



Microscopic activity sequence generation: a multiple correspondence analysis to explain travel behavior based on socio-demographic person attributes

Usman Ahmed¹ · Ana Tsui Moreno¹ · Rolf Moeckel¹

© The Author(s) 2020

Abstract

Activity sequencing is a crucial component of disaggregate modeling approaches. This paper presents a methodology to analyse and predict activity sequence patterns for persons based on their socio-demographic attributes. The model is developed using household travel survey data from Germany. The presented method proposes an efficient approach to replace complex activity-scheduling modules in activity-based models. First, the paper describes a multiple correspondence analysis technique to identify the correlation between activity sequence patterns and socio-demographic attributes. Secondly, a probabilistic model is developed, which could predict likely activity sequence patterns for an agent based on the results of the multiple correspondence analysis. The model is predicting activity sequence patterns fairly accurately. For example, the activity sequence pattern home–work–home is well predicted ($R^2 = 0.99$) for all the workers, and the activity sequence pattern home–education–home is rather well predicted ($R^2 = 0.90$) for students. The model predicts the 112 most common activity sequence patterns reasonably well, which covers 72% of all activity sequence patterns observed.

Keywords Activity sequencing · Activity sequence patterns · Activity-based modeling · Multiple correspondence analysis

Introduction

Since decades, urban travel forecasting models have played a crucial role in supporting infrastructure investment decisions. The majority of these models are based on the classical four-step modeling approach (Mcnally 2007). The mobility patterns have become more

✉ Usman Ahmed
um.ahmed@mail.utoronto.ca

Ana Tsui Moreno
ana.moreno@tum.de

Rolf Moeckel
rolf.moeckel@tum.de

¹ Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany

complex, which motivated planners to build more sophisticated models. This results in the development of disaggregate modeling approaches known as activity-based models. The rationale for this approach is that people travel to fulfil the demand to conduct activities at different locations. While the trip-based modeling approach is simpler to implement than activity-based models, it poorly handles activity sequence patterns by disconnected non-home based trips (Adler and Ben-Akiva 1979; Axhausen and Gärling 1992).

One of the components of activity-based modelling is to build activity sequences of people. This paper presents a new method to generate activity sequence patterns for activity-based modeling frameworks. Compared to traditional activity sequencing procedures that often require to model separately mandatory activity generation, discretionary activity generation and scheduling the order of these activities, the proposed approach is straightforward to implement and runs much faster. In the literature, the definition of activity sequence patterns vary. For example, Thomas F. Golob (2000) defines activity sequence patterns as simple and complex, such as complex home-based trip chains. The current study defines an activity sequence pattern as a sequence of activities performed during a day (e.g. home to work to shopping to home), regardless of the duration and location of activities.

State of the art of activity sequencing

The crucial component of activity-based modeling is the generation of the complete activity diary of an individual or a household unit. This mainly involves three sub-components. The first is the type of activity and its sequential arrangement; the second is the destination/location of activities and the third is the allocation of start time, end time and duration of the activity (Charypar and Nagel 2005). The current research is focused upon the first component: i.e. type and sequence of activities, known as activity sequencing or activity pattern generation model. The problem of activity sequencing is solved using different modeling methodologies and methods. In some studies, activity sequencing is modeled integrated with activity destination or activity time allocations or both. Some studies model activity sequence independently. The methods used can be broadly classified as discrete choice-based methods, machine learning-based methods and other methods.

One of the early works was done by Pas (1984) who studied the effect of sociodemographic characteristics on activity-travel patterns. He applied cluster analysis to identify five travel-activity patterns. He then used a linear logit model to analyse a contingency table to examine the relationship between activity sequences and socio-demographic variables. Recker et al. (1986a, b) provided a theoretical framework to model complex travel behavior. Individuals select an activity pattern from the choice set that maximizes their utility. The concept was later operationalized in the STARCHILD modeling system by Recker et al. (1986b). Bowman and Ben-Akiva (2001) estimated a nested logit model for daily activity pattern choice as part of the integrated activity-based model system. An activity pattern was defined by the primary activity, primary tour type and number and purpose of secondary tours. A choice set of 54 activity patterns were defined. The Boston household survey was used to estimate the model. Another study by Bhat (2001) modeled activity patterns of workers during evening hours using a multinomial logit model. The stops, activity duration and travel time deviations were modeled explicitly. Gangrade et al. (2002) used a nested logit model to model activity sequences

only for commuters. Poisson and negative binomial regression models were used to predict activity frequencies for each individual, while a simple heuristic rule was applied to determine the sequence of activities. Ordóñez Medina (2015) focused on sequencing of activities, their start and end times, duration of activities, location of activities and mode used to perform activities. A spatio-temporal network method was used to form activity sequences while utility functions were used to select an activity sequence. Habib et al. (2017) studied activity-travel patterns of non-workers using a random utility maximizing approach in sequential nested structures. Activity type, destination location and time expenditure choice were modeled. Activities were sequentially added to form activity-travel behaviour for non-workers. They explicitly considered travel time budget in the formation of activity-travel pattern. Paleti et al. (2017) developed a rank-ordered logit model to model activity sequence, location and tour structure in an integrated manner. The alternatives in the choice set were defined as the combination of activity type and location. Recently, Xu et al. (2018) used a combination of mixed integer linear programming (MILP) and random utility maximization (RUM) architecture to model activity patterns. The household activity pattern problem (HAPP) structure was used as an optimization function and parameters were estimated using path-size logit models. The choice set of each traveller was generated and individualized using clustering algorithm and goal-programming based on the individual's spatial and temporal constraints.

Several studies used machine learning algorithms to analyse travel behaviour. Those include Charypar and Nagel (2005) who applied a genetic algorithm to model activity schedules of individuals. Activity type, time allocations and location of activity was part of the activity schedule. A utility function was developed that was used as a fitness function to compare and choose the best activity schedule. Hanson and Huff (1986) used principle component analysis (PCA) and k-means clustering algorithms to classify individuals based on the multi-day travel-activity patterns. The analysis yields five clusters that were distinguished by discriminant analysis based on socio-demographic variables. PCA was also used by Pitombo et al. (2009) to reduce the dimension of the data from 39 to 10 to suggest a taxonomy of travellers regarding their travel behaviour. Multiple linear regression analysis was used to model different travel patterns using the reduced ten dimensions as independent variables. Diana and Pronello (2010) used correspondence analysis to segment travellers' profiles regarding their stated mode choices. Age, person type and attitude towards transit choice were among the variables used for analysis. Jiang et al. (2012) used the k-means clustering via PCA to cluster daily activity patterns on weekdays and weekends. PCA was used to obtain eigen-activities. This way, a thorough understanding of the trip behaviour could be deduced. Hafezi et al. (2018) used random forest theory to model daily activity patterns. Socio-demographic and prior activities were used as predictor variables. The model has an activity sequence replication accuracy of up to 77%. Hafezi et al. (2019) developed a pattern recognition modeling framework to model activity schedules. A fuzzy C-means clustering algorithm was used to cluster data into twelve clusters of homogeneous activity schedules. The schedule is defined by activity type, start time, end time, duration and sequential arrangement of activities. Individuals were assigned to particular clusters based on their socio-demographic attributes. Recently, Drchal et al. (2019) introduced the concept of a data-driven activity scheduler using machine learning methods. The schedule consists of activity type, duration, location and mode choice. Four core modules were developed to model each component of the schedule in a sequential manner.

