Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 34 – 41

Procedia Computer Science

The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)

A generic data-driven sequential clustering algorithm determining activity skeletons

Wim Ectors[a,*], Bruno Kochan[a], Luk Knapen[a], Davy Janssens[a], Tom Bellemans[a]

[a]*Transportation Research Institute, Hasselt University, Wetenschapspark 5 Box 6, B-3590 Diepenbeek, Belgium*

Abstract

Many activity-based models start by scheduling inflexible or mandatory activities (if present), before more flexible activities. Often work and educational activities are assumed as most stringent and recognized as the only mandatory activities. According to this definition, only 45% of all schedules contains a mandatory activity (OVG single-day travel survey in Flanders, Belgium). This means 55% of schedules does not have a traditional mandatory-flexible activity structure. This research proposes a completely data-driven approach to reveal the real basic structure of individuals' schedules, i.e. the skeleton schedule sequence. To this end, a sequential clustering algorithm was developed. Furthermore, an in-depth analysis of the parameter settings was performed. The proposed method reveals a set of skeleton activity schedules and confirms the importance of work and education.

Peer-review under responsibility of the Conference Program Chairs

Keywords: skeleton schedules; activity patterns; mandatory activities; activity-based modeling

1. Introduction

Despite having some limitations, the traditional four-step model is still abundantly used by small- and medium-sized cities and consultancy agencies across the world. It is relatively easy to implement and has a low complexity. Activity-based models (ABMs) have been under development for many years and remedy many of the traditional four-step model's limitations. They are characterized by the fact that the base unit is an activity rather than a trip or tour.

In 1986 the STARCHILD system of models was introduced. It used a utility-maximizing approach to determine the expected activity pattern (syn. schedule). It distinguished between planned and unplanned activities. Planned activities were predicted and inserted into the schedule first and subsequently used as boundary conditions in the unplanned activity planning problem[1,2]. Also in most recent ABMs mandatory, inflexible activities are determined around which more flexible activities can be scheduled[3,4]. Often only work and school (university) activities are considered for a mandatory activity. This approach and assumption is used in i.a. SWDAS (TASHA), ALBATROSS, CEMDAP, FAMOS, the CT-RAMP family of ABMs in general and recently also in the daily-activity pattern model. However, e.g. the HAPP framework introduced in 1995 does *not* differentiate between flexible or inflexible activities.

* Corresponding author. Tel.: +32-11-269114.
E-mail address: wim.ectors@uhasselt.be

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.096

However, only 33% of all activity schedules in Flanders, Belgium, contain a work or business activity. Including educational activities, still only 45% of all schedules contains a mandatory activity. Admittedly, one does not necessarily have a mandatory activity scheduled every day. However considering modeling purposes, there could still be meaningful activity patterns in those remaining 55% of schedules which do not obey the popular mandatory- vs flexible activity-structure. Extracting these is the purpose of the current research: revealing *skeleton* activity patterns which can be used as nuclei for subsequent scheduling decisions, in a generic data-driven fashion.

Some previous researchers found indications that the concept of skeleton patterns is plausible. Doherty[5] revealed that approximately 34% of activity time is considered routine or planned weeks to years in advance. This substantial fraction of preplanned and routinized (recurring) activity time can be viewed as a base for a skeleton pattern in one's schedule. Arentze et al.[6] postulate that activity patterns are the outcome of a learning process, of which the result is a set of rules. They reject the notion that individuals systematically compare all possible activity patterns. A skeleton activity pattern implicitly captures some of these rules and would therefore limit the activity pattern choice set. Additionally, Arentze et al.[7] found that predicting tour skeletons instead of assuming them given within the model did not significantly reduce the goodness of fit of their ABM: ALBATROSS 2. They however did not detail the process by which these skeletons are generated.

A k-means-based clustering technique was developed by Allahviranloo et al.[8], ultimately in order to accurately forecast activity patterns of an individual. Individuals with their socio-demographic information are clustered in two stages according to their observed activity patterns. A surprising result was the discovery of a core set of sequences often serving as a centroid for other patterns. This discovery strengthens the belief that a core set of skeleton schedules exists and that these could be used for activity pattern modeling purposes.

