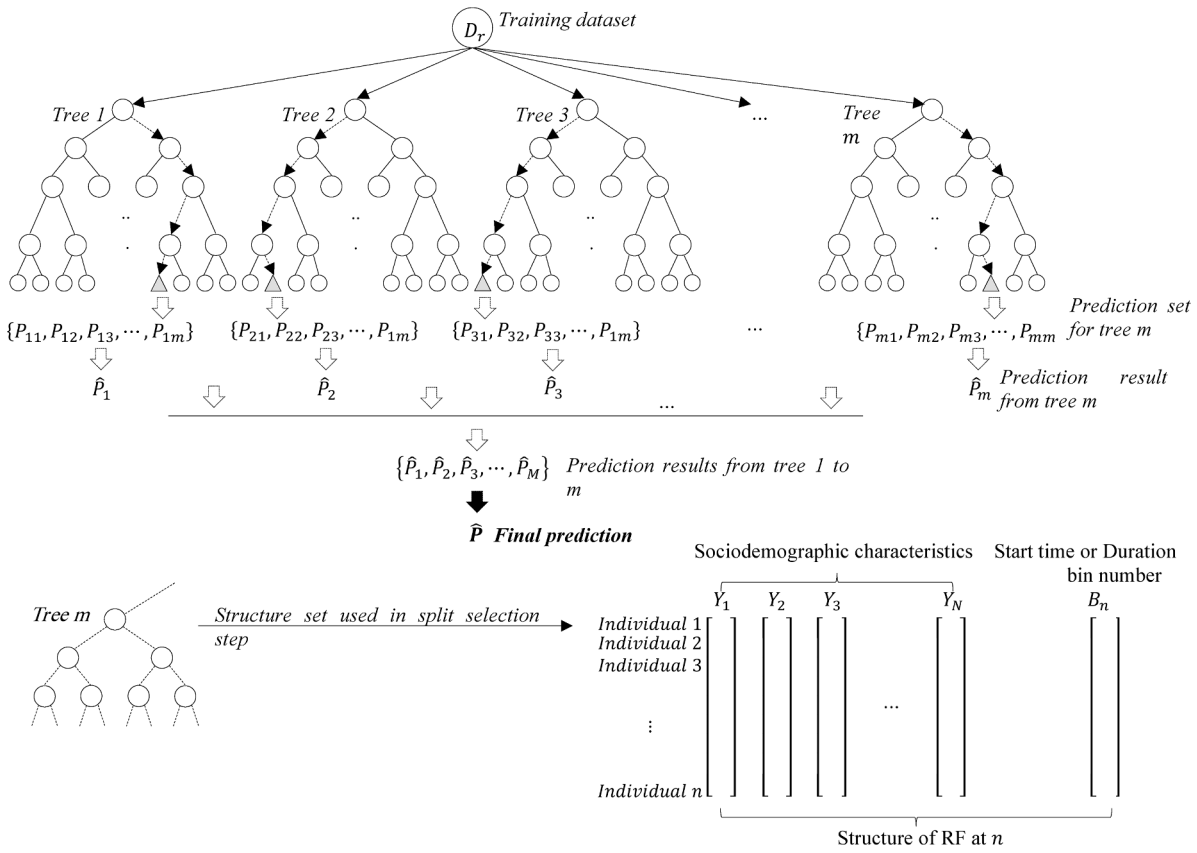


Fig. 1. Random Forest Structure for Predicting Temporal Information Associated with the Traveler's Daily Activity.

selected from all the available predictor variables (Q). The default number of selected predictor variables is the square of the total number of existing predictor variables ($\sqrt{Q}$). Lastly, a split at each node is achieved using the split predictor selection search. We undertake $E_n(C)$ as a subset of training data D_r where C cut in C is as pair (u, w). u is the randomly nominated predictor variables. w is the place of the split along the u-th correspondent, within the limits of C. We hypothesize the set of all such possible cuts in C as T_C. The split condition $S_{clas,n}(u, w)$ is computed as follows (Breiman, 2001; Biau and Scornet, 2016):

$$S_{clas,n}(u,w) = \frac{1}{E_n(C)}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_C\right)^2 \exists_{X_i \in C} - \frac{1}{E_n(C)}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_{C_L}\exists_{X_i^{(u)} < w} - \overline{Y}_{C_R}\exists_{X_i^{(u)} \geq w}\right)^2 \exists_{X_i \in C} \quad (1)$$

Table 3
Proposed Predictor Variables for Predicting Temporal Information Associated with the Traveler's Daily Activity.

Predictor	Subcategories
Gender	Male, female
Age	15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85+
Marital status	Married, living common-law, widowed, separated, divorced, single-never married,
Household size	1,2,3,4,5,6
Highest education level	Masters or earned doctorate, bachelor or undergraduate degree, diploma or certificate, some university, some community college, trade, technical or business college, high school/secondary, other
Full/part-time student	Full-time student, part-time student
Paid/self employed	A paid worker, self-employed, an unpaid family worker
Flexible work schedule	No, yes
Work at home	No, yes
Total personal income	Under $20,000, $20,000–$39,999, $40,000–$59,999, $60,000–$79,999, $80,000–$99,999, $100,000 or more
Total household income	Under $20,000, $20,000 - $39,999, $40,000 - $59,999, $60,000 - $79,999, $80,000 - $99,999, $100,000 or more
Dwelling type	Single unit residential, duplex or semi-detached, townhouse, multi-unit residential-less than 6 stories, multi-unit residential-6 or more stories, mobile dwelling, other-specify
Dwelling owned/rented	Owned, rented
Valid driver's license	No, yes
Buss pass	No, yes
Number household vehicles	0,1,2,3,4,5,6
Number household motorcycles	0,1,2,3,4,5,6,7,8,9,10
Number household bicycles	0,1,2,3,4,5,6,7,8,9,10
Usual mode to work	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
Usual mode to school	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
State of health	Excellent, very good, good, fair, poor
Activity start time*	Corresponding duration bin number for each activity type in the agenda
Activity duration*	Corresponding start time bin number for each activity type in the agenda
Prior activities**	Home chores *(working at home, eating/meal preparation, indoor or outdoor cleaning, interior or exterior home maintenance, child care, other in home activities)*, Home leisure *(watching tv/listening to radio, reading books/newspapers, etc.)*, Night sleep, Workplace *(work/job, all other activities at work, work related, etc.)*, Shopping & services *(shopping for goods and services, routine shopping)*, School/college *(class participation, all other activities at school)*, Organizational/hobbies *(organizational, voluntary, religious activities. hobbies done mainly for pleasure, cards, board games, all other hobbies activities)*, Entertainment *(eat meal outside of home, all other entertainment activities)*, Sports *(walking, jogging, bicycling, all sports related activities)*

* Use only in predicting start time and activity duration
** Use only in predicting activity agendas and activity sequences

$$X_i = \left(X_i^{(1)}, \cdots, X_i^{(p)}\right) \forall (u, w) \in T_C \tag{2}$$

$$C_L = \left\{x \in C : x^{(u)} < w\right\} \tag{3}$$

$$C_R = \left\{x \in C : x^{(u)} \geq w\right\} \tag{4}$$

where $\overline{Y}_C$ is the average of the Y_i such that X_i belongs to C.