Other methods include the nonlinear canonical correlation analysis by Golob (1986) to study the relationship between home-based trip chains and socioeconomic characteristics of the person without considering the spatial and temporal aspect of activities. Travellers

were categorized into the three segments workers, students and others. He found various explanatory variables for each segment that effects trip-chaining behaviour. Kitamura et al. (2000) used a microsimulation approach to develop travel patterns. Monte Carlo simulation was used to generate daily travel patterns. The study showed that it is possible to generate synthetic travel patterns through microsimulation. Han and Sohn (2016) used continuous hidden Markov models while sequence alignment methods are used by Kim (2014) and Sammour et al. (2012). Li and Lee (2017) used probabilistic grammars to model daily activity patterns.

The most common problem with behavioural choice modeling is the determination of choice sets of alternatives. The possible sequence of activities could result in huge number of alternatives in the choice set. On one hand, the high number of alternatives may not be suitable to use discrete choice models, especially if activity sequences are combined with activity locations. On the other hand, some of the resulting combinations may not be representative of travel behaviour and not observed in household travel surveys. Moreover, the problem with machine learning algorithms is that they are designed as a black box and are computationally expensive. Therefore, the current study develops a methodology that considers complete activity sequence patterns of individuals based on their socio-demographic attributes. Further, among the most favourable alternatives for an individual, preference is given to those alternatives that are more frequently selected in the household travel survey.

Data and modeling framework

This chapter explains the data analysis and modeling framework used in this study. A two-step method is used to build the model. In the first part, activity sequence patterns are analyzed and a relationship between activity sequence patterns and socio-demographic variables is established. In the second step, a prediction framework is developed that uses the relationships defined in the first step to calculate the probabilities of all activity sequence patterns for a given person. Overall, the modeling framework is a non-home activity focused model. The home is not simulated as a location of activities but just a place to return. The activity sequences defined does not necessary start and end at home but rather could be incomplete as represented in the survey data for a 24-h period.

The activity sequence patterns are modeled at person level instead of household level. There is a debate in the literature as to which analysis unit is better, person-level of analysis or household-level of analysis. Some studies (Downes et al. 1978; Supernak et al. 1983) support a person-level analysis, because most activities are conducted by individual persons (joint activities form an exception). Others (Badoe and Chen 2004) supports a household-level analysis, because activities of one household member (such as, pick up the child or do grocery shopping) will affect the need to perform these activities of other household members. Pokhrel (2017) compared both approaches econometrically and found no conclusive recommendation which approach would perform better. The current study modeled activity sequence patterns at the person level because the methodology used for model building can provide better results at the person level than at the household level as the unit of analysis. The person level can provide a greater heterogeneity in activity sequencing behaviour. As intra-household interactions are not modeled, accompanying trips are modeled as an activity. If a member of the household is accompanying another member, this type of activity sequence pattern is modeled, and the activity is termed as ‘accompanying’. For example, there is an activity sequence pattern as ‘home-accompanying-home’ used in

the model. However, the link to the other household member that the person is accompanying with is not modeled. From our household travel survey, it was unfortunately not possible to identify whether the person is accompanying a household member or some other person.

Generation of activity sequence patterns

The German household travel survey for the year 2008 called ‘Mobilität in Deutschland (MiD 2008)’, or ‘Mobility in Germany’, was used as a primary data source for this model. The MiD 2008 contains travel data reported in the form of individual trips performed by the household members. For every household, travel data of each member of the household is collected on the same day (Cyganski et al. 2013). This study considers seven types of activities for the generation of activity sequence patterns. They are “Home (H)”, “Work (W)”, “Education (E)”, “Shop (S)”, “Leisure (L)”, “Private errands (PE)” and “Accompanying (AC)”. A total of 5683 unique activity sequence patterns were generated from 45,746 individuals. Each individual corresponds to one activity sequence pattern and some patterns are also repeated. For instance, 4467 individuals performed the same activity sequence pattern as ‘home-work-home’ (Table 1) on the survey day. The top 5 activity sequence patterns are shown in Table 1.

A detailed explanation of the activity sequence generation process is described by Ahmed (2018). These activity participations are reported for a single day; therefore, activity sequence patterns also represent a day which is assumed to be an average weekday.

Activity sequence pattern analysis

To study the relationship between activity sequence patterns and socio-demographic attributes, the generated activity sequence patterns were checked for missing values of socio-demographic attributes. Person records with missing value of any of the attribute was omitted. This reduced the number of individuals to 39,080 with 5136 activity sequence patterns. Most of the activity sequence pattern types were rare types with very low number of survey records. For multiple correspondence analysis, it is important to have a sufficiently large sample size for each category to obtain meaningful relationships. One approach to overcome this problem is to group categories (Husson et al. 2017, p. 150). However, it is impossible to decide whether, for example, H–W–S–H–L–H should be grouped with

Table 1 Activity sequence patterns observed

Rank	Activity sequence pattern	Frequency	% Share
1	H–W–H	4467	9.8%
2	H–L–H	4324	9.5%
3	H–S–H	2360	5.2%
4	H–E–H	2091	4.6%
5	H–PE–H	1524	3.3%
...			
5683	W–S–W–H	1	0.0%
	Total	45,746	100%

H home, *W* work, *E* education, *L* leisure, *S* Shop, *PE* private errands

H–W–S–H–S–H or H–W–S–H–PE–H. To avoid a manual (and arbitrary) grouping process, this problem is approached by selecting activity sequence patterns with 30 or more person records.

For household type definition in trip generation, it is state of practice to use 30 survey records as a threshold for each household type to be considered (Moeckel et al. 2017). This criterion was used in the current study and 112 activity sequence patterns were found with survey records of 30 and more. It is important to note that these 112 activity sequence patterns represent 72% of the dataset, as shown in Table 2. It can also be seen from Table 2 that criteria of 20 and above would add around 60 more activity sequence patterns while the dataset representation would only increase by 4%. The criteria of 10 and above would add 193 activity sequence patterns with dataset representation of 8% more.

The inclusion of more than 70% of the dataset with a reasonable number of activity sequence patterns was considered to be a reasonable representation of the most common activity sequencing behaviour. These 112 activity sequence patterns were generated by 28,172 individuals. For simplicity, each activity sequence pattern is coded as TC (TC1–TC112), where TC1 is the most frequent activity sequence pattern while TC112 is the least frequent activity sequence pattern. All 112 activity sequence patterns with respective codes are described by Ahmed (2018). The remaining 5024 of the activity sequence patterns were grouped together into a single group. The group can be assigned a probability value and the selection of activity sequence patterns from that group may be done randomly based on the frequency of activity sequence pattern. Given the infrequent occurrence of those activity sequence patterns and the lack of data to describe who commonly would conduct such rare activity sequence patterns, a random selection is considered to be plausible. For the 112 much more frequent activity sequence patterns, true probabilities that are based on socio-demographic attributes are calculated for every individual.