Roorda and Ruiz[9] found evidence against the common assumption that a skeleton schedule only consists of mandatory activities (such as work or education). They used a CHASE-based dataset from the TAPS longitudinal survey.

The current research forms the first step towards a multi-agent, rule-based traffic demand model. Similar to ALBATROSS[6] and FEATHERS[10] it might be based on a system of sequential decisions using decision trees. Note that within this research, skeleton activity patterns are defined as the fundamental activity patterns that are the backbone for many complete sequences of activities. Some authors used *skeleton activity patterns* in a slightly different meaning, i.e. that of a set of strictly inflexible or mandatory activities. In the current research activity skeletons are identified based on frequency rather than flexibility. This criterion reveals important patterns that may be used as fundamental patterns for modeling purposes.

2. Methodology

2.1. Data Description

The data used in this research resulted from the *Transportation Behavior Research Flanders* (*Onderzoek Verplaatsingsgedrag* or OVG) travel survey conducted in the Flanders region, Belgium. This large-scale survey is funded by the Ministry of Mobility and Infrastructure. Single-day travel diaries (including weekends) enriched with individual and household socio-demographical information were collected for approximately 17,300 individuals. Participants were asked to keep a detailed log of their trip purpose, departure and arrival time, and trip origin and destination on a randomly assigned day. Individual weights were calculated in order that the OVG sample may correctly represent the true population of Flanders. These weights were used in all reported figures of this paper. The survey was conducted in multiple phases from 2007 until 2013. A new phase started in 2015.

Of the 17,300 participants, approximately 13,200 conducted at least one trip. Using the reported trip purposes of Table 1, single-day activity schedules were created. All of them are assumed to start at home, but do not necessarily end at home. A limitation is that only activities for which a trip was performed are included (i.e. no at-home activities). Multiple identical consecutive activities were merged into a single occurrence, a practice which can be defended when purely studying activity patterns. It is beneficial as it removes some of the variance present in the data.

In total, approx. 2,600 unique single-day schedules could be constructed. The 14 schedules having the highest frequency account for 45% of the observations. None of the other schedules has a share larger than one percent. This shows that quite a large fraction of the observed travel patterns can be grasped in only 14 distinct schedules, but that

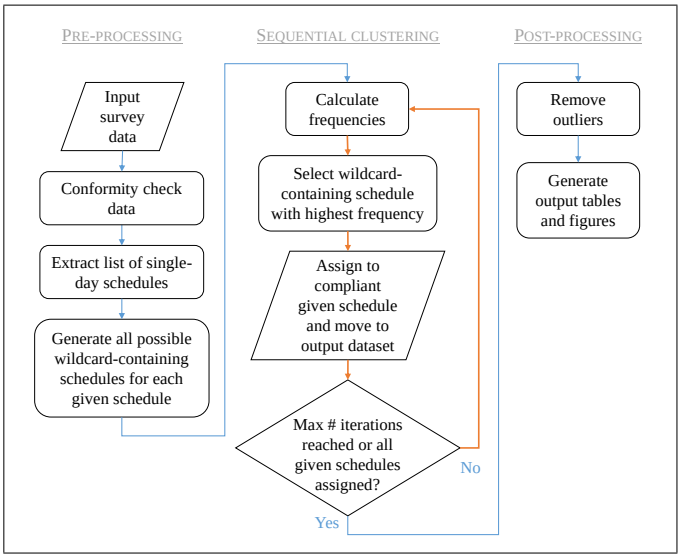

Fig. 1. Simplified flowchart of the proposed sequential clustering algorithm.

still 55% of the observed activity patterns occur more rarely and represent more complex behavior. It is especially from this fraction that the skeleton scheduling algorithm will attempt to extract common, single-day activity patterns.

2.2. Algorithm Description

This paper presents a sequential clustering algorithm that reveals common activity patterns in otherwise highly heterogeneous activity schedules. The data analysis and clustering algorithm were developed in SAS utilizing macros.