One of the main challenges in most decision tree models is to select the best split predictor method (Breiman, 2001; Biau and Scornet, 2016). Compared to the ID3 and C4.5 decision tree models, the CART classifier can handle both numeric and categorical variables and it can easily handle outliers. In this study, we used the CART algorithm for decision splits in the RF model. The best cut (u_n^*, w_n^*) is computed by maximizing $S_{clas,n}(u, w)$ over B_{lit} and T_C.:

$$\left(u_n^*, w_n^*\right) \in \underset{\substack{u \in B_{lit} \\ (u,w) \in T_C}}{argmax} S_{clas,n}(u, w) \tag{5}$$

This process is terminated when all the information is saturated. The new input data at the testing stage $t_n \in \{1, \cdots, e\}$ is propagated down to all of the trees in the RF model, and a set of prediction results from all trees is obtained $\left\{\widehat{P}_1, \widehat{P}_2, \widehat{P}_3, \cdots, \widehat{P}_M\right\}$. The final prediction result gained is based on majority votes $\widehat{P}$.

4.2. Variable importance measures: Mean decrease accuracy (MDA)

One of the advantages of the RF model is its ability to measure the importance of variables and subsequently rank them in order to guide a decision split algorithm for finding the best cut points. This process in the RF model can be performed through Mean Decrease Accuracy (MDA) or Mean Decrease Impurity (MDI) computations. In this study, we measured the importance level of predictor var-

iables using the Mean Decrease Accuracy (MDA) method that builds on the out-of-bag (OOB) error estimate (Breiman, 2001; Biau and Scornet, 2016). The variables for predicting temporal information associated with the traveler's daily activity are selected from Table 3 and their importance level for incorporating in the RF model is estimated through the MDA method. The MDA of the variable $X^{(u)}$ is computed by balancing the difference in OOB error (O_n) prior and subsequent to the permutation over all trees. The primary OOB error of each tree is calculated by testing the RF model using OOB data. The later OOB error of each tree is computed by adding noise to the sample data of the feature randomly and retesting the OOB error. The MDA for randomly nominated variable $X^{(u)}$ is computed as follows:

$$\mathrm{MDA}(X^{(u)}) = \frac{1}{m}\sum_{l=1}^{m}\left[O_n\left[d_n(;\Theta_m), I^{u}_{m,n}\right] - O_n\left[d_n(;\Theta_m), I_{m,n}\right]\right] \tag{6}$$

$$O_n[m_n(;\Theta_l), I] = \frac{1}{|I|}\sum_{i:(X_i,Y_i)\in I}\left(Y_i - m_n(X_i;\Theta_l)\right)^2 \tag{7}$$

$I_{m,n}$ is the out-of-bag data set of the m-tree, $I = I^{u}_{m,n}$ is the permutated data for variable u and $O_n(.;\Theta_m)$ is the estimation for the m-th tree. In total 70% of the dataset was used for training the model and 30% for testing model performance. For simplicity's sake in this study, we refer to the setting with random predictors selection as RF_CART_I and the setting with selected (important) predictors as RF_CART_II.

4.3. Model calibration and validation

In the RF algorithm several parameters need to be initialized and calibrated. These are the initial number of trees, the node split principle, and the cutoff vector. The initial number of trees m is set to 1000, and was calibrated by out-of-bag error estimation to verify if the model can be converged within this value and to address possible variables' multicollinearity in the feature selection. In this study, the best split at each node B_{lit} is determined with the CART algorithm. Another alternative approach to find the best split in the RF model is the Curvature Search technique. The cutoff vector T_C is a vector of length equivalent to the number of classes. T_C includes three sub-parameters ($c1$, $c2$, $c3$) that are originally each randomly set in the range between [0,1] with the requirement that the total sum is equal to 1. These parameters were cross-validated through the OOB error rate in order to achieve the best parameter values for the proposed RF model.