Socio-demographic attributes

As socio-demographic attributes, household characteristics and person characteristics were considered. Table 3 shows attributes included in this study and their frequency after omitting incomplete records. The column “Number of person records” shows the number of surveyed households or persons in the dataset. The column “Expanded number of records” describes how many households/persons are represented after the respective expansion factors of the survey were applied. The attributes in Table 3 were selected based on state-of-practice in transportation modeling and are assumed to explain activity sequencing behavior. All attributes used were transformed into categorical variables. To the best of the authors’ knowledge, there is no formal method in multiple correspondence analysis to

Table 2 Activity sequence pattern statistics

Minimum number of person records per activity sequence pattern	Activity sequence pattern	Individuals	Share of all activity sequence pattern records (%)
Total	5136	39,080	100
30 and above	112	28,172	72
20 and above	171	29,553	76
10 and above	305	31,379	80

Table 3 Attributes considered in the model

Attributes		Code	Category description	Number of person records	Expanded number of person records	
Household	Household size (number)	H ₁	1	2065	233,291	
		H ₂	2	9794	408,485	
		H ₃	3 or more	16,313	625,304	
	Household auto ownership (number)	C ₀	0	2175	150,971	
		C ₁	1	12,858	595,368	
		C ₂	2	10,349	423,089	
		C ₃	3 or more	2790	97,652	
	Monthly income (€/month)	I ₁	0 to <2000	7863	430,803	
		I ₂	2000 to <3000	9177	389,882	
		I ₃	3000 to <5000	8719	346,577	
		I ₄	> 5000	2473	99,818	
	Person	Gender	M	Male	14,381	642,140
			F	Female	13,791	624,940
Age (years)		A ₁	0 to 22	7215	311,476	
		A ₂	23 to 47	7498	434,664	
		A ₃	48 to 64	7943	280,120	
		A ₄	65 and above	5516	240,820	
Driving license holder		L	Yes	20,079	901,980	
		NL	No	8093	365,100	
Occupation		O ₁	Worker	13,059	610,850	
		O ₂	Student	5084	207,822	
	O ₃	Other	10,029	448,408		

detect which variables are statistically most significant. However, the contribution of each variable to the multiple correspondence analysis dimensions can give an idea of the most important categories which explains variability in the dataset.

Multiple correspondence analysis (MCA)

Methodology

Correspondence analysis (CA) and its extension Multiple Correspondence analysis (MCA) identifies relationships among variables in the survey data. While the more common Principal Component Analysis (PCA) is suited for metric data, the similar method MCA is used with categorical data. Multiple correspondence analysis is typically used to analyse surveys where individuals respond to questions with specific answers (Husson et al. 2017, p. 127). Commonly, rows represent respondents and columns represent variables. For multiple correspondence analysis, each question is considered as a variable and each answer is considered as a category of that variable (Husson et al. 2017).

This method reduces the dimensions of the data by creating new principal dimensions. Each dimension represents a certain amount of variance in the dataset. Each dimension

can also be interpreted as a synthetic variable (Husson et al. 2017, p.15), which is formed by the contribution of each variable category. Some dimensions may explain variability of a certain category more than other categories. The method finds the principal dimensions, which explain the variability between individuals and associations among variable categories. The dimensions are orthogonal to each other. Thus, instead of K dimensions for individuals and I dimensions for categories, the reduced number of dimension, i.e. principal dimensions, are formed. Thus the data can be plotted in two dimensions, which helps to study the relationships in the data.

For model estimation, the dataset was prepared in a way that each row represented an individual and each column represented socio-demographic attributes and an activity sequence pattern, as reported in the MiD 2008. These variables naturally fall into two groups, one defining the person and the other defining travel behaviour. The model uses socio-demographic attributes as active variables and activity sequence patterns as supplementary variables. The dimensions were formed using active variables, i.e. socio-demographic attributes, to establish the principal dimensions of the multiple correspondence analysis. Then, supplementary variables, i.e. activity sequence patterns, were added to show the relationship between the principal dimensions and supplementary variables.

MCA results

The model was estimated using the attributes defined in the section on socio-demographic attributes. The results of the model estimation are described below.

Dimensions

The model estimation resulted in the formation of 15 principal dimensions as shown in Fig. 1 below:

The number of dimensions to be formed is estimated by the multiple correspondence analysis method itself. Each dimension explains a certain amount of variance in the dataset. For example, dimension 1 explains 18.5% of the variance in the dataset, while dimension 2 explains 15.4% of the variance in dataset. Every consecutive dimension explains less variance than the previous one, and all 15 dimensions together can explain 100% of the variance.

Socio-demographic attributes

Each variable category may be represented in all the 15 dimensions. Table 4 shows the estimation results for the first two dimensions. The columns 'Dimension 1' and 'Dimension 2' show the coordinate points of each category for the first and second dimension. The coordinates describe for each dimension a relative position a given category has in comparison to other categories. These coordinates can be used to visualize results (compare Fig. 2) and to calculate the most likely activity sequence pattern. The column 'contribution' represents the contribution of each category in the formation of the respective dimension. For instance, A_1 (age between 0 and 15) contributes the most in the definition of dimension 1 with the share of more than 15%, while O_1 (employed), contributes the least in the formation of dimension 1 with the share of only 0.032%. The column v-test is a

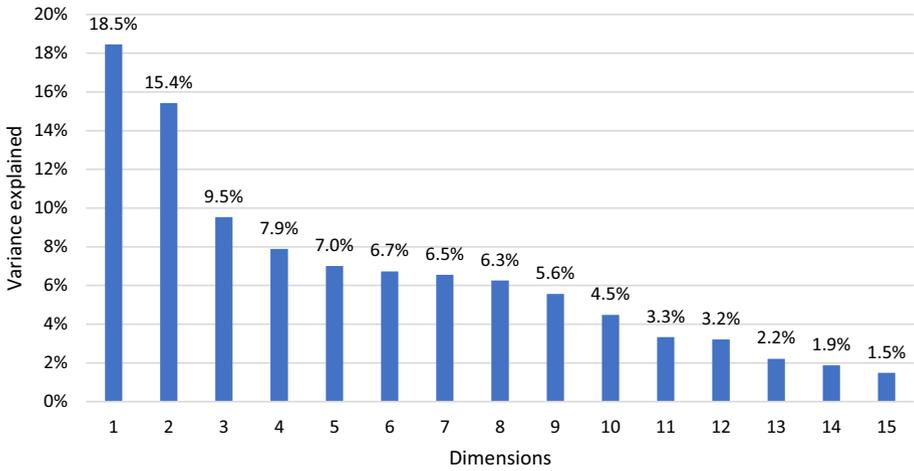


Fig. 1 Variable decomposition per dimension

Table 4 Results of categories in first two dimensions

Category	Dimension 1				Dimension 2			
	Coordinate	Contribution	V-test	Cosine ²	Coordinate	Contribution	V-test	Cosine ²
H ₁	- 1.14	8.71	- 611.95	0.30	0.46	1.71	247.72	0.05
H ₂	- 0.58	3.89	- 448.96	0.16	- 0.20	0.57	- 156.95	0.02
H ₃	0.81	11.54	894.11	0.63	- 0.04	0.04	- 45.31	0.00
C ₀	- 0.99	4.22	- 409.83	0.13	1.13	6.57	467.71	0.17
C ₁	- 0.30	1.56	- 321.23	0.08	0.10	0.20	105.98	0.01
C ₂	0.60	4.31	476.26	0.18	- 0.36	1.84	- 284.52	0.06
C ₃	0.79	1.74	256.78	0.05	- 0.81	2.18	- 263.34	0.06
A ₁	1.31	15.16	839.84	0.56	0.90	8.64	579.46	0.27
A ₂	- 0.07	0.06	- 56.76	0.00	- 0.76	8.60	- 619.36	0.30
A ₃	- 0.35	0.96	- 207.37	0.03	- 0.58	3.26	- 350.39	0.10
A ₄	- 1.16	9.27	- 633.65	0.32	0.89	6.48	484.09	0.19
M	- 0.08	0.11	- 86.33	0.01	0.16	0.57	180.78	0.03
F	0.08	0.11	86.33	0.01	- 0.16	0.55	- 180.78	0.03
O ₁	- 0.04	0.03	- 46.48	0.00	- 0.84	14.63	- 909.85	0.65
O ₂	1.49	13.07	740.59	0.43	0.99	7.00	495.28	0.19
O ₃	- 0.63	5.08	- 524.93	0.22	0.68	7.09	567.23	0.25
I ₁	- 0.80	7.89	- 647.64	0.33	0.53	4.11	427.46	0.14
I ₂	0.20	0.45	150.20	0.02	- 0.14	0.26	- 104.17	0.01
I ₃	0.57	3.22	394.00	0.12	- 0.34	1.37	- 235.33	0.04
I ₄	0.70	1.38	229.53	0.04	- 0.56	1.06	- 183.80	0.03
L	0.71	5.17	504.64	0.20	1.15	16.58	826.25	0.54
NL	- 0.29	2.09	- 504.64	0.20	- 0.47	6.71	- 826.25	0.54
Total		100	72.51	4.02	1.75	100	221.91	3.68