Observing the single-day activity patterns, often there are few differences between schedules. Replacing one (or a multiple) activities with a 'wildcard'-activity allows to describe multiple schedules with a single, wildcard-containing schedule. For example, an activity schedule 'home-shopping-home-shopping-home' and 'home-shopping-home-recreation-home' can both be described by 'home-shopping-home-X-home', where the 'X' represents the wildcard. Optimizing the location of such a wildcard can be coded in an algorithm. The resulting skeleton schedules may be used as a starting point in ABMs. Figure 1 describes the algorithm in a simplified manner.

2.2.1. Pre-Processing

Initially the input data is checked for irregularities such as missing data and erroneous entries. Subsequently, single-day activity schedules are constructed. For each schedule, all possible wildcard-containing schedules are calculated according to a set of rules. For instance, in **setting 1** one has to specify the minimum number of activities within a schedule not replaced by a wildcard. This is a setting of major consequence for the algorithm's numerical complexity since the total number of possible wildcard-containing schedules per given schedule is determined by:

$$N = \sum_{r=s}^{n} \frac{n!}{r!(n-r)!} \tag{1}$$

where N is the number of possible wildcard-containing schedules, n the number of activities in the given schedule (i.e. the schedule length), s the minimum number of remaining activities according to setting 1 and r is the number of remaining activities in the summation. The subtraction $n - r$ represents the number of wildcards in a schedule. For instance, each schedule with five activities (22% of the schedules) has 26 distinct wildcard-containing schedules when $s = 2$. However, with the same settings, each schedule with 10 activities will produce 1013 distinct alternatives.

One observes that this could potentially become problematic in case of very large datasets and many long schedules. Other settings partially address this issue, as well as other measures taken in the code to avoid an unacceptable numeric complexity. **Setting 2** and **setting 3** state, respectively, that a home-activity and a work-activity cannot be replaced

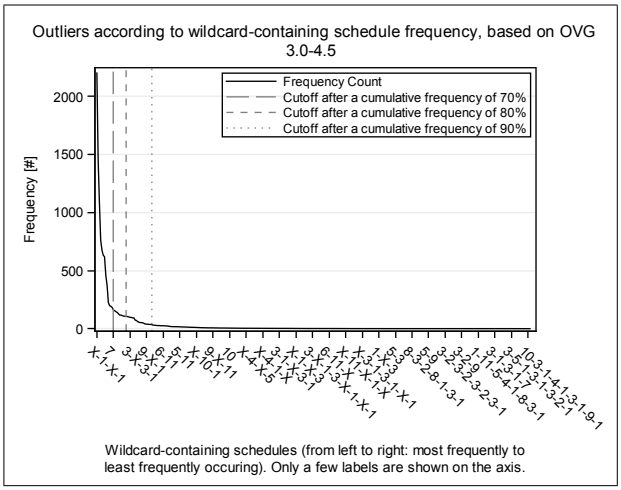

Fig. 2. Illustration of the skeleton schedule outlier identification process (*odd patterns*). The initial "home"-activity of all schedules was dropped. An "X" represents a wildcard.

Table 1. Travel motive encoding in OVG 3.0 - 4.5

Symbol	Travel Motive
1	Going home
2	Business trip
3	Work
4	Shopping
5	Visit someone
6	Education
7	Walk, tour, run
8	Bring/get someone/something
9	Recreation, sport, culture
10	Services (physician, bank...)
11	Other
99	No response

by a wildcard. By enabling this setting, one in fact enforces that these type of activities are part of the final skeleton pattern. This is conform the current understanding of how individuals use these activities as anchor points (or often called "pegs") whilst planning activities with a lower priority or higher degree of flexibility. **Setting 4** allows that multiple consecutive wildcards may be merged. If enabled, a single wildcard will replace multiple activities. This significantly reduces the complexity of the schedules, offering the simplification that is desired for skeleton activity patterns. **Setting 5**, allows to put a maximum complexity (i.e. a maximum number of activities) on the input schedules. Schedules with a very large number of activities occur very rarely and can be considered outliers (i.e. *odd patterns*). Removing these from the start will reduce the number of generated wildcard-containing schedules significantly since these extreme activity patterns produce the largest number of alternative schedules.