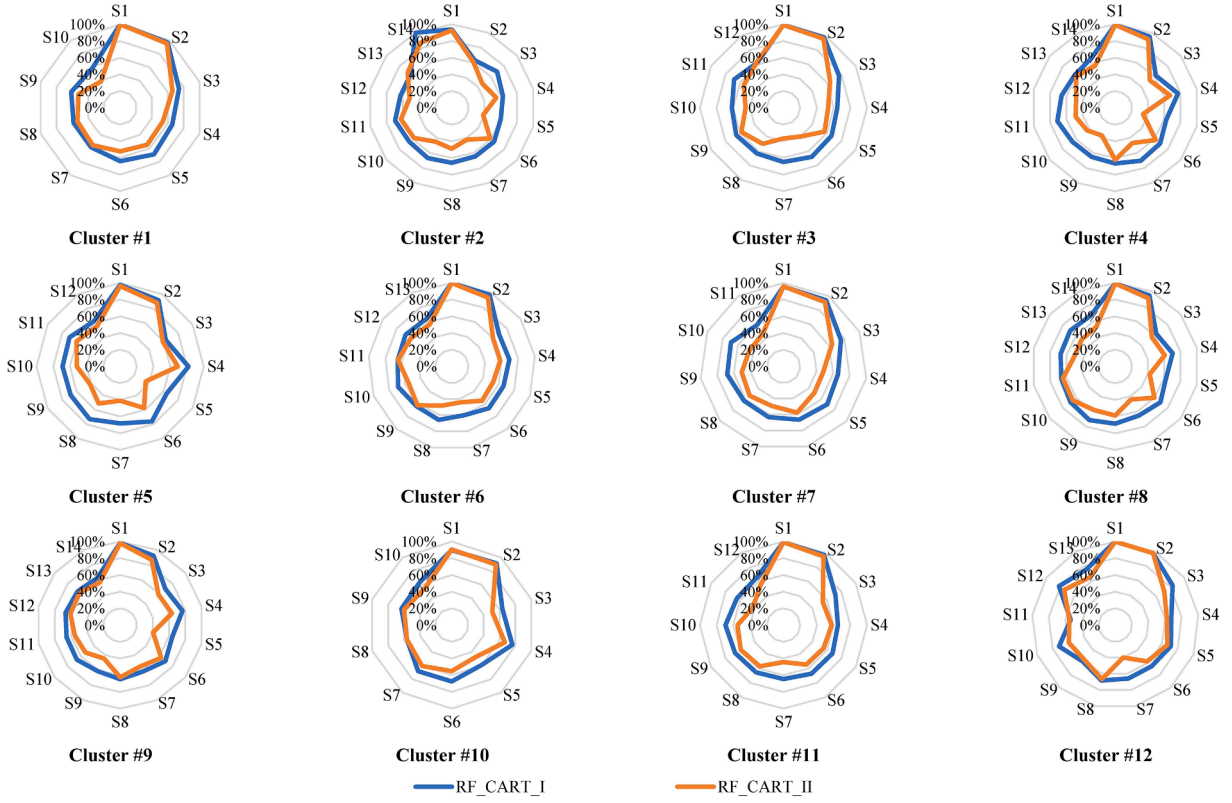


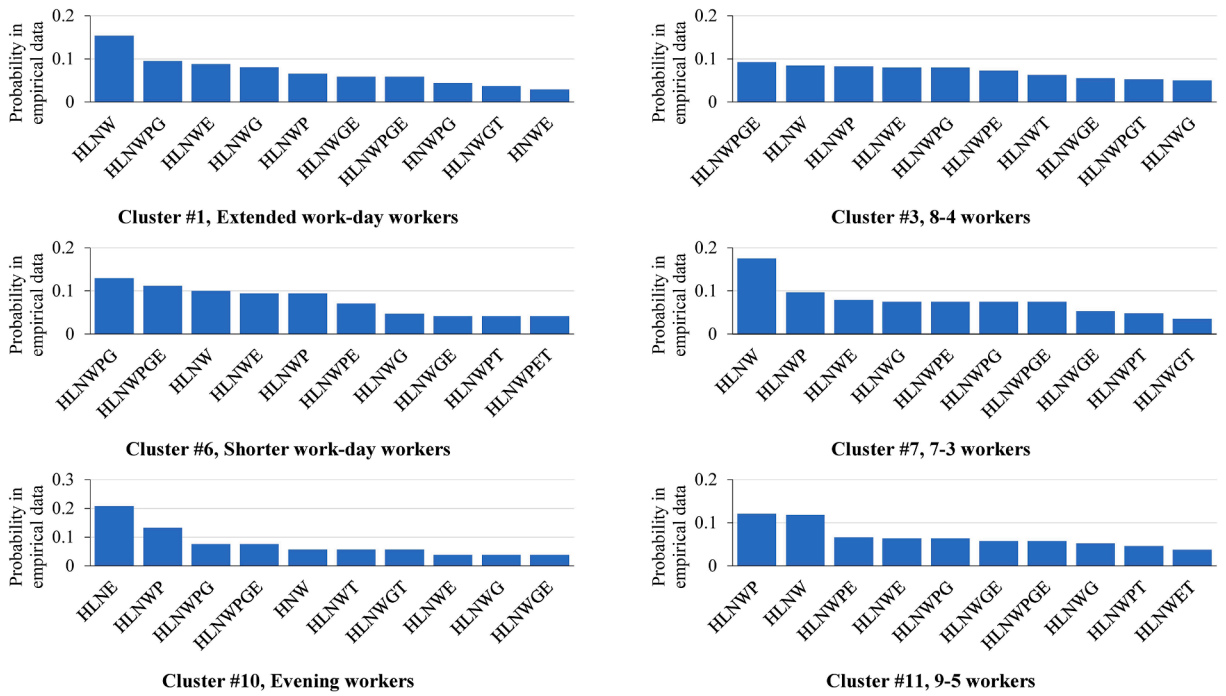
Fig. 2. Comparing the Estimation Precision between Two RF Models: RF_CART_I (Random Predictors) RF_CART_II (Important Predictors).

5. Discussion of results

In order to analyze the efficiency and performance of the RF model under different conditions and complex activity patterns, we applied the model to 12 clusters drawn from the Space-Time Activity Research (STAR) survey. In total, six clusters were recognized as relating to out-of-home workers, four clusters to non-worker non-students, and separate clusters for students and individuals who mostly spend their time at home ("stay-at-homes"). The membership size of clusters varied in the range of 48 to 419. Analysis of cluster members showed statistically significant differences between clusters in terms of their distributions of start time, activity duration, activity type, and socio-demographic characteristics (Daisy et al., 2020; Millward et al., 2020).

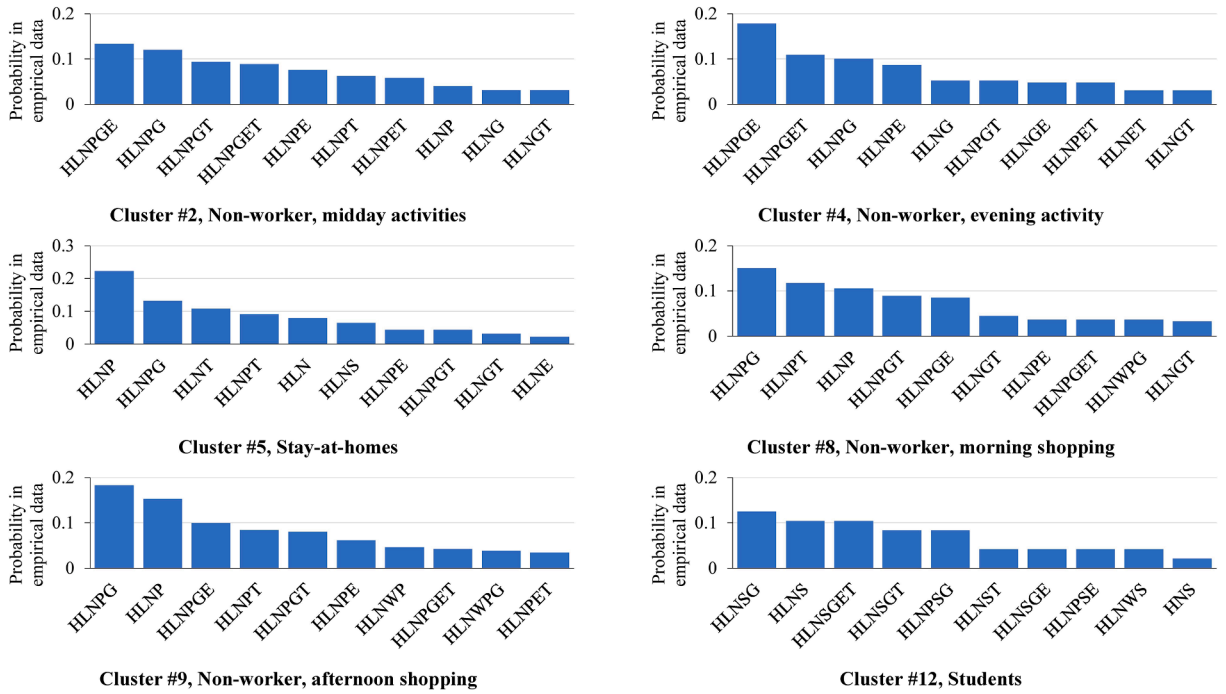
Employing data for the 12 clusters, activity agendas and activity sequences of travelers were predicted using the advanced RF model. Comparison of the estimation accuracy between two RF models, including RF_CART_I (random predictors) and RF_CART_II (important predictors) for all 12 clusters is shown in Fig. 2. Results showed that the model with random selection of predictor variables performed better than including only high importance predictor variables. Moreover, the number of activities for the simulated travelers in the RF_CART_I model have a similar distribution (same median and inter-quartile) to the observed number. Fig. 3 and Fig. 4 illustrate the distribution of the 10 most frequent combinations of agenda in the travelers' daily activity patterns for identified worker and non-worker clusters, respectively. For all groups, the most frequent 24-hour agenda combinations (regardless of sequence) include home-chores, leisure, and night sleep (HLN). For worker groups (Fig. 3), workplace activity (W) obviously features in nearly all frequent combinations, with the exception of the most typical combination for evening workers, which is HLNE. Out-of-home non-work activities (particularly shopping and hobbies, S and G) occur more frequently for shorter workday workers, and least frequently for evening workers and 9–5 workers. For non-worker groups (Fig. 4) both shopping and hobbies occur in nearly all the most frequent combinations, with even the stay-at-home group having some out-of-home activities. Frequent combinations for the student group all include school/college (S), but shopping is notably absent from their agenda combinations.

