A Bioinspired Sequence Alignment Approach for Modelling Weekly Activity Patterns

Md. Rifat Hossain Bhuiyan*[1], Muhammad Ahsanul Habib[2]

[1] PhD Student, Department of Civil and Resource Engineering, Dalhousie University, Halifax, Nova Scotia, Canada

[2] Professor, School of Planning, and Department of Civil and Resource Engineering, Dalhousie University, Halifax, Nova Scotia, Canada

SHORT SUMMARY

This research proposes a multi-module framework to derive weekly representative travel patterns from single-day travel diaries. It employs hierarchical clustering to group samples with homogeneous activity patterns, followed by progressive multiple sequence alignment to construct day-level representative patterns. These day-level patterns are then adjoined based on similarity to obtain week-level representative activity patterns, ultimately producing archetypal weekly pseudo-diaries. The analysis of weekly activity patterns based on the formulated longitudinal data revealed diverse insights into weekly travel behaviours. Results show that a distinct difference exists in weekly activity patterns and time-use behaviour across population groups. For working groups, shorter work durations on Fridays were observed, and the weekly work duration for teleworkers was found to be lower than that of workplace workers. The proposed framework represents a significant step towards multi-day activity-based travel demand modelling, especially in the face of data limitations.

Keywords: Travel Survey, Multi-day Data, Multiple Sequence Alignment, Weekly Travel Pattern.

1. INTRODUCTION

Transportation planners and professionals largely rely on Household Travel Surveys (HTS) to examine peoples' travel behaviour and system performances as these surveys provide rich household, individual, and travel related information. These surveys are crucial for understanding the underlying influence of socio-economic factors on travel choices. Besides, data from the travel diaries in HTS reveals key trip attributes such as distance, time, mode, purpose, and cost, which are essential for developing travel demand models. Most HTS, especially those used for mainstream modelling, are cross-sectional, meaning they are conducted as one-off surveys at a specific point in time (Stopher et al., 2008). The cross-sectional travel surveys generally collect respondents' activity-travel information for a 'typical day' of the week. The underlying assumption behind such approach is that the travel patterns for weekdays are mostly similar (Verreault & Morency, 2011). If travel behaviour is reported for a randomly chosen day within a longer period, it provides an unbiased sample of behaviour for that timeframe (Pas & Sundar, 1995). Single-day travel surveys are commonly conducted to gather information for different weekdays, leading to unbiased samples of average weekday travel behaviour. Thus, the implicit assumption is that, by randomly sampling households and individuals on a random weekday, the resulting data will accurately represent the travel behaviour of the overall population (Stopher & Zhang, 2011). Though single-day cross-sectional surveys remain appealing due to the ease of acquiring a large sample size and acceptable population representation, they cannot capture the variability in individual and household travel patterns across different days of the week. Hence, to model travel demand over an extended period, while accounting for day-to-day variability, multi-day travel data is necessary.

An alternative approach to multi-day data collection involves formation of multi-day pseudo-diaries from single-day data. This can be achieved through data fusion with large passive datasets or sampling from existing single-day household travel survey data to generate multi-day activity-travel data. These methods aim to leverage all available data sources while mitigating their respective drawbacks. This research proposes a novel framework to derive weekly representative travel patterns from single-day travel diaries. The approach uses multiple sequence alignment with hierarchical clustering, followed by progressive sequence alignment to construct day-level and week-level representative patterns, ultimately producing archetypal weekly pseudo-travel diaries. While this research does not seek to enhance the sequence alignment approach itself, it introduces a novel framework to address a critical concern in travel behaviour research: maximizing the utility of single-day activity-travel data. By effectively sampling a diverse set of activity-travel patterns, the proposed framework allows them to serve as surrogates for multi-day activity-travel data.