2.2.2. Sequential Clustering

Having created a dataset containing single-day activity schedules together with their compatible wildcard-containing schedules, a sequential clustering stage will determine the largest groups of unique wildcard-containing patterns. First the frequencies of occurrence of given and wildcard-containing schedules are calculated. Secondly, the wildcard-containing schedule with the highest frequency is selected. If it has a higher frequency than the compliant given schedule with highest frequency, then this wildcard-containing schedule is taken as the skeleton schedule and assigned to all given schedules that comply with this skeleton pattern. If frequencies of given and wildcard-containing schedules are equal, the given schedule is retained and subsequently assigned to itself as the skeleton schedule. The given schedules that were assigned a skeleton schedule are then moved to a separate dataset. Consequently they are excluded from the sequential clustering loop, as are all the other wildcard-containing schedules that were generated for that given schedule. The loop ends if all given schedules have been moved to the result dataset, or if a predefined number of skeletons has been found.

2.2.3. Post-Processing

In the final stage, outliers are removed. Within this research, outliers are *odd patterns* that occur at a low frequency. Outlier-detection is quite easily performed for numeric variables, but for nominal variables (as is the case here) rather few approaches are common. In this single-dimensional problem, it was opted to simply define outliers based on their frequency of occurrence. Only the fraction up to a certain cumulative "cutoff" frequency is retained, all others are disregarded as *odd patterns*. Figure 2 shows this for one particular run of the algorithm. The cutoff percentage is prone to optimization. The post-processing stage also produces a number of output tables and charts.

2.3. Sensitivity Analysis of Algorithm Settings

One could imagine that different settings of settings 1-5 might produce diverse results. In an attempt to better understand the influence of these parameters a sensitivity analysis was performed. Considering the ambition to use these skeleton schedules in ABMs, the ultimate goal is to most accurately 'predict' the skeleton schedule based on socio-demographic data. For this reason a decision tree classifier, which is also the basic discrete choice model in rule-based ABMs such as ALBATROSS [6] and FEATHERS [10], will classify the generated skeleton schedules based on socio-demographic data. Various algorithm settings will produce different sets of skeleton schedules, which directly affect classification accuracy of the decision trees. The sensitivity analysis investigates the influence of these various parameter settings on the prediction accuracy. The analysis was executed in two stages:

1. Use different sets of settings to generate sets of skeleton schedules. The sequential clustering algorithm in SAS is used to this end.
2. Use an in-house Python implementation of the ID3 decision tree classification algorithm to estimate train- and test set contingency matrix accuracy (CMA) figures. A CMA estimation is an indication of the classification accuracy and should be as high as possible for modeling purposes.

Since no founded assumptions on the relation between settings and CMA can be assumed, it was opted to explore the multidimensional problem by generating sets of skeleton schedules for a wide range of settings and then calculating the CMA for each of them. This is sub-optimal, but acceptable within this research' context.

Firstly, different sets of skeleton schedules were generated using the sequential clustering algorithm. A number of settings were available: 1) the minimum number of remaining activities in a schedule, 2) whether the home-activity can be replaced by a wildcard, 3) whether identical sequential wildcards are merged, 4) the cutoff threshold for outlier exclusion (cumulative frequency), 5) the maximum number of activities allowed per schedule after pre-processing and 6) if a single-day schedule should be split in multiple home-based tours. In total, 2,520 sets of skeleton schedules were produced according to various combinations of these settings.

Secondly, all of the skeleton schedules from the first step were classified with an ID3 decision tree algorithm. The generated skeleton schedules were the classification attribute, whilst socio-economic attributes connected to each individual of the given schedule were used to construct the decision tree. The dataset was split in a train (75%) and test set (25%). The CMA was calculated for each of the sets. In this stage, another parameter became available, i.e. the minimum number of instances per leaf of the decision tree. In general, this parameter needs to be chosen wisely to prevent overfitting. In total, $\pm 44,000$ ID3 trees were fitted.