The RF model for predicting start time and activity duration was run under eight different bin settings for a twofold purpose: to test the efficiency level of the model and to compare results with other alternative techniques. The OOB rate error cross-validation specifies that after $m > 850$ (m is the number of trees), the OBB error rate tends to be constant. Therefore, it is reasonable to accept m as 1000 at first. In view of this, we achieve the best optimal parameters for RF models as follows: $c1 = 0.25$, $c2 = 0.35$ and $c3 = 0.40$. For each cluster, the activity start time and activity duration (bin numbers) for all activity types in travelers' agenda are predicted. The model results for start time and activity duration prediction are presented in Table 4 and Table 5, respectively. The prediction accuracy is obtained from comparison of observed and predicted bin numbers for every activity type. The best result for predicting start time was obtained by setting IV (8 bins, each of duration 180 min), at 60.10% accuracy, followed by setting III (48 bins, each of duration 30 min), at 36.28%. Similarly, the best prediction result for activity duration was found under setting $\widehat{\text{IV}}$ (4 bins, each of duration 360



*Horizontal axis: Occasions (H=Home chores, L=Home leisure, N=Night sleep, W=Workplace, P=Shopping & services, S=School/college, G=Organizational/hobbies, E=Entertainment, T=Sports). **Agenda combinations shown in Figure 4 are regardless activity sequences

Fig. 3. Distribution of 10 Most Frequent Combinations of Agenda in the 24-Hour Day for Six Identified Worker Clusters.



*Horizontal axis: Occasions (H=Home chores, L=Home leisure, N=Night sleep, W=Workplace, P=Shopping & services, S=School/college, G=Organizational/hobbies, E=Entertainment, T=Sports). **Agenda combinations shown in Figure 5 are regardless activity sequences

Fig. 4. Distribution of 10 Most Frequent Combinations of Agenda in the 24-Hour Day for Six Identified Non-Worker Clusters.

min), at 98.65%, followed by setting $\widehat{\text{III}}$ (24 bins, each of duration 60 min), at 67.32%.

Although there is little value in employing highly aggregated bins to gain overall accuracy, the tables show that it is possible to predict for certain groups and certain activities with bins of short duration. For application to trip modeling, the focus should be on predicting start times of out-of-home activities to within 15 min, and predicting their durations to within 30 min (and preferably 15 min). For most groups and most activities, start times are more difficult to predict than duration times.

For activity start times (Table 4), both night sleep and sports activities are highly predictable, even with times spaced at 10 and 15 min. Individuals tend to schedule these at regular times of the day. In contrast, shopping, entertainment, and hobbies have low predictability for most groups, although there are exceptions, such as entertainment for group 9 (non-workers with afternoon shopping). Workplace and school start times, which are critically important for trip modeling, are highly predictable to within 15 min for groups 12 and 10 (students, evening workers), moderately predictable for group 7 (7–3 workers), but much less predictable for groups 1, 3, 6, and 11 (extended-day, 8–4, shorter-day, and 9–5 workers).

For activity durations (Table 5), in-home activities are only moderately predictable, with the exception of group 12 (students). Of the out-of-home activities, shopping is most predictable for most groups, though less so for groups 2 and 4 (non-workers with midday and evening outside activities). Durations for sports, entertainment, and hobbies are also highly predictable for some groups. Workplace and school durations are particularly important for trip modeling, and there is high predictability of 15-minute durations for group 12 (students). Groups 6 and 7 (shorter workday and 7–3 workers) show moderate predictability at the 15-minute level. However, even with bins of 30-minute duration the large 8–4 and 9–5 workers (groups 3 and 4) remain highly unpredictable. This probably reflects the variable timing of lunch breaks, and also the greater self-management of schedules for these worker groups.

The empirical results show that with increases in the time interval (increasing bin numbers), the RF efficiency also increased. The lowest model result was reported for predicting start time with 144 bins, each of duration 10 min (setting I), at only 32.95% accuracy.

As was discussed earlier, in total 70% of the dataset was used for training the model and 30% for testing model performance. In order to evaluate the performance of the heuristic rule-based algorithm, the estimation errors in minutes on a continuous scale and in percentage were computed to show the duration of misclassification, and are shown in Table 6. For every activity type in each cluster, the error is estimated by calculating the edit-distance between the observed activity pattern and projected temporal pattern in the test set. Results show that the highest misclassification error in each cluster is for those activities with a shorter duration in the traveler's daily activity patterns. For example, in the extended work-day workers cluster, entertainment activity has the highest misclassified error by 32.74 min. Further studies to overcome this limitation associated with activity types with shorter duration are recommended. The total error in each cluster was estimated based on the summation of all misclassification errors over 24-hours of projected traveler's activity. The highest error percentage was found for the student cluster, at 36.71%, followed by the evening worker cluster, at 25.50%. Compared to other clusters, students and evening worker clusters had the lowest sample size in our model. Empirical results

Table 4
Percentage Accuracy of Activity Start Time Estimation for Test Dataset.