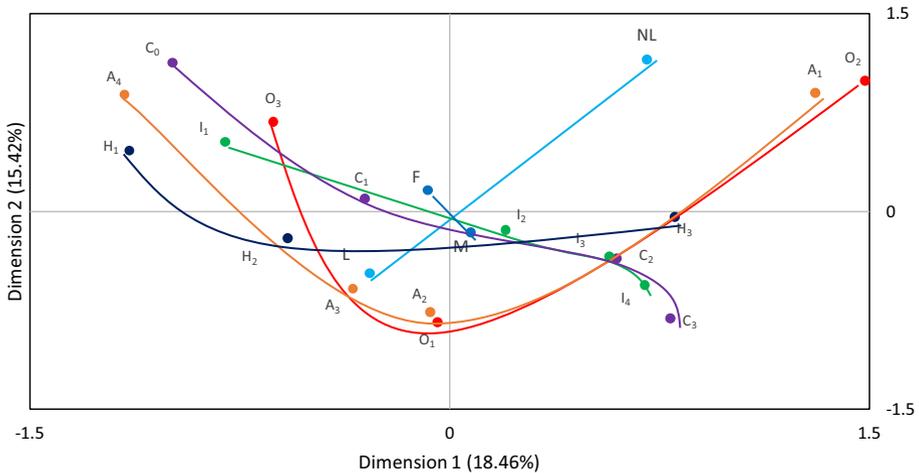


Fig. 2 Factor map of categories

test-value that follows a Gaussian distribution. Values below -1.96 or above 1.96 indicate that the coordinate is statistically significantly different from zero.

The cosine² value describes to which degree a dimension explains a category. For example, in dimension 1 the cosine² for category H1 is 0.30 (or 30%) and in dimension 2 it is 0.05 (or 5%). The sum of the cosine² values across all 15 dimensions is equal to 1 (or 100%) for each category. The detailed calculation procedure of these parameters can be found in Husson et al. (2017). Not all categories are displayed well in any particular dimension. The squared correlation of categories for all 15 dimensions shows that H₃ (household size category 3) is well represented in dimension 1 while I₃ (income group 3 i.e. from 3000 to 5000 euros per month) is best represented in dimension 6.

The relationship among attribute categories in the first two dimensions is shown in the factor map in Fig. 2. The closer the categories, the more related they are with each other. The categories are coloured, and the lines (drawn manually) show the distribution of each category along the map. The map shows that the socio-demographic attributes can be broadly grouped into three clusters. The top right part of the map shows that students (O₂) are more likely to fall between 0 and 22 years (A₁) and to not have a driver's license (NL). NL is a little further away from this group of three points as only people with the age 17 or older can have a driver's license in Germany. Overall, the upper right quadrant of the map mostly represents a cluster of students. The middle portion of the map shows that workers (O₁) are more likely to be between 23 and 64 years (A₂ and A₃), have high categories of 2000 euros and above (I₂, I₃ and I₄), with a higher likelihood to own a car and a driver's license (L). This portion represents the workers' cluster.

The top left portion of the graph mainly shows retired people (O₃) with lower income and an age (A₄) of 65 years and above. They could also be homemakers or currently unemployed people with lower income and more likely to not own a car. This portion of the map could be called the cluster others.

The relationship of household income and auto ownership is well depicted. It shows that people with low income (I₁) tend to either not own a car (C₀) or own one car (C₁). People with medium income (I₂) are more likely to own one or two cars, while people with high income (I₃ and I₄) tend to own two and more cars (C₂ and C₃). The location of household

sizes is also interesting. Students (O_2) are more likely to live with their parents (represented by household size of 3 or more people, H_3).

The gender categories are close to each other and near the origin, which shows that gender has little influence on other characteristics in dimensions 1 and 2. The interpretation of the above map seems intuitive, but it should be kept in mind that Fig. 2 only shows the relationship among attribute categories in two dimensions.

Activity sequence pattern

The distribution of activity sequence patterns in the first two dimensions is shown in Fig. 3. Only the top and bottom 10 activity sequence patterns are labelled for visual reasons. Similar parameters as shown in Table 4 were generated by multiple correspondence analysis for activity sequence patterns except ‘contribution’, as activity sequence patterns did not contribute towards the formation of dimensions.

The map above shows that the activity sequence patterns can be broadly grouped into three clusters which correspond to the clusters defined for socio-demographic attributes (Fig. 2). The one at the top right belongs to the activity sequence patterns that contain education activities (e.g. TC4, TC6 and TC190) and can be called education cluster, the cluster at the centre bottom belongs to activity sequence patterns that contain work activities (e.g. TC1, TC9, TC103, TC107, TC108, TC110 and TC 112) and can be called work cluster, and the cluster in the middle belongs to activity sequence patterns containing every other type of activity modeled (e.g. TC2, TC3, TC111) and can be called the other cluster. To better understand the relationship between clusters of socio-demographic attributes and activity sequence patterns, a combined map of attribute categories and activity sequence patterns is shown in Fig. 4. The activity sequence patterns are colour coded in three clusters according to the type of activity they contain. It is also interesting to note that activity sequence pattern TC1 (H–W–H), which is the most frequently occurring activity sequence pattern in the dataset, is very close to activity sequence pattern TC112 (H–W–H–S–H–PE–H), which is the least common activity sequence pattern in the dataset shown in bold and italic

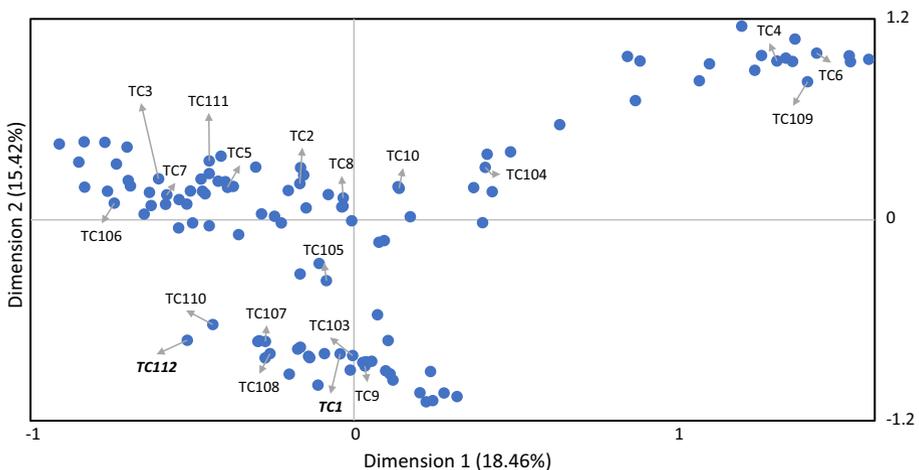


Fig. 3 Factor map of activity sequence patterns

sequence patterns. The likelihood (expressed by MCA distance) for individual i to select activity sequence pattern j is calculated by Eq. 1

$$d_{ij} = \sqrt{\sum_{k=1}^n [(I_i - TC_j)_k]^2 \times w_k} \tag{1}$$

where d_{ij} = distance between individual i and activity sequence pattern j ; I_i = coordinate value of individual in k th dimension; TC_j = coordinate value of activity sequence pattern j in k th dimension; w_k = weight of k th dimension (equal to percentage of variance explained by dimension k); k = dimension number; n = total number of dimensions