The results were analyzed in a multiple linear regression model with transformed variables (see list below). A satisfying adj. R-square of 0.82 was found, illustrating a good understanding of the algorithm settings. Some findings:

- The most influential setting is that of the minimum number of non-wildcard activities in a wildcard-containing schedule (setting 1). A low value results in a higher average CMA (inversely correlated). Fewer remaining activities result in a reduced complexity of the resulting skeleton patterns. This reduced complexity results in higher-frequency skeleton patterns and therefore fewer alternatives for the decision tree classification attribute, which increases train and test set CMA.
- The second most important variable explaining test set CMA is the cutoff percentage for outlier (*odd pattern*) detection. The CMA is strongly correlated to the inverse of the cutoff percentage. This setting directly controls the number of alternatives of the classification attribute and hence has a large impact on CMA.
- The maximum number of activities allowed per schedule (setting 5, i.e. a filter in the pre-processing stage) has a slightly negative, but marginal effect. This setting begins to negatively influence CMA if its value is set to fewer than nine activities. Schedules with more activities apparently can be safely removed without significantly affecting CMA. This is a major finding in order to reduce computational complexity. Highly complex schedules are very rare and therefore cannot have a large impact on CMA. Additionally, there is a large probability that these complex schedules will be excluded in the post-processing stage.
- The effect of imposing that a home-activity cannot be replaced by a wildcard is slightly negative but very minimal. This setting effectively limits the reduction in complexity of a schedule that the sequential clustering algorithm can reach. It is reassuring that this only marginally impacts classification accuracy. It strengthens the belief that home-activities are truly part of a skeleton schedule and that this setting can safely be imposed.

Considering all of these effects and other practical issues, a practical optimum in test set-CMA of 32% was found. This figure may be compared to the *null model accuracy* of 13.3%. The latter can be interpreted as the accuracy according to a random draw from the classification attribute distribution. One observes that the choice problem is intrinsically difficult, but that the practical optimum does improve the decision process considerably.

3. Results

Taking into account the conclusions from the analysis stage, together with some practical and intuitive considerations, the results of two distinct runs are presented in Table 2. Both runs of the algorithm differed only by the setting of the minimum number of remaining activities per wildcard-containing schedule (setting 1). In the first run this minimum was set to three and in the second run it was set to two. Settings 2, 3 and 4 were enabled. No input data was excluded according to setting 5.

Of the approx. 2,600 unique single-day schedules that were constructed from the OVG survey data, only 733 unique skeleton schedules remained after the first run. After the second run, even only 341 remained. Of these 341 unique skeleton schedules, merely 14 unique skeleton schedules are needed to comply with approx. 70% of all given schedules (this was only 45% in the original data). Observe how none of the skeleton schedules can be a subset of another one while still complying to all the settings. The exemplary outlier (*odd pattern*) identification process in Figure 2 is that of the second run.

One observes that well-defined schedules containing few activities are retained by the algorithm, complying to setting 1. These occur with a relative high frequency, therefore this is justified. Complex schedules seem to be broken down in distinct home-based subtours, where in case of run 1 very common subtours such as a home-based shopping tour are maintained (e.g. 4-1-X-1), whilst in run 2 these are aggregated into the very basic skeleton of multiple home-based subtours (e.g. X-1-X-1). This behavior is sensitive to setting 1.

Figure 3 illustrates the skeleton pattern frequency over the days of the week. Even at a the aggregation level of run 2 some interesting patterns are visible. For instance, at Wednesdays the schedule 'home-educational-home' seems to be reduced in favor of the double bare home-based subtour schedule (X-1-X-1). In Belgium children only go to school for half a day on Wednesdays, leaving more time available for other activities. Often parents also make time to escort their children on these activities. Also on Fridays this effect is noticeable. People may perform more activities and later in the evening, most likely because of the start of the weekend.

This trend extends into the weekend during which more diverse and complex schedules skeletons with multiple bare home-based tours have a higher probability. Only on Fridays and Saturdays does the skeleton pattern 'X-1-X-1-X-1-X-1' fall within the topmost 70%. On Saturday most shops and other venues are opened, other than on Sunday when very few are opened. This is also observed in the skeleton schedules: 'home-shopping-home' has a high probability to occur on a Saturday, as well as 'home-recreational-home' schedules. On Sunday, the latter occurs even more frequently, but shopping activities as the only out-of-home activity of the day are however less common, having a similar probability as on weekdays. Sunday also seems most convenient to do a visiting activity.