Activity type	Estimation accuracy for setting I (144 bins, each of duration 10 min)												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	7.98	27.40	25.01	7.53	7.94	10.77	30.46	7.65	26.00	37.96	12.26	33.78	***Mean accuracy 32.95%***
Home leisure	14.81	16.91	12.01	5.56	53.57	16.47	41.92	9.62	27.19	–	7.42	–	
Night sleep	48.41	39.70	36.69	56.25	55.00	44.91	37.67	40.74	36.21	85.00	33.08	34.30	
Workplace	10.21	–	16.67	–	–	13.75	35.92	–	–	66.67	16.55	82.00	
Shopping & services	–	16.67	32.00	13.35	–	22.86	25.00	9.30	13.69	–	10.00	33.33	
School/college	–	–	–	–	–	–	–	–	–	–	–	31.25	
Organizational/Hobbies	–	19.64	24.07	60.00	50.00	21.59	35.00	25.00	14.29	–	43.06	–	
Entertainment	–	–	22.26	–	–	–	27.78	–	62.50	–	31.25	50.00	
Sports	60.00	33.33	50.00	52.70	–	50.00	–	–	–	86.00	46.73	–	
Activity type	**Estimation accuracy for setting II (96 bins, each of duration 15 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	9.71	28.09	23.78	6.63	6.57	9.72	22.64	7.16	26.42	37.96	12.63	26.14	***Mean accuracy 33.25%***
Home leisure	24.07	18.44	62.50	19.44	36.88	17.43	38.47	12.23	19.62	–	12.70	–	
Night sleep	43.65	38.29	38.35	48.61	40.83	49.07	35.29	51.85	33.60	46.80	30.90	37.90	
Workplace	8.27	–	26.67	–	–	18.75	35.67	–	–	66.67	21.04	46.30	
Shopping & services	–	16.24	42.56	12.34	50.00	21.52	–	12.34	8.12	–	–	66.67	
School/college	–	–	–	–	–	–	–	–	–	–	–	30.00	
Organizational/Hobbies	14.29	18.65	21.30	60.00	50.00	27.27	27.50	11.11	19.64	–	66.67	33.33	
Entertainment	16.67	11.11	20.03	25.00	65.70	49.80	33.33	–	62.50	–	37.50	50.00	
Sports	78.60	23.81	66.67	85.60	–	–	33.33	33.33	–	67.30	47.86	50.00	
Activity type	**Estimation accuracy for setting III (48 bins, each of duration 30 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	19.96	21.91	25.16	10.90	9.99	11.73	25.70	12.99	19.82	37.96	20.39	40.13	***Mean accuracy 36.28%***
Home leisure	25.00	16.78	30.80	14.33	25.62	19.91	28.50	17.90	25.62	62.50	26.39	–	
Night sleep	47.02	39.55	56.25	62.50	41.88	51.39	45.29	41.67	50.30	88.00	35.30	76.00	
Workplace	16.40	–	30.67	–	–	20.85	34.15	–	–	55.56	23.42	69.90	
Shopping & services	25.00	20.89	34.46	15.99	27.04	18.05	25.00	16.47	19.18	–	14.29	66.67	
School/college	–	–	–	–	–	–	–	–	–	–	–	42.50	
Organizational/Hobbies	14.48	20.55	25.00	44.76	50.00	45.96	42.50	27.04	29.76	–	36.50	–	
Entertainment	16.67	33.33	17.55	17.14	–	57.14	27.78	–	58.33	–	43.75	50.00	
Sports	72.50	32.14	50.00	52.78	–	41.67	66.67	26.11	–	79.60	49.32	50.00	
Activity type	**Estimation accuracy for setting IV (8 bins, each of duration 180 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	43.90	43.39	51.53	37.73	36.27	41.52	52.04	38.03	47.72	45.12	55.33	62.05	***Mean accuracy 60.10%***
Home leisure	43.33	33.73	56.14	36.14	39.28	48.21	59.67	49.15	46.02	43.75	44.68	77.78	
Night sleep	68.06	67.64	48.10	43.75	42.76	57.78	49.66	63.89	64.48	76.67	44.83	80.00	
Workplace	35.12	100.0	51.02	90.00	83.33	59.49	47.36	90.00	75.00	71.67	47.02	100.0	
Shopping & services	44.58	40.86	53.35	35.30	33.75	49.98	51.19	54.49	49.46	52.38	54.56	86.11	
School/college	–	–	–	–	–	100.0	–	100.0	100.0	–	100.0	57.92	
Organizational/Hobbies	48.15	50.78	59.03	50.63	43.60	56.43	57.60	54.49	50.64	83.33	63.33	63.33	
Entertainment	37.04	50.83	52.64	45.15	77.78	55.36	53.33	55.21	71.90	50.00	51.48	66.67	
Sports	53.81	48.21	46.39	59.26	50.00	59.52	56.67	57.78	78.33	100.0	57.14	80.00	

'–' indicates that algorithm didn't predict start time due to missing the particular activity type in model's input.

Table 5
Percentage Accuracy of Activity Duration Estimation for Test Dataset.