Not all the dimensions explain the same amount of variability, but some dimensions are more important than other ones. Therefore, a weighted factor (w_k) is included in Eq. 1 for a weighted calculation of distance. These weights are equal to the percentage of variance explained by each dimension formed by multiple correspondence analysis. The distance should be computed considering all the dimensions in order to capture the whole variability of data. By using Eq. 1, a matrix can be formed that provides the distances between individuals and activity sequence patterns.

Calculation of utilities

The utility of an activity sequence pattern for a given individual depends on the impedance or distance between the individual and the activity sequence pattern. The utility of an activity sequence pattern closer to the individual is larger than the activity sequence pattern that is farther from the individual. For this purpose, an exponential function is used. Given an individual i , the utility of each activity sequence pattern j can be calculated by Eq. 2 below:

$$U_{ij} = e^{-\alpha \times d_{ij}} \tag{2}$$

where U_{ij} = utility of individual i for activity sequence pattern j ; d_{ij} = distance between individual i to activity sequence pattern j ; α = calibrated parameter

By using Eq. 2, a utility matrix can be formed between all individuals and all observed activity sequence patterns.

Calculation of probabilities

Based on the calculated utilities, the probability of individual i choosing activity sequence pattern j can be calculated using Eq. 3 below:

$$P_{ij} = \frac{e^{\beta \times u_{ij}} \times \frac{f_j}{d_{ij}}}{\sum_{j=1}^n \left(e^{\beta \times u_{ij}} \times \frac{f_j}{d_{ij}} \right)} \tag{3}$$

where P_{ij} = probability of individual i choosing activity sequence pattern j ; u_{ij} = utility of individual i for activity sequence pattern j ; f_j = relative frequency of activity sequence pattern j as observed in the survey data; d_{ij} = distance between individual i to activity sequence pattern j ; n = number of activity sequence patterns; β = calibrated parameter

Equation 3 multiplies the exponentiated utility with a scaling factor consisting of the relative frequency of activity sequence pattern j divided by the distance between activity

sequence pattern j and individual i . This scaling factor ensures that very frequent activity sequence pattern (such as home–work–home) are chosen more frequently than less common ones, even if this less common activity sequence pattern is relatively close to a given individual. Two activity sequence patterns can be at the same distance from the individual based on its socio-demographic attributes, but one activity sequence pattern could be more frequent than the other one. The distance calculated by the MCA method merely determines that a given activity sequence pattern is closer to one person type than to another, but this distance does not indicate the frequency of an activity sequence pattern. For example, TC1 (H–W–H) and TC112 (H–W–H–S–H–PE–H) are both activity sequence patterns that are selected mostly by workers. Therefore, the distance calculated by MCA of these two activity sequence patterns is close to workers. However, TC1 is the most frequently selected activity sequence pattern, while TC112 is a less frequently selected activity sequence pattern. If the selection was made based on distance only, TC1 and TC112 would be selected with similar frequency. While the weights help to select the right frequency of activity sequence patterns, the MCA method helps to assign the most likely activity sequence patterns to a right person based on their socio-demographic attributes.

Predictive model results

For model calibration, unique person types were determined. A total of 899 unique types were found. All individuals were assigned to the three categories of workers, students and others. Initially, the model was calibrated for these three groups. The coordinates of the individuals that were used to build the multiple correspondence analysis model were already estimated by the same analysis. Given the coordinates of activity sequence patterns from the multiple correspondence analysis model, the distance between individuals and activity sequence patterns was calculated using Eq. 1. The parameter α of the utility equation (Eq. 2) and the parameter β of the probability equation (Eq. 3) were calibrated for each group. The α parameter defines to which degree individuals mostly consider activity sequence patterns that are very close in terms of MCA distance, or whether individuals also consider activity sequence patterns that are a little bit further away. The α value should be calibrated to different values for different individuals, because different user groups are likely to have a different “tolerance” for less common activity sequence patterns. The calibration suggests that students stick more closely to the most likely activity sequence patterns (expressed by the larger α value of 0.6) than workers ($\alpha = 0.45$). Most students go to school and back, while workers tend to have more variety in activity sequence patterns, and therefore, will choose activity sequence patterns with larger MCA distances as well. The beta parameter is used to translate utilities into probabilities. The beta parameter defines how much individuals focus on the highest utilities, or whether people are willing to (occasionally) select options with lower utilities.

Coefficient of determination (R^2) and root-mean-square-error (RMSE) were used as a measure for calibrating the parameters, maximizing R^2 and minimizing RMSE. These goodness-of-fit measures describe how closely the model predicts activity sequence patterns aggregated to educational, work and leisure clusters. The results are shown in Table 5 and Fig. 5. For this comparison, all activity sequence patterns that contain the activity work were grouped to the activity cluster work. All activity sequence patterns that contain the activity education were grouped to the education cluster. The remaining activity sequence patterns were aggregated to other. The observed probability of 2.04% of the education

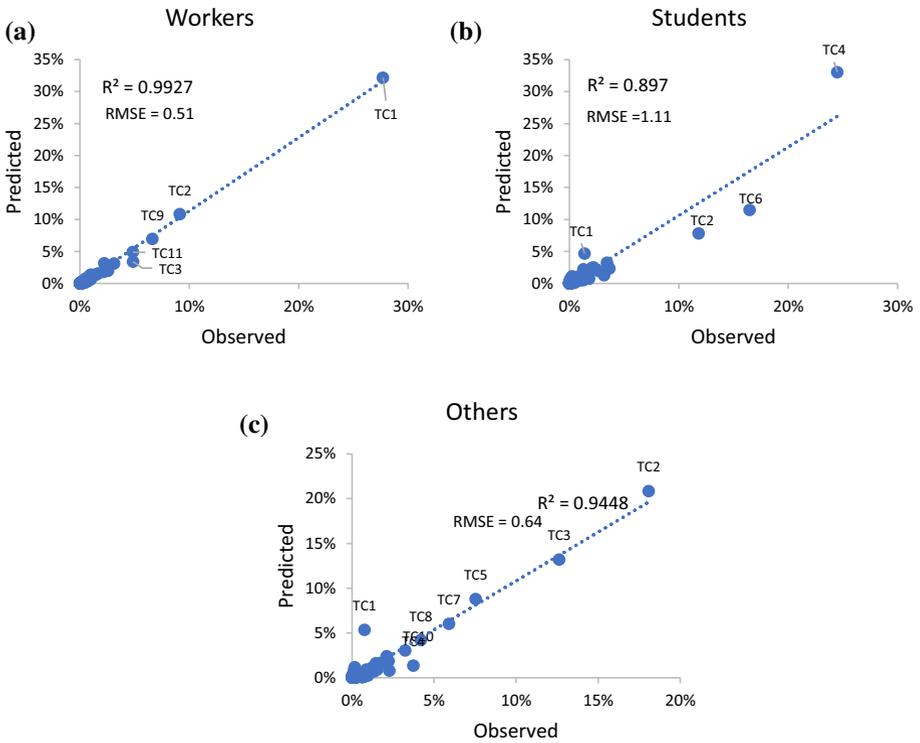


Fig. 5 a–c Model calibration results for the three categories workers, students and others