4. Discussion

To gain more insight into the clustering process, the distribution of activities that were replaced by wildcards in a wildcard-containing schedule can be checked. As can be expected, the activity distribution is quite complex. After all, the intent of the proposed sequential clustering algorithm was to reveal common patterns and separate them from complex behavior. The distributions contain many combinations of activities, especially since a wildcard can replace multiple activities. In most cases there is no single distinct combination of activities with a considerable higher frequency than subsequent combinations. This confirms that the algorithm did not discard common travel behavior from the skeleton schedules.

The presented methodology has some limitations towards the use in ABMs. The fact that currently only single day schedules of individuals are taken into account potentially ignores some multi-day or household-related patterns. Unfortunately the OVG survey data does not support such research, however these decision levels might be incorporated elsewhere in an ABM and therefore are not insurmountable. This is a limitation in many ABMs.

Table 2. Schedules after pre-processing and sequential clustering algorithm runs. The list was prematurely ended after 30 schedules. Records in **bold** and underscored highlight a cumulative frequency higher than 70%, respectively 80%. An "X" represents a wildcard. Table 1 lists the activity encoding. The initial "home"-activity of all schedules was dropped.

	After pre-processing			Run 1 (setting 1=3)			Run 2 (setting 1=2)		
#	Pattern	Freq. [%]	Cum. Freq. [%]	Pattern	Freq. [%]	Cum. Freq. [%]	Pattern	Freq. [%]	Cum. Freq. [%]
1	3-1	11.14	11.14	3-1	11.14	11.14	X-1-X-1	16.94	16.94
2	4-1	8.30	19.44	4-1	8.30	19.44	3-1	11.14	28.07
3	6-1	5.84	25.28	6-1	5.84	25.28	4-1	8.30	36.37
4	5-1	4.85	30.13	4-1-X-1	5.48	30.75	6-1	5.84	42.21
5	9-1	4.76	34.89	X-1-X-1-X-1	5.19	35.94	X-1-X-1-X-1	5.19	47.40
6	10-1	1.73	36.62	5-1	4.85	40.79	5-1	4.85	52.25
7	4-1-5-1	1.50	38.12	9-1	4.76	45.55	9-1	4.76	57.01
8	2-1	1.40	39.52	3-1-X-1	3.53	49.08	3-1-X-1	3.53	60.54
9	7	1.24	40.75	X-1-9-1	2.86	51.94	3-X-1	2.87	63.41
10	8-1	1.19	41.94	10-1	1.73	53.67	10-1	1.73	65.14
11	4-1-9-1	1.16	43.10	X-1-5-1	1.53	55.20	X-4-1	1.53	66.67
12	6-1-9-1	1.09	44.20	2-1	1.40	56.60	X-5-1	1.48	68.15
13	4-1-4-1	1.03	45.22	7	1.24	57.83	2-1	1.40	69.55
14	3-1-9-1	0.89	46.12	X-1-4-1	1.24	59.07	**7**	**1.24**	**70.78**
15	11-1	0.84	46.95	8-1	1.19	60.26	**8-1**	**1.19**	**71.97**
16	3-2-1	0.79	47.74	X-1-X-1-X-1-X-1	1.08	61.34	**X-9-1**	**1.12**	**73.10**
17	3-1-3-1	0.75	48.48	X-1-8-1	0.91	62.24	**X-1-X-1-X-1-X-1**	**1.08**	**74.17**
18	3-1-4-1	0.71	49.19	X-4-1-X-1	0.90	63.14	**X-9-X-1**	**0.98**	**75.15**
19	8-1-8-1	0.61	49.80	11-1	0.84	63.98	**X-3-1**	**0.90**	**76.05**
20	3-4-1	0.57	50.37	3-X-1-X-1	0.82	64.80	**X-1-7**	**0.90**	**76.95**
21	9-1-9-1	0.48	50.85	3-2-1	0.79	65.59	**X-3-X-1**	**0.86**	**77.81**
22	3-1-5-1	0.47	51.32	X-1-3-1	0.78	66.37	**11-1**	**0.84**	**78.65**
23	7-1	0.47	51.80	3-X-3-1	0.76	67.12	**3-X-1-X-1**	**0.82**	**79.47**
24	4-5-1	0.44	52.24	3-1-3-1	0.75	67.87	**X-11-X-11**	**0.81**	**80.29**
25	8-3-8-1	0.43	52.67	3-1-X-1-X-1	0.72	68.59	**X-4-X-1**	**0.80**	**81.09**
26	3-1-8-1	0.43	53.10	8-3-X-1	0.65	69.24	**X-1-3-1**	**0.78**	**81.87**
27	10-4-1	0.43	53.53	3-4-1	0.57	69.81	**3-X-3-1**	**0.76**	**82.62**
28	6-1-6-1	0.42	53.95	**6-1-X-1**	**0.56**	**70.37**	**3-1-3-1**	**0.75**	**83.37**
29	9	0.41	54.35	**7-1**	**0.47**	**70.84**	**3-1-X-1-X-1**	**0.72**	**84.09**
30	8-3-1	0.40	54.75	**8-X-8-1**	**0.47**	**71.31**	**X-1-X-1-X**	**0.71**	**84.80**