Activity type	Estimation accuracy for setting I (96 bins, each of duration 15 min)												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	28.60	16.65	29.70	13.06	13.31	20.54	26.54	16.48	16.85	27.94	33.64	31.22	***Mean accuracy 42.86%***
Home leisure	29.05	19.47	20.58	15.92	11.15	31.30	38.36	16.29	15.71	41.67	20.85	61.11	
Night sleep	19.40	15.03	24.25	11.91	13.93	18.29	24.74	15.43	10.03	10.53	18.76	66.45	
Workplace	9.59	–	12.50	75.00	50.00	31.56	31.06	83.33	67.60	–	19.86	79.60	
Shopping & services	87.22	36.52	43.74	30.93	48.47	49.07	63.89	44.91	39.65	58.33	48.57	77.78	
School/college	–	85.00	–	75.00	–	–	–	57.60	–	–	–	12.50	
Organizational/Hobbies	39.32	43.23	51.04	45.32	35.60	55.12	52.74	48.62	36.35	50.00	58.61	61.11	
Entertainment	28.70	37.22	43.95	41.44	59.00	32.54	29.17	62.50	52.33	65.30	55.28	50.00	
Sports	58.89	49.40	77.78	58.33	79.17	–	88.89	75.00	30.67	75.90	42.86	76.80	
Activity type	**Estimation accuracy for setting II (48 bins, each of duration 30 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	35.57	33.70	45.18	28.98	21.97	36.16	39.16	32.45	33.21	35.87	46.80	41.84	***Mean accuracy 52.33%***
Home leisure	35.83	36.13	32.65	29.36	29.52	32.94	41.29	35.57	27.82	52.08	42.65	73.96	
Night sleep	22.49	17.55	35.43	17.53	21.90	25.77	29.22	44.44	12.91	26.32	27.55	69.08	
Workplace	12.55	49.60	17.30	75.00	97.60	35.87	41.88	87.50	75.00	33.33	17.37	79.60	
Shopping & services	80.71	58.67	66.00	55.08	70.02	65.73	70.83	61.01	61.23	66.67	72.17	83.33	
School/college	–	68.70	76.60	–	–	76.80	–	72.30	–	–	–	12.50	
Organizational/Hobbies	49.79	53.37	55.56	48.07	68.27	65.76	76.02	69.77	53.05	83.33	74.69	64.29	
Entertainment	36.57	26.48	43.81	38.70	80.56	49.05	61.90	61.11	57.71	84.60	66.46	75.00	
Sports	72.78	40.48	68.33	47.92	80.00	68.75	66.67	48.67	55.74	75.60	60.54	89.60	
Activity type	**Estimation accuracy for setting III (24 bins, each of duration 60 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	59.05	50.12	68.60	50.33	42.55	53.40	63.89	49.65	50.96	57.47	64.19	50.56	***Mean accuracy 67.32%***
Home leisure	68.64	56.03	64.91	54.43	62.38	54.64	62.27	51.95	58.07	56.25	54.33	77.27	
Night sleep	20.19	23.73	39.83	21.93	35.75	32.41	46.42	43.67	52.82	35.53	39.23	69.08	
Workplace	17.47	90.00	22.91	95.83	76.80	42.58	40.02	79.00	88.89	66.67	22.15	96.00	
Shopping & services	85.00	83.53	82.29	86.53	87.79	76.82	96.88	76.04	89.29	79.17	84.05	94.44	
School/college	–	89.60	75.60	97.60	–	73.60	88.30	76.80	–	–	94.00	25.00	
Organizational/Hobbies	65.19	63.76	62.04	65.08	77.46	84.48	93.00	85.88	67.96	92.86	93.00	75.00	
Entertainment	68.65	76.11	67.42	65.55	79.30	63.69	70.37	72.22	77.22	90.00	80.00	66.67	
Sports	89.52	64.29	68.70	58.33	80.56	80.00	90.48	75.83	80.19	87.60	71.31	87.50	
Activity type	**Estimation accuracy for setting IV (4 bins, each of duration 360 min)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	99.15	100.0	100.0	97.75	95.60	99.26	96.15	100.0	100.0	100.0	100.0	100.0	***Mean accuracy 98.65%***
Home leisure	98.77	100.0	98.70	99.57	100.0	98.38	99.29	100.0	100.0	100.0	100.0	100.0	
Night sleep	89.65	100.0	87.90	96.47	91.42	92.53	68.38	100.0	100.0	100.0	100.0	100.0	
Workplace	76.11	100.0	91.85	100.0	100.0	93.40	90.71	100.0	100.0	100.0	100.0	100.0	
Shopping & services	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
School/college	–	100.0	100.0	100.0	–	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Organizational/Hobbies	96.88	100.0	100.0	99.04	100.0	100.0	97.50	100.0	100.0	100.0	100.0	100.0	
Entertainment	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Sports	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	

'–' indicates that algorithm didn't predict activity duration due to missing the particular activity type in model's input.

therefore reveal that the RF model can predict response variables with more precision when trained with a larger dataset. The mean estimation error for all 12 clusters in our model is 18.38% in the 24-hour period. Further comparison regarding the model performance can be made on the level of the algorithm/method used where other techniques such as CHAID and AdaBoost algorithms are used instead of the RF algorithm on the same dataset.

Compared to the observed temporal patterns, the algorithm could mostly re-assemble different activities of travelers in each cluster. However, in cases where the dominant activity (e.g. work) occupied a large portion of the traveler's daily activity pattern, the accuracy level for scheduling activities with smaller durations decreased. Improving the performance of the heuristic rule-based algorithm with more insertion and adjustment constraints for scheduling activities with shorter durations is recommended for future studies.

6. Conclusions

Complexities in activity-travel behavior of population groups in the study region vary according to their socio-demographic and socio-economic characteristics. For instance, homemakers and retirees have lower variances in time expenditure choices compared to worker and student groups. Accordingly, the best approach is to predict or model travel behavior for a set of model individuals, who represent homogeneous cohorts. The significant original contribution of this study is to develop a new modeling framework that is able to learn and predict temporal attributes of activities for use in activity-based travel demand models. We modeled the weekday in-home and out-of-home activity temporal features of 12 clusters containing individuals with homogeneous activity patterns, drawn from the large Halifax STAR travel diary survey. Activity start time and activity duration for every activity type were allocated to a set of bins. With respect to the pattern complexity of activity sequences and sample size of person-days in the clusters, eight different bin structures, varying in the time interval, were considered. The model was trained with 70% of the dataset and the remaining 30% was used for testing the model performance.

In the modeling framework proposed in this study, activity agendas and activity sequences of travelers are predicted using an advanced Random Forest (RF) algorithm. The algorithm comprises numerous prediction decision trees, in which each tree plays as a weak learner and is able to make a prediction. The aggregation of these weak inputs provides a powerful ensemble learning model. A final prediction result is obtained from the majority votes obtained from each ensemble tree. Independent variables were selected from the socio-demographic characteristics of travelers and the corresponding start times or duration bin numbers for each activity type in the agenda. Bin numbers were defined as the response variable in the RF model. Consequently, for each activity type, activity start time and activity duration were generated from a uniform distribution within their interval time range of predicted bin numbers. In the next step, activities were inserted into the skeleton schedule using a heuristic decision rule-based algorithm, and a 24 h schedule was constructed with respect to two constraints: the importance level of the activity established from the cluster's representative pattern characteristics, and the duration of non-mandatory activities.

The estimation accuracy of the proposed RF model was examined under different bin settings. The best estimation results for predicting activity start time were found for setting IV (8 bins, each of duration 180 min), 60.10%, followed by setting III (48 bins, each of duration 30 min), 36.28%. Similarly, the best model estimation results for predicting activity durations were found for setting $\widehat{\text{IV}}$ (4 bins, each of duration 360 min), 98.65%, followed by setting $\widehat{\text{III}}$ (24 bins, each of duration 60 min), 67.32%. Results show that the proposed model is able to assemble the traveler's schedule with an average 81.62% accuracy in the 24-hour period. When the RF model is trained with a larger dataset, it is expected to predict response variables with more precision. As a follow-up experiment it would be an idea to use unequal intervals based on Tables 4 and 6 (i.e. a hybrid structure of I-IV). This way accuracy could be further improved. (e.g. time periods could be longer during the night than during the day).