2. DATA AND CONCEPT

Data Description

This study utilizes data from the Halifax Travel Activity Survey (HaliTRAC) conducted from July 2022 to March 2023**.** In total, 2,295 individuals responded, detailing activities for a designated travel day (3 AM to 3 AM), including locations, nature of activities, departure and arrival times, travel modes, and other trip-related information. The data from the 24-hr travel log was processed to generate unidimensional (only temporal dimension, no spatial consideration) activity sequences for each individual. By dividing the 24-hr (1440 min) time budget into 15-minute intervals and 10 distinct activities, 96-character activity-sequences were prepared. This approach included defining 10 activity types, character coding each activity, examining time-use per activity from the travel log to represent each 15-minute time slot with an activity character. Activities lasting less than 15 minutes were ignored, and the remaining activities were rounded up or down to the nearest 15-minute interval. Figure 1 shows a generic representation of an activity-sequence with time stamps and Table 2 presents the 10 activity categories, their character codes, and description along with daily time use statistics.

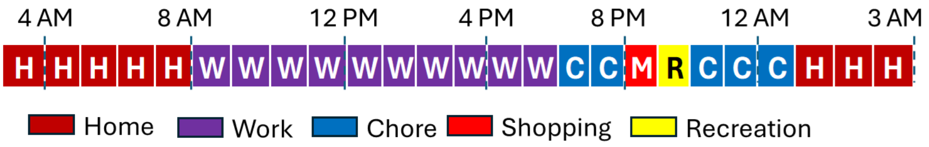


Figure 1: Representation of an activity pattern sequence

Table 2: Descriptive Statistics of demographics and daily activities

Demography		**N**	**%**		**Census (%)**	**Description**
Age	<25	292	12.72		12.9	Respondents’ age distribution.
	25-55	845	36.82		43.7	
	55-75	951	41.44		34	
	>75	207	9.02		9.1	
Gender						
	Male	1180	51.42		48.7	Respondents’ self-reported gender identity.
	Female	1115	48.58		51.3	
Employment						
	Full-time	1022	44.53		51.9	Respondents' employment status.
	Part-time	73	3.18			
	Retired	930	40.52		40.5	
	Student	270	11.76		7.6	
Education						
	University	1271	55.38		33.7	Respondents' education level.
	College	377	16.43		19.8	
	High-school	494	21.53		25.3	
	No degree	153	6.67		12.1	
Daily Activity	**Code**	**Total 15-Min Episodes**	**Sample Average Duration (hr.)**	**S.D. (hr.)**	**Census Average Duration (hr.)**	**Description**
In home	H	151103	16.46	4.51	8.9	In-home activities such as home leisure and sleep, excluding working from home and in-home chores. (census considers only sleep activities)
Work at workplace	W	19699	2.15	3.8	3.5	Work activities, including regularly scheduled work and work-related tasks such as calls and meetings.
Work from home	T	7317	0.8	2.41	-	Performing paid work from home
School	S	5448	0.59	2.01	0.6	Daycare, school, or activities related to school.
Shopping	M	5875	0.64	1.55	0.4	Shopping for goods and services, routine shopping.
Escort	E	1315	0.14	0.68	0.9	Pickup/drop off.
Recreation	R	9050	0.99	1.9	0.8	Eating out, meeting friends, and other entertainment activities.
Home chore	C	15259	1.66	3.01	2.1	Eating or meal preparation, cleanin, home maintenance, childcare, and other in-home activities.
Personal business	P	3380	0.37	1.24	0.8	Fitness activities, medical, and banking.
Other activities	O	1874	0.2	0.94	0.2	All non-routine, non-traditional activities.

While Table 2 demonstrates that the sample data is demographically representative of the population at an acceptable level (~5%), this research further leverages the sequential nature of the dataset by visualizing the start-time, end-time, and duration of daily activities for all defined activities, including in-home activities.