Table 5 Model Calibration Results for three categories

Person type		Share of population (%)	Cluster1 (Education) (%)	Cluster2 (Work) (%)	Cluster3 (Others) (%)	R ²	RMSE	α	β
Workers	Observed	48.2	2.04	59.34	38.62	0.99	0.51	0.45	14
	Predicted		1.37	59.84	38.78				
Students	Observed	16.4	58.30	2.66	39.04	0.90	1.11	0.6	20
	Predicted		58.14	9.79	32.07				
Others	Observed	35.4	8.86	1.93	89.21	0.94	0.64	0.54	12.3
	Predicted		3.18	11.20	85.62				

cluster for a worker is calculated by adding all the observed probabilities of all activity sequence patterns present in that cluster. The predicted probabilities are calculated by the addition of predicted probabilities within each cluster. The results show that for workers, 59.34% of the activity sequence patterns were predicted for cluster 2 (work-related activity sequence patterns), while the model is predicting 59.84% from the same cluster. Similarly, education and other activity sequence patterns were predicted reasonably well for workers.

For students, the model is predicting correctly education activity sequence patterns but overpredicts activity sequence patterns in the work cluster.

A reason for this reduced fit could be the larger number of work activity sequence patterns (31 activity sequence patterns) compared to education activity sequence patterns (15 activity sequence patterns). The others category predicts reasonably well the activity sequence patterns from cluster 3 (others cluster), while underpredicting the education and overpredicting the work activity sequence patterns.

Secondly, the model was calibrated for the top 10 most frequently occurring person types to compare the aggregate calibration with a calibration for individual person types. (Table 6).

For instance, person type 1 represents a group of individuals who are workers, with household size 1, auto-ownership category 1, age group category 1, male, income category 1 and own a driver’s license. This person type represents 2.0% of the population. The parameters α and β were calibrated individually for these ten person types as shown in Table 7 and Fig. 6. The figure also points out the outliers for every person type. The parameters calibrated for the three broad groups (Table 5) applied to person types shows that the person types are rather sensitive to calibration factors.

R^2 drops significantly and RMSE increases significantly for person type 5, 6, 7 and 10 when individually calibrated parameters are used. For instance, R^2 drops from 0.92 to 0.79 and RMSE increase from 0.79 to 2.26 for person type 5. The cluster probabilities also do not match as well for these person types when parameters aggregated to three groups are used. Therefore, it seems advisable to calibrate each person type separately to obtain best results. However, infrequently occurring person types will not have enough records in the survey for calibration. It is therefore suggested to use individually calibrated parameters for all person types with at least 30 records and aggregated parameters for the three person groups for all other person types.

The described methodology can simulate the selection between those 112 activity sequence patterns for which at least 30 observations were found in the survey. With this method, 72% of all persons and their activity sequence patterns are covered. The other less frequently observed activity sequence patterns, which often are the more complex patterns, also need to be included in application to represent the full pattern of activity sequences. As those remaining 28% cannot be related with statistical confidence to socio-demographic attributes, they are assigned randomly. They are assigned randomly based on the observed

Table 6 Person type definitions

Person type	Person definition	Category	Population (%)
1	$H_1 C_1 A_2 M O_1 I_1 L$	Worker	2.0
2	$H_3 C_2 A_2 M O_1 I_3 L$	Worker	1.7
3	$H_3 C_2 A_2 M O_1 I_2 L$	Worker	1.6
4	$H_3 C_1 A_2 M O_1 I_2 L$	Worker	1.3
5	$H_3 C_2 A_2 F O_1 I_2 L$	Worker	1.3
6	$H_3 C_2 A_1 M O_2 I_3 NL$	Student	1.3
7	$H_2 C_1 A_4 M O_3 I_1 L$	Other	2.3
8	$H_1 C_0 A_4 F O_3 I_1 NL$	Other	2.0
9	$H_2 C_1 A_4 M O_3 I_2 L$	Other	1.5
10	$H_1 C_1 A_4 F O_3 I_1 L$	Other	1.3

Table 7 Model Calibration results for top 10-person types based on individual calibration factors

Person Type		%age	Cluster1 (Education) (%)	Cluster2 (Work) (%)	Cluster3 (Others) (%)	R ²	RMSE	α	β
<i>Workers</i>									
1	Observed	2.0	0.66	61.93	37.42	0.81	1.44	0.3	15
	Predicted		0.19	56.79	43.02				
2	Observed	1.7	0.24	64.95	34.81	0.97	0.55	0.5	5
	Predicted		3.90	62.18	33.92				
3	Observed	1.6	1.09	64.96	33.95	0.99	0.37	0.5	5.5
	Predicted		3.23	64.74	32.03				
4	Observed	1.3	0.35	66.95	32.70	0.99	0.48	0.5	6.5
	Predicted		2.29	65.56	32.15				
5	Observed	1.3	0.36	50.55	49.09	0.92	0.79	0.5	3.2
	Predicted		4.90	51.32	43.78				
<i>Students</i>									
6	Observed	1.3	58.65	0.22	41.13	0.90	1.05	0.6	10
	Predicted		50.20	10.89	38.90				
<i>Others</i>									
7	Observed	2.3	0.00	0.88	99.12	0.70	1.49	0.5	9
	Predicted		1.22	4.97	93.81				
8	Observed	2.0	0.00	0.15	99.85	0.80	1.52	0.3	16
	Predicted		1.11	3.48	95.41				
9	Observed	1.5	0.26	0.66	99.08	0.81	1.14	0.8	22.5
	Predicted		1.60	5.93	92.47				
10	Observed	1.3	0.00	2.07	97.93	0.72	1.35	0.2	5
	Predicted		1.37	7.04	91.59				

frequency, under the constraint that education-related patterns are assigned to students only and work-related patterns are assigned to workers only.

Conclusions

The proposed methodology works reasonably well to explain frequently observed activity sequences. In the given dataset, 72% of all observed activity sequence patterns were conducted by at least 30 respondents, and therefore, were deemed to be represented sufficiently well by this survey. The remaining 28% percent of all observed activity sequence patterns are more complex and rather rare, and therefore, cannot be represented by this method as well-founded. While the assignment of rare activity sequence patterns has a random component, frequently observed activity sequence patterns are assigned realistically based on socio-demographic data.