Numerous aspects of temporal information on activities, such as activity start time, activity duration, and activity end time, can be predicted for various population groups with various activity sequence patterns. Such precise information is essential for the scheduling phase of activity-based travel demand modeling. The proposed method improves on previous methods, and provides more accurate temporal information, especially for individuals with high pattern complexity of activity sequences. Compared to other decision tree techniques such as boosted decision trees, the proposed RF model in this study is able to automatically handle missing values in the algorithm. Furthermore, variables do not need to be transformed, very few parameters need to be adjusted, and the algorithm does not overfit easily. However, compared to conventional hypothesis-driven approaches which yield interpretable results, such as multinomial logit regression, most machine learning based approaches are designed as a black box. The model is trained to evaluate the labels of the data and its results are a set of support data-points and their respective weights. This is potentially problematic if the intention is to understand how elements of the activity-travel system interact, but it is not an issue if the purpose is simply accurate prediction. In addition, machine learning based approaches are very efficient in computational time, with a high degree of reproducibility. The methods employed in this study can also be adapted for modeling other components of activity-based travel demand models, such as transport mode, and work and residential location choice models.

To build on this study and further demonstrate the potential of our proposed method, we are proposing several avenues of research. Firstly, it is possible to integrate the proposed model with a dynamic traffic assignment model and increase the model's ability for use in rescheduling activities. This will require updating the algorithm with new data on congested travel times. Secondly, and in line with growing worldwide interest in developing activity-based travel demand model at the household level, we aim to explicitly model intra-household interactions using the proposed modeling framework in this study. Thirdly, the large STAR survey data that has been used for building the RF model in this paper includes business hours survey data. Therefore, one potential extension would be to include operation hours in the modeling process.

In the current modeling framework, the median stop number is used to predict the number of activities within each cluster. For

Table 6
Mean Scheduling Error for Test Dataset (Duration of Misclassification*).

Activity type	Mean estimation error (minutes)											
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12
Home chores	24.44	56.30	48.02	86.96	105.79	32.03	4.69	86.36	61.81	61.50	32.91	45.64
Home leisure	118.24	31.15	29.09	15.57	69.49	70.54	53.38	2.48	52.30	15.20	50.81	231.62
Night sleep	49.31	54.92	20.76	14.93	30.02	44.04	55.30	28.28	31.18	174.83	28.22	186.85
Workplace	18.08	2.09	164.05	10.43	0.36	58.99	134.91	3.38	1.21	101.59	60.35	8.47
Shopping & services	11.29	48.63	12.28	8.46	2.21	0.35	7.60	20.01	3.44	4.11	0.40	10.02
School/college	–	0.61	1.13	0.52	0.04	4.16	1.69	4.61	5.60	3.81	0.56	13.71
Organizational/hobbies	16.40	26.55	7.85	36.26	2.44	3.71	2.39	9.30	7.77	0.87	2.24	13.31
Entertainment	32.72	1.69	28.10	54.13	2.28	3.65	7.19	5.50	6.39	1.21	3.89	18.51
Sports	14.89	3.28	19.08	7.65	0.24	4.21	12.05	12.79	2.74	4.12	10.03	0.46
Total error in 24-h	285.38	225.22	330.37	234.91	212.86	221.69	279.20	172.71	172.43	367.23	189.43	528.60
Mean estimation error (%)	19.82	15.64	22.94	16.31	14.78	15.40	19.39	11.99	11.97	25.50	13.15	36.71

* Each cell with the length of 5 min.

some travelers, this assumption may be inappropriate, and hence the variation in number of activities cannot be accurately predicted. Further investigation should be made on this aspect of the scheduling process. The RF model presented in this study predicted activities with shorter durations with lower estimation accuracy compared to activities with longer durations. Therefore, improvement in the model structure for predicting and scheduling activities with shorter durations remains an area of future investigation. Furthermore, since behavior of some predictor variables might be similar and somewhat difficult to interpret, further study might combine some of the variable classes together such as widowed, separated, and divorced, or married, living common-law.

In summary, the modeling framework presented in this study yields a straightforward and easy-to-implement tool for urban and transport modelers to predict and model temporal information of activities for various population groups within a region. The results of this study are expected to be implemented within the activity-based travel demand model for Halifax, Nova Scotia, Scheduler for Activities, Locations, and Travel (SALT).

CRediT authorship contribution statement

Mohammad Hesam Hafezi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing - original draft, Writing - review & editing. **Naznin Sultana Daisy:** Data curation, Investigation, Validation, Writing - review & editing. **Hugh Millward:** Data curation, Resources, Validation, Writing - review & editing. **Lei Liu:** Funding acquisition, Resources, Writing - review & editing.

Declaration of Competing Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgments

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and DFO-MPRI for their contribution in supporting the research. We also wish to thank the Dalhousie Transportation and Environmental Simulation Studies (TESS) group members for their valuable suggestions. Data for this research were provided by the Halifax STAR Project, supported through the Atlantic Innovation Fund from the Atlantic Canada Opportunities Agency, Project No.181930. We would also like to thank anonymous reviewers for their very useful comments and suggestions.