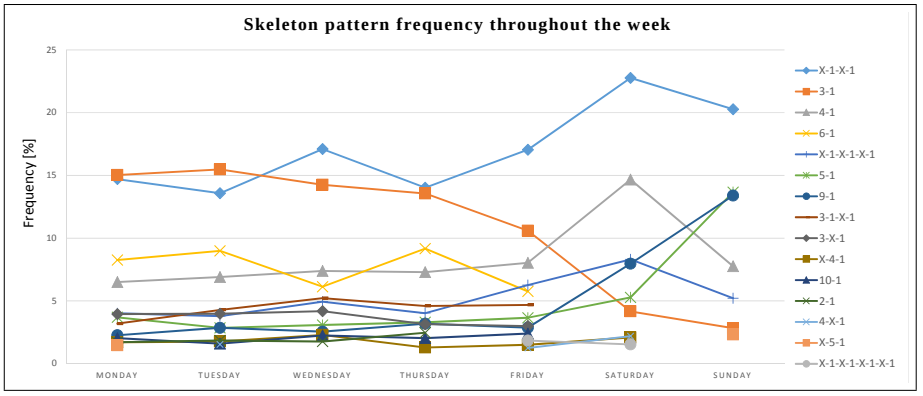

Fig. 3. Comparison of skeleton activity pattern frequency [%] throughout the week, based on subsets of OVG 3.0-4.5. Identical settings as in run 2 were used: the minimum number of non-wildcard activities is set to two and all other settings were enabled. A cutoff of 70% (cumulative) was used. An "X" represents a wildcard. Table 1 lists the activity encoding. The initial "home"-activity of all schedules was dropped.

Additionally, the temporal component of the activities within the schedule is not determined. Still, this should not be a limitation as, upon closer inspection, the temporal distribution of the activities within the skeleton schedules are quite well-defined, especially when taking into account socio-demographic profiles and the day of the week. This discussion is however out of the scope of the current paper.

Furthermore, the modeler has to pay special attention to the number of trips within a schedule. This is an important evaluation criterion of an ABM. The step of merging multiple identical consecutive activities into a single occurrence

augments the risk to underestimate the number of performed trips. Consequently, this requires to be compensated in subsequent modeling stages.

Despite these restrictions, the currently proposed sequential clustering algorithm excels through its simplicity and general applicability. In essence, the algorithm consists only of one iterative process with a simple objective, and some optional boundary conditions. It is fully compatible with a staged activity schedule prediction (first a skeleton schedule, subsequently completed with more discretionary activities), which is currently the most widely adopted approach. By contrast, this is not the approach in the HAPP framework. An extension by Allahviranloo et al.[8] mitigates this 'shortcoming', but is built from more complex algorithms such as affinity propagation and k-means clustering. Although some results are very similar, the added value of a more complex methodology and implementation is not guaranteed. Additional research might compare several methods on practicality and modeling performance.