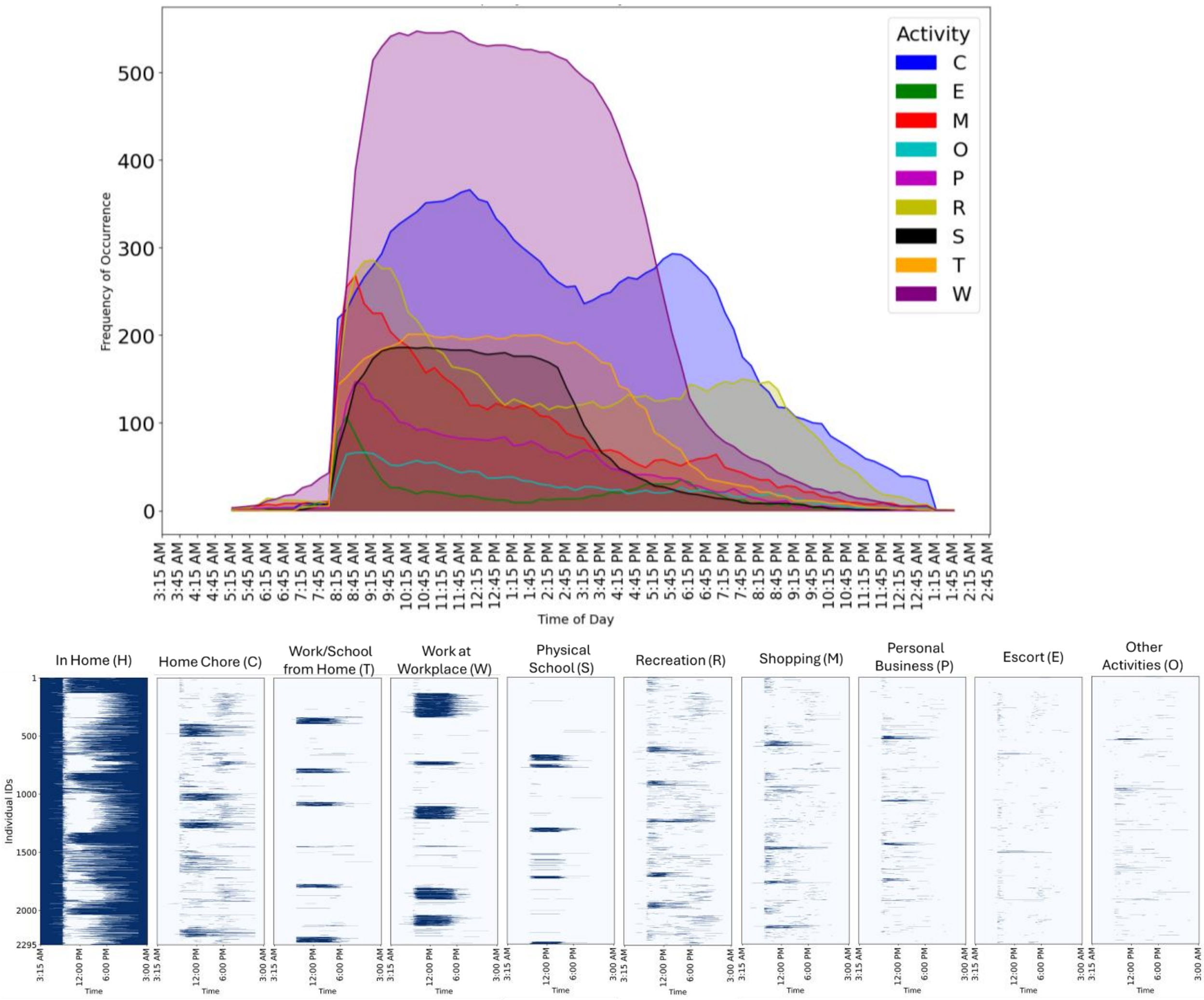


Figure 2: (a) Activity start-time, end-time and duration (b) density plot of the dataset

Figure 2(a) illustrates the frequency of occurrence of specific activities at various times throughout the day, providing some interesting insights about activity start time, end-time, duration. Figure 2(b), on the other hand, presents the density plot of activity sequences for all 2,295 individuals, revealing time spans with higher durations of both in-home and out-of-home activities.

Conceptual Framework

The framework of this study comprises of four modules. The modules are briefly discussed in this section. Figure 3 below provides a visual demonstration of the framework.