References

Allahviranloo, M., Recker, W., 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transport. Res. Part B: Methodol. 58, 16–43.
Allahviranloo, M., Regue, R., Recker, W., 2017. Modeling the activity profiles of a population. Transportmetr. B: Transport Dyn. 5 (4), 426–449.
Arentze, T., Timmermans, H., 2000. ALBATROSS: a learning based transportation oriented simulation system. European Institute of Retailing and Services Studies (EIRASS), Technische Universiteit Eindhoven, Netherlands.
Arentze, T.A., Timmermans, H.J., 2004. A learning-based transportation oriented simulation system. Transport. Res. Part B: Methodol. 38 (7), 613–633.
Auld, J., Mohammadian, A., Doherty, S.T., 2009. Modeling activity conflict resolution strategies using scheduling process data. Transport. Res. Part A: Policy Pract. 43 (4), 386–400.
Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. Transportat. Lett. 1 (3), 245–255.
Ben-Akiva, M.E., Bowman, J.L., 1998. Activity Based Travel Demand Model Systems. In: Marcotte, P., Nguyen, S. (Eds.), Equilibrium and Advanced Transportation Modeling. Springer US: Boston, MA. p. 27–24.
Bhat, C., Guo, J., Srinivasan, S., Sivakumar, A., 1894. Comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record: Journal of the Transportation Research Board. No. 1894. Transportation Research Board of the National Academies, Washington, D.C., pp. 57-66.
Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25 (2), 197–227.
Breiman, L., 2001. Random Forests. Machine Learn. 45 (1), 5–32.
De Palma, A., Lindsey, R., Quinet, E., Vickerman, R., 2011. A Handbook of Transport Economics. Edward Elgar.
Drchal, J., Čertický, M., Jakob, M., 2019. Data-driven activity scheduler for agent-based mobility models. Transport. Res. Part C: Emerg. Technol. 98, 370–390.
Daisy, N.D., Liu, L., Millward, H., 2018a. Trip chaining propensity and tour mode choice of out-of-home workers: evidence from a mid-sized Canadian city. Transportation 47, 763–792.
Daisy, N.D., Millward, H., Liu, L., 2018b. Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. J. Transport Geogr. 69, 150–162.
Daisy, N.D., Millward, H., Liu, L., 2018c. IIndividuals' Activity-Travel Behavior in Travel Demand Models: A Review of Recent Progress. Intelligence, Connectivity, and Mobility. American Society of Civil Engineers (ASCE) 2615–2625.
Daisy, N.D., Hafezi, M.H., Liu, L., Millward, H., 2018d. Understanding and Modeling the Activity-Travel Behavior of University Commuters at a Large Canadian University. J. Urban Plann. Develop. 144 (2).
Daisy, N.D., Millward, H., Liu, L., 2020. Modeling activity-travel behavior of non-workers grouped by their daily activity patterns. Mapping the Travel Behavior Genome 339–370.
Ettema, D., Borgers, A., Timmermans, H., 1993. Simulation model of activity scheduling behavior. In: Transportation Research Record: Journal of the Transportation Research Board. No. 1413, Transportation Research Board of the National Academies, Washington, D.C., pp. 1-11.
Garling, T., Kwan, M.-P., Golledge, R.G., 1994. Computational-process modeling of household activity scheduling. Transport. Res. Part B: Methodol. 28 (5), 355–364.
Goran, J., 2001. Activity based travel demand modeling-a literature study. Technical report, Danmarks Transport-Forskning.
Hafezi, M.H., Liu, L., Millward, H., 2017. Learning Daily Activity Sequences of Population Groups using Random Forest Theory. Transport. Res. Record: J. Transport. Res. Board 2672 (47), 194–207.
Hafezi, M.H., Liu, L., Millward, H., 2017. Identification of Representative Patterns of Time Use Activity Through Fuzzy C-means Clustering. Transportation Research Record. J. Transport. Res. Board 2668, 38–50.
Hafezi, M.H., Millward, H., Liu, L., 2018. Activity-based travel demand modeling: Progress and possibilities. Planning, Sustainability, and Infrastructure Systems. American Society of Civil Engineers (ASCE) 4, 138–147.

Hafezi, M.H., Liu, L., Millward, H., 2018. Modeling Activity Scheduling Behavior of Travelers for Activity-Based Travel Demand Models. Presented at the Transportation Research Board 97th Annual Meeting Transportation Research Board. 18-05680.
Hafezi, M.H., Daisy, N.D., Liu, L., Millward, H., 2018. Daily activity and travel sequences of students, faculty and staff at a large Canadian university. Transport. Plann. Technol. 41 (5), 536–556.
Hagerstrand, T., 1970. What about people in regional science? Papers Regional Sci. 24 (1), 7–24.
Hayes-Roth, B., Hayes-Roth, F., 1979. A Cognitive Model of Planning. Cognit. Sci. 3 (4), 275–310.
Jiang, S., Ferreira, J., Gonzalez, M.C., 2012. Clustering daily patterns of human activities in the city. Data Min. Knowl. Discov. 25 (3), 478–510.
Jones, P.M., Dix, M.C., Clarke, M.I., Heggie, I.G., 1983. Understanding travel behavior. J. Forecast. 4, 315–316.
Kitamura, R., Chen, C., Pendyala, R., 1997. Generation of Synthetic Daily Activity-Travel Patterns. In: Transportation Research Record: Journal of the Transportation Research Board. No. 1607. Transportation Research Board of the National Academies, Washington, D.C., pp. 154-163.
Kitamura, R., Pas, E.I., Lula, C.V., Lawton, T.K., Benson, P.E., 1996. The sequenced activity mobility simulator (SAMS): an integrated approach to modeling transportation, land use and air quality. Transportation 23 (3), 267–291.
Kitamura, R., Chen, C., Pendyala, R.M., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation 27 (1), 25–51.
Li, S., Lee, D.-H., 2017. Learning daily activity patterns with probabilistic grammars. Transportation 44 (1), 49–68.
Liao, L., Patterson, D.J., Fox, D., Kautz, H., 2007. Learning and inferring transportation routines. Artif. Intell. 171 (5), 311–331.
Liu, F., Janssens, D., Cui, J., Wets, G., Cools, M., 2015. Characterizing activity sequences using profile hidden Markov models. Expert Syst. Appl. 42 (13), 5705–5722.
Miller, E., Roorda, M., 2003. Prototype model of household activity-travel scheduling. Transportation Research Record: Journal of the Transportation Research Board. No. 1831. Transportation Research Board of the National Academies, Washington, D.C., pp. 114–121.
Millward, H., Hafezi, M.H., Daisy, N.D., 2020. Activity travel of population segments grouped by daily time-use: GPS tracking in Halifax. Canada Travel Behaviour and Society 16, 161–170.
Oberkampf, W.L., DeLand, S.M., Rutherford, B.M., Diegert, K.V., Alvin, K.F., 2002. Error and uncertainty in modeling and simulation. Reliab. Eng. Syst. Saf. 75 (3), 333–357.
Rasouli, S., Timmermans, H., 2012. Uncertainty in travel demand forecasting models: literature review and research agenda. Transport. Lett. 4 (1), 55–73.
Recker, W.W., McNally, M.G., Root, G.S., 1986. A model of complex travel behavior: Part II-An operational model. Transport. Res. Part A: General 20 (4), 319–330.
Recker, W.W., McNally, M.G., Root, G.S., 1986. A model of complex travel behavior: Part I-Theoretical development. Transport. Res. Part A: General 20 (4), 307–318.
Roorda, M.J., Miller, E.J., Habib, K.M.N., 2008. Validation of TASHA: A 24-h activity scheduling microsimulation model. Transport. Res. Part A: Policy Pract. 42 (2), 360–375.
Shafique, M.A., Hato, E., 2015. Use of acceleration data for transportation mode prediction. Transportation 42 (1), 163–188.
Suthaharan, S., 2015. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Springer, US.
Timmermans, H.J.P., Zhang, J., 2009. Modeling household activity travel behavior: Examples of state of the art modeling approaches and research agenda. Transport. Res. Part B: Methodol. 43 (2), 187–190.
TURP, Turp (Time Use Research Program), 2008. Halifax regional space time activity research (STAR) survey: A GPS-assisted household time-use survey, survey methods. Saint Mary's University, Halifax.
You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. J. Modern Transport. 25 (2), 116–123.
Zhang, A., Kang, J.E., Axhausen, K., Kwon, C., 2018. Multi-day activity-travel pattern sampling based on single-day data. Transport. Res. Part C: Emerg. Technol. 89, 96–112.