One of the most challenging problems for activity-based models is the determination of activity sequences. Commonly, a two-step choice is implemented, where first a main activity is selected, and subsequently, stops on this tour may be added. The hypothesis of this common approach for activity-based models is that every activity sequence is driven by

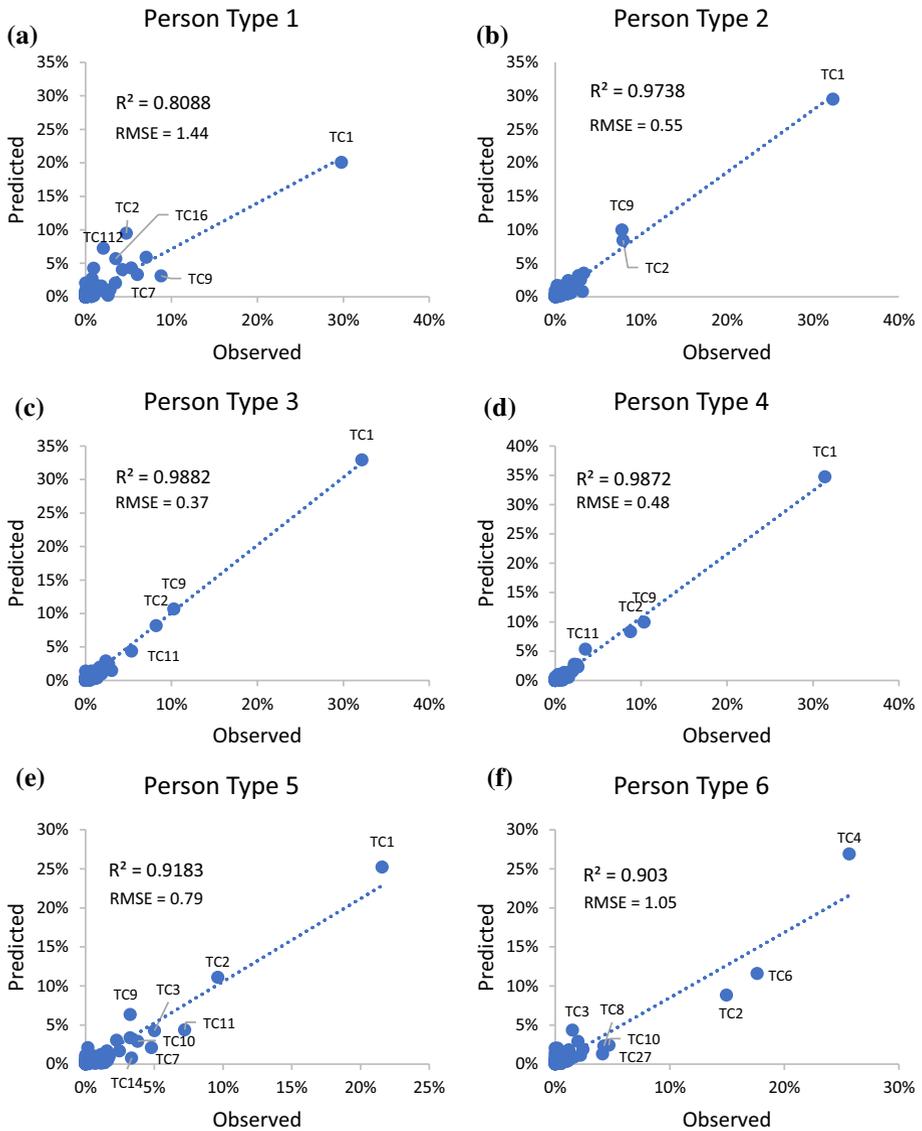


Fig. 6 a–j Model calibration results for top 10-person types

one dominant activity. Furthermore, activity sequences are commonly modeled from home to home, i.e., the sequence H–W–H–S–H would be modeled as two independent sequences (H–W–H and H–S–H). With the exception of time constraints, most activity-based models do not consider the attributes of the first tour when modeling the second tour. Furthermore, most activity-based models limit the number of intermediate stops to one or two stops, thereby limiting the diversity of modeled activity sequences. In practice, activity-based models solve this problem by a very large number of discrete choice problem sets that are ordered based on heuristic assumptions (such as, a main tour purpose is chosen first).

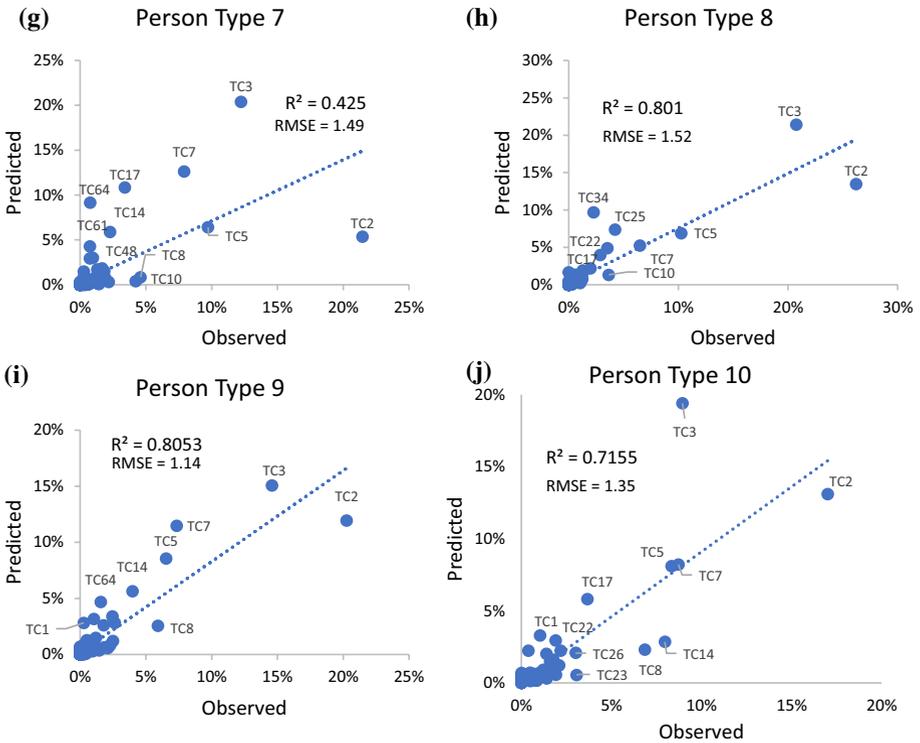


Fig. 6 (continued)

The study described here developed a methodology that considers activity sequence patterns of individuals based on socio-demographic attributes. No individual discrete choice models needed to be estimated, but a probability for 112 different activity sequence patterns was calculated based on the agents' socio-demographic attributes. Among the most favourable alternatives for an individual, preference is given to those alternatives that are more frequently selected according to the household travel survey. The whole day is simulated as one consistent sequence, and the number of intermediate stops is only limited by the sample size of complex activity sequence patterns found in the survey data. Moreover, the methodology is flexible and could include land-use and transportation related attributes to further explain activity sequence patterns. With this new method, a very complex module of every activity-based model was simplified into a single step that creates a large variety of activity sequence patterns based on socio-demographic data.

For a full travel demand model, the presented methodology needs to be expanded by a location choice model for each activity, a mode choice model and an activity timing model. The location choice model needs to work similar to a traditional destination choice model, with the exception that the locations of each activity on a tour tends to be in spatial proximity (i.e., for a H-W-S-H activity sequence pattern, someone would most likely select a shopping location near work or near home). The mode choice model works similarly to a traditional mode choice model of a four-step model, with the exception that commonly a main mode is chosen for an entire tour, and sub modes are chosen for individual trips.

The activity timing model needs to model a preferred arrival time and an activity duration. While there is still some model development needed, the method presented here is considered to be a major step towards significantly simpler activity-based models.

Acknowledgement Open Access funding provided by Projekt DEAL. This research was completed with the support of the Technische Universität München, Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no. 291763.

Author contributions UA: Study conception, data curation and analysis, interpretation of results and draft manuscript preparation. ATM: Study conception, supervision and interpretation of results. RM: Study conception, supervision and interpretation of results. All authors reviewed the results and approved the final version of the manuscript.

Compliance with ethical standards

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adler, T., Ben-Akiva, M.: A theoretical and empirical model of trip chaining behavior. *Transp. Res. Part B Methodol.* **13**, 243–257 (1979)
- Ahmed, U.: Development of a microscopic activity sequence modeler, Master's Thesis at the Technical University of Munich https://www.bgu.tum.de/fileadmin/w00blj/msm/masterThesis/Ahmed_2018_rev.pdf, (2018)
- Axhausen, K.W., Gärling, T.: Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transp. Rev.* **12**, 323–341 (1992)
- Badoe, D.A., Chen, C.-C.: Unit of analysis in conventional trip generation modelling: an investigation. *Can. J. Civ. Eng.* **31**, 272–280 (2004). <https://doi.org/10.1139/l03-098>
- Bhat, C.: Modeling the commute activity-travel pattern of workers: formulation and empirical analysis. *Transp. Sci.* **35**, 61–79 (2001)
- Bowman, J.L., Ben-Akiva, M.E.: Activity-based disaggregate travel demand model system with activity schedules. *Transp. Res. Part A Policy Pract.* **35**, 1–28 (2001)
- Charypar, D., Nagel, K.: Generating complete all-day activity plans with genetic algorithms. *Transportation (Amst)*. **32**, 369–397 (2005)
- Cyganski, R., Justen, A., Schulz, A., Köhler, K.: Generation of coherent trip chains for travel behaviour analysis and modelling. In: European Transport Conference (2013)
- Diana, M., Pronello, C.: Traveler segmentation strategy with nominal variables through correspondence analysis. *Transp. Policy*. **17**, 183–190 (2010)
- Downes, J.D., Johnsen, S., Morrell, D.: Household and Person Trip Generation Models. Transport and Road Research Laboratory (TRRL) Report 401. Wokingham (Berkshire), United Kingdom (1978)
- Drchal, J., Čertický, M., Jakob, M.: Data-driven activity scheduler for agent-based mobility models. *Transp. Res. Part C Emerg. Technol.* **98**, 370–390 (2019)
- Gangrade, S., Pendyala, R.M., McCullough, R.G.: A nested logit model of commuters' activity schedules. *J. Transp. Stat.* **5**, 19–36 (2002)

- Golob, T.F.: A nonlinear canonical correlation analysis of weekly trip chaining behaviour. *Transp. Res. Part A Gen.* **20**, 385–399 (1986)
- Golob, T.F.: A simultaneous model of household activity participation and trip chain generation. *Transp. Res. Part B Methodol.* **34**, 355–376 (2000). [https://doi.org/10.1016/S0191-2615\(99\)00028-4](https://doi.org/10.1016/S0191-2615(99)00028-4)
- Habib, K.N., El-Assi, W., Hasnine, M.S., Lamers, J.: Daily activity-travel scheduling behaviour of non-workers in the National Capital Region (NCR) of Canada. *Transp. Res. Part A Policy Pract.* **97**, 1–16 (2017)
- Hafezi, M.H., Liu, L., Millward, H.: Learning daily activity sequences of population groups using random forest theory. *Transp. Res. Rec.* **2672**, 194–207 (2018)
- Hafezi, M.H., Liu, L., Millward, H.: A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation (Amst)*. **46**, 1369–1394 (2019)
- Han, G., Sohn, K.: Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transp. Res. Part B Methodol.* **83**, 121–135 (2016)
- Hanson, S., Huff, J.: Classification issues in the analysis of complex travel behavior. *Transportation (Amst)*. **13**, 271–293 (1986)
- Husson, F., Lê, S., Pagès, J.: *Exploratory Multivariate Analysis by Example Using R*. Chapman and Hall/CRC, Boca Raton (2017)
- Jiang, S., Ferreira, J., González, M.C.: Clustering daily patterns of human activities in the city. *Data Min. Knowl. Discov.* **25**, 478–510 (2012)
- Kim, K.: Discrepancy analysis of activity sequences: What explains the complexity of people's daily activity-travel patterns? *Transp. Res. Rec. J. Transp. Res. Board.* **2413**, 24–33 (2014)
- Kitamura, R., Chen, C., Pendyala, R.M., Narayanan, R.: Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation (Amst)*. **27**, 25–51 (2000)
- Li, S., Lee, D.-H.: Learning daily activity patterns with probabilistic grammars. *Transportation (Amst)*. **44**, 49–68 (2017)
- McNally, M.G.: The four step model. *Handb. Transp. Model.* **1**, 35–52 (2007)
- Moeckel, R., Huntsinger, L., Donnelly, R.: From macro to microscopic trip generation: representing heterogeneous travel behavior. *Open Transp. J.* **1**, 11 (2017)
- Ordóñez Medina, S.A.: Recognizing personalized flexible activity patterns. In: 14th International Conference on Travel Behavior Research (IATBR 2015) (2015)
- Paleti, R., Vovsha, P., Vyas, G., Anderson, R., Giaimo, G.: Activity sequencing, location, and formation of individual non-mandatory tours: application to the activity-based models for Columbus, Cincinnati, and Cleveland. *Transportation (Amst)* **44**, 615–640 (2017)
- Pas, E.I.: The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environ. Plan. A.* **16**, 571–581 (1984)
- Pitombo, C., Sousa, A., Filipe, L.: Classification and regression tree, principal components analysis and multiple linear regression to summarize data and understand travel behavior. *Transp. Lett.* **1**, 295–308 (2009)
- Pokhrel, A.: Comparative Analysis of Person-Category and Household-Category Trip Generation Models, Master's Thesis at the Technical University of Munich (2017). https://www.bgu.tum.de/fileadmin/w00blj/msm/theses/pokhrel_2017.pdf
- Recker, W.W., McNally, M.G., Root, G.S.: A model of complex travel behavior: part I—theoretical development. *Transp. Res. Part A Gen.* **20**, 307–318 (1986a)
- Recker, W.W., McNally, M.G., Root, G.S.: A model of complex travel behavior: part II—an operational model. *Transp. Res. Part A Gen.* **20**, 319–330 (1986b)
- Sammour, G., Bellemans, T., Vanhoof, K., Janssens, D., Kochan, B., Wets, G.: The usefulness of the Sequence Alignment Methods in validating rule-based activity-based forecasting models. *Transportation (Amst)*. **39**, 773–789 (2012)
- Supernak, J., Talvitie, A., DeJohn, A.: Person-category trip-generation model. *Transp. Res. Rec.* **944**, 74–83 (1983)
- Xu, Z., Kang, J.E., Chen, R.: A random utility based estimation framework for the household activity pattern problem. *Transp. Res. part A policy Pract.* **114**, 321–337 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Usman Ahmed is a PhD student in the Department of Civil and Mineral Engineering at the University of Toronto. He received a M.Sc. degree from the Department of Civil, Geo and Environmental Engineering

at the Technical University of Munich. His research is focused on both passenger and freight demand modeling.

Ana Moreno is a senior research associate at the Research Group on Modeling Spatial Mobility at the Technical University of Munich in Germany. She completed her doctorate in Transportation Engineering in 2015 at the Universitat Politècnica de València in Spain. Her research interests include travel demand models; integrated land use/transport models; traffic operations and road safety on two-lane highways and traffic calming in urban areas.

Rolf Moeckel is an Assistant Professor in the Department of Civil, Geo and Environmental Engineering at the Technical University of Munich in Germany. He leads a research group called Modeling Spatial Mobility that focuses on travel behavior research and transport modeling. He conducted research as a PostDoc at the National Center for Smart Growth at the University of Maryland and worked as a consultant with Parsons Brinckerhoff in New York and Albuquerque in the US. He holds a doctorate in spatial planning from the University of Dortmund in Germany.