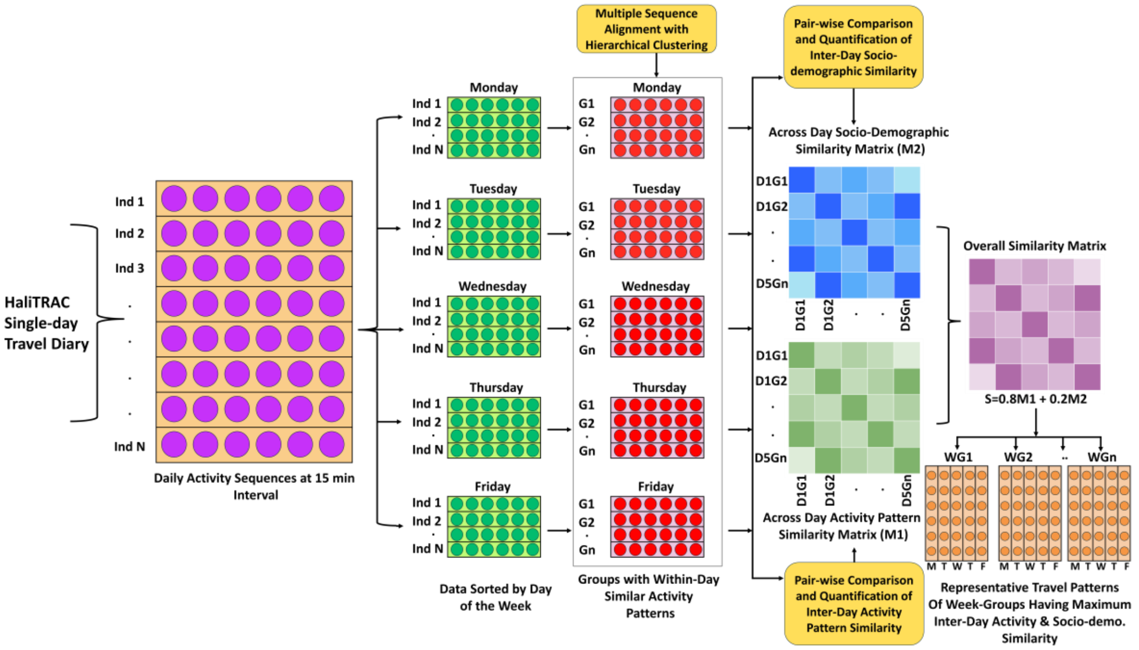

Multiple Sequence Alignment with Hierarchical Clustering
Pair-wise Comparison and Quantification of Inter-Day Socio-demographic Similarity
HaliTRAC Single-day Travel Diary
Ind 1
Ind 2
Ind 3
Ind N
Daily Activity Sequences at 15 min Interval
Monday
Tuesday
Wednesday
Thursday
Friday
Ind 1
Ind 2
Ind N
Data Sorted by Day of the Week
G1
G2
Gn
Groups with Within-Day Similar Activity Patterns
Across Day Socio-Demographic Similarity Matrix (M2)
D1G1
D1G2
D5Gn
Across Day Activity Pattern Similarity Matrix (M1)
Pair-wise Comparison and Quantification of Inter-Day Activity Pattern Similarity
Overall Similarity Matrix
S=0.8M1 + 0.2M2
WG1
WG2
WGn
M T W T F
Representative Travel Patterns Of Week-Groups Having Maximum Inter-Day Activity & Socio-demo. Similarity

Figure 3: The Conceptual Framework of this Research

Firstly, the 2295 activity sequences were divided into 5 data sets based on the day for which the responses were recorded. Hierarchical clustering using Levenshtein distance was then applied to each dataset to form 'within-day similar' groups with homogeneous activity-travel patterns. Secondly, the Needleman-Wunsch algorithm was used for pairwise alignment within these groups to measure the levels of similarity and dissimilarity. Thirdly, progressive multiple sequence alignment was conducted on each 'within-day similar' group to create single representative activity sequences. For instance, if hierarchical clustering identifies 11 groups (each group contains several similar sequences), progressive alignment generates 11 representative sequences. This process involves utilizing pairwise alignment scores from module 2, followed by sequence alignment of the closet pairs based on a dynamically updating distance matrix. Detailed methodology and processes are discussed in the methodology and analysis section. The fourth module involves quantifying across-day similarities of representative patterns (using pairwise alignment) and demographic profiles (based on the Jaccard Index). These similarity scores are then merged to obtain overall similarity scores, identifying population groups with high similarity levels across different days to construct the longitudinal data structure.

3. METHODS

Hierarchical Clustering Using Levenshtein Distance

This method combines Levenshtein distance, a measure of dissimilarity between sequences, with hierarchical clustering to group sequences based on their similarity. The goal was to cluster 'within-day similar' groups exhibiting homogeneous activity-travel patterns. Levenshtein distance, also known as edit distance, is the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one sequence into another. Given two sequences a and b of lengths |a| and $|b|$respectively, the Levenshtein distance $lev_{a,b}(i,\ j)$ is defined as-

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if \min(\mathrm{i},j) = 0 \\ min \begin{cases} lev_{a,b}\ (i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+\delta(a_i,b_j) \end{cases} & \text{otherwise,} \end{cases} \quad (1)$$

Where, $\delta(a_i, b_j)$ is the indicator function that equals to 0 when $a_i = b_j$ and 1 otherwise.

Pairwise Alignment

Pairwise alignments compare two sequences, the source and target, to align equivalent elements, known as matches. To equalize lengths, gaps are inserted. Alignments are evaluated using similarity or distance measures, with algorithms aiming to maximize similarity or minimize distance. This process can be local or global. Global alignment compares sequences across their entire length. This paper uses the Needleman-Wunsch algorithm for pairwise alignment scores, which involves initialization, matrix filling, and traceback. While the mathematical framework is not detailed here (Jackson & Aluru, 2005), we focus on applying the method to activity-travel patterns with a simple example. Two unidimensional 8-character activity sequences are-

$Seq1 = HHMMSSHH$; and (2)
$Seq2 = HHMSSSHH$

While results in classifying sequences with known patterns can be achieved by setting any gap penalties, this research aligns with the scoring schemes used in previous similar studies (Saneinejad & Roorda, 2009; C. Wilson et al., 1999). Hence, the following scoring-scheme for match, mismatch, gap insertion, and gap extension were followed- Match score: 10; Mismatch score: 0; Gap insertion: -3; Gap extension: -3.

To determine the optimal alignment, a traceback is performed from (8,8) to (0,0), following the path of maximum scores. This alignment shows: Matches: H, H, M, S, S, H, H (7 matches), and Mismatch: M!= S (1 mismatch). Since the maximum possible score is $8 \times 10 = 80$ given all positions matched and the obtained alignment score is 70, the percentage similarity is-

$$Percentage\ Pairwise\ Similarity = \left(\frac{F(8,8)}{Max.Score}\right) \times 100 = \left(\frac{70}{80}\right) \times 100 = 87.5\% \quad (3)$$

Progressive Alignment

Progressive alignment is a method used in bioinformatics to create multiple sequence alignments (MSA) incrementally. It involves three main steps: calculating pairwise alignment scores, constructing a guide tree, and aligning sequences based on the tree. Typically, guide trees are built using methods like UPGMA or Neighbor-Joining, reflecting evolutionary distances. However, for large datasets, the computational load is high. This research simplifies the process by merging sequences based on pairwise distances without using a guide tree. The iterative merging follows four steps: identifying the closest pair, aligning them, creating a new sequence, and updating the distance matrix, repeated until all sequences are aligned.

Transcending to Multi-day

To extend this analysis across multiple days, we performed pairwise alignment on all identified group representative patterns (from Monday to Friday) to examine across-day similarities, following the methodology described in Section 3.2. This analysis revealed new 'week-groups' exhibiting similar activity patterns throughout the week. However, forming 'week-groups' based solely on activity pattern similarities is not comprehensive, as literature suggests that socio-demographic profiles significantly influence travel behaviour (Bhat, 1996; Hanson, 1982; Hensher & Rose, 2007; Meloni et al., 2009; Pas, 1984; Strathman et al., 1994; Veterník & Gogola, 2017; Xianyu et al., 2017). To address this, we traced back the socio-demographic profiles of the groups and quantified the socio-demographic similarities between these groups using the Jaccard Similarity Index. Jaccard Similarity is a widely used metric for comparing the similarity and diversity of sample sets particularly having large numbers of categorical data. The mathematical formulation of this approach is-

$$J\,(A,B) = \left|\frac{A \cap B}{A \cup B}\right| \tag{4}$$

Attributes	Ind1 (A)	Ind2 (B)	A∩ B	A∪ B
Age (<25)	1	0	0	1
Age (>25)	0	1	0	1
Male	1	1	1	1
Female	0	0	0	0
Student	1	0	0	1
Worker	0	1	0	1
		Sum	1	5

Where, A and B are the comparison sets of attributes. $|A \cap B|$ represents the number of elements in the intersection of sets A and B, whereas $|A \cup B|$ represents the number of elements in the union of A and B. For example, the matrix above lists the demographic attributes of two individuals (Ind1, Ind2). From the attribute matrix, we find that $(A \cap B) = 1$ and $(A \cup B) = 5$. Hence, $J\,(A,B) = \left|\frac{A \cap B}{A \cup B}\right| = \frac{1}{5} = 0.2$. Thus, from analysis, 20% similarity between the demographic attributes was observed. After performing demographic and activity pattern similarity measurements, the scores were combined to achieve an overall similarity score, defined as- $S = 0.8\ \times Pattern\ Similarity + 0.2\ \times demographic\ similarity$

4. RESULTS

Clustering Results

The hierarchical clustering process identified 'within-day similar' groups exhibiting homogeneous activity-travel patterns. To keep the results succinct, cluster analysis for all five datasets is not included here. Instead, we present a dendrogram to visualize the arrangement of clusters produced by the hierarchical clustering algorithm specifically for Thursday (Day 4). In hierarchical clustering, the elbow plot is utilized to determine the optimal number of clusters by plotting the distance at which clusters merge at each step of the clustering process. The point where the rate of increase in ‘merge distance’ sharply changes indicates the most appropriate number of clusters. For this case, nine distinct clusters were identified. This study defines representative activity patterns as generalized sequences that capture common features within a group’s activity-travel patterns. Using pairwise similarity matrices and progressive multiple sequence alignment, we derived these patterns for identified groups. Figure 5 specifically illustrates Tuesday’s (Day 2) patterns, simplified for easier plotting from 15-minute interval data.

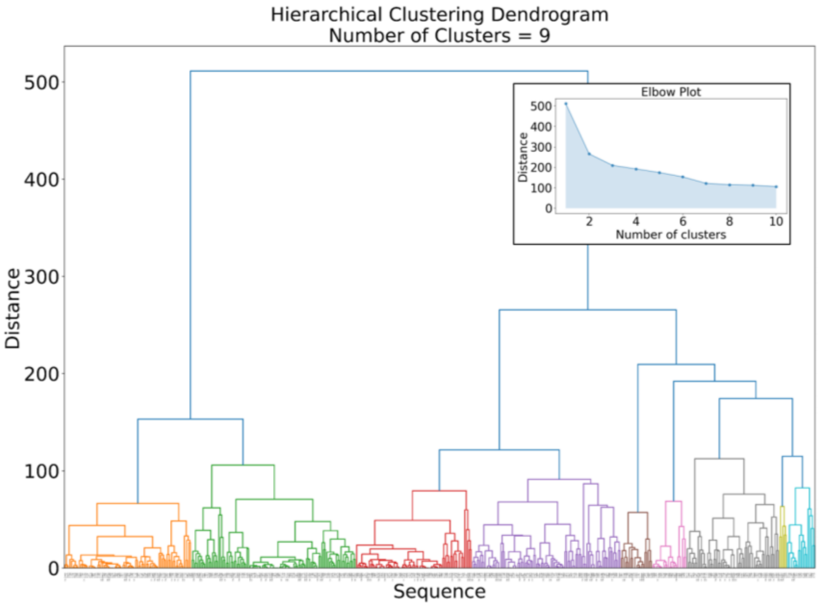