5. Conclusions and Further Research

Traditionally, activity-based models first schedule work and educational activities as these are assumed as most inflexible and of high priority. However, it was shown that there is more information available in the form of skeleton schedules that can be used as nuclei in activity-based modeling. To accommodate such an approach, a completely data-driven method to reveal skeletons of common activity sequences in single-day schedules was devised. In other words, the concept of mandatory activities was extended to skeleton sequences, where skeleton sequences take the form of frequently occurring sequences including wildcards.

To this end, a sequential clustering algorithm was developed, consisting of a pre-processing step, a sequential clustering step and a post-processing step. The effects of the algorithm settings on the classification accuracy of the skeleton patterns by the ID3 decision tree classifier were analyzed. This analysis approach closely resembles the decision process in ABMs such as ALBATROSS[6] and FEATHERS[10]. Two representative runs of the algorithm were analyzed and discussed. The proposed method reveals a set of skeleton activity schedules and confirms the importance of work and education.

Further research on this topic should consider the practical integration into an activity-based modeling environment, also addressing the limitations discussed before. Furthermore, alternatives to the ID3 classification model should be explored to improve 'prediction' accuracy. Additionally, the algorithm could be executed on different study areas and see the methodology to reveal single-day skeletons validated.

References

1. Recker, W., McNally, M., Root, G.. A model of complex travel behavior: Part I - Theoretical development. *Transportation Research Part A: General* 1986;**20**(4):307–318. doi:10.1016/0191-2607(86)90090-7.
2. Recker, W., McNally, M., Root, G.. A model of complex travel behavior: Part II - An operational model. *Transportation Research Part A: General* 1986;**20**(4):319–330. doi:10.1016/0191-2607(86)90090-7.
3. Miller, E.J.. Propositions for Modelling Household Decision - Making. In: Lee-Gosselin, M., Doherty, S., editors. *Presented at International Colloquium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Assumptions and New Conceptual Frameworks*. Oxford: Elsevier; 2002, p. 21–60.
4. Garikapati, V.M., You, D., Pendyala, R.M., Vovsha, P.S., Livshits, V., Jeon, K.. Multiple Discrete-Continuous Model of Activity Participation and Time Allocation for Home-Based Work Tours. *Transportation Research Record: Journal of the Transportation Research Board* 2014;**2429**:90–98. URL: http://trb.metapress.com/openurl.asp?genre=article{&}id=doi:10.3141/2429-10. doi:10.3141/2429-10.
5. Doherty, S.. 2005 Fred Burggraf Award, Planning and Environment: How Far in Advance Are Activities Planned?: Measurement Challenges and Analysis. *Transportation Research Record* 2005;**1926**(1):40–49. doi:10.3141/1926-06.
6. Arentze, T., Hofman, F., van Mourik, H., Timmermans, H.. ALBATROSS: Multiagent, Rule-Based Model of Activity Pattern Decisions. *Transportation Research Record* 2000;**1706**(1):136–144. doi:10.3141/1706-16.
7. Arentze, T., Hofman, F., Timmermans, H.. Reinduction of Albatross Decision Rules with Pooled Activity-Travel Diary Data and an Extended Set of Land Use and Cost-Related Condition States. *Transportation Research Record* 2003;**1831**(1):230–239. doi:10.3141/1831-26.
8. Allahviranloo, M., Regue, R., Recker, W.. Pattern Clustering and Activity Inference. In: *TRB 93rd Annual Meeting Compendium of Papers*. 2014, .
9. Roorda, M.J., Ruiz, T.. Interpersonal Commitments and the Travel/Activity Scheduling Process. In: *11th World Conference on Transport Research*. Berkeley CA; 2007, URL: http://www.wctrs-society.com/wp/wp-content/uploads/abstracts/berkeley/D5/1103/Roorda_Ruiz_WCTR2007paper_CDresubmission.doc.
10. Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., Timmermans, H.. Implementation Framework and Development Trajectory of FEATHERS Activity-Based Simulation Platform. *Transportation Research Record: Journal of the Transportation Research Board* 2010; **2175**(-1):111–119. doi:10.3141/2175-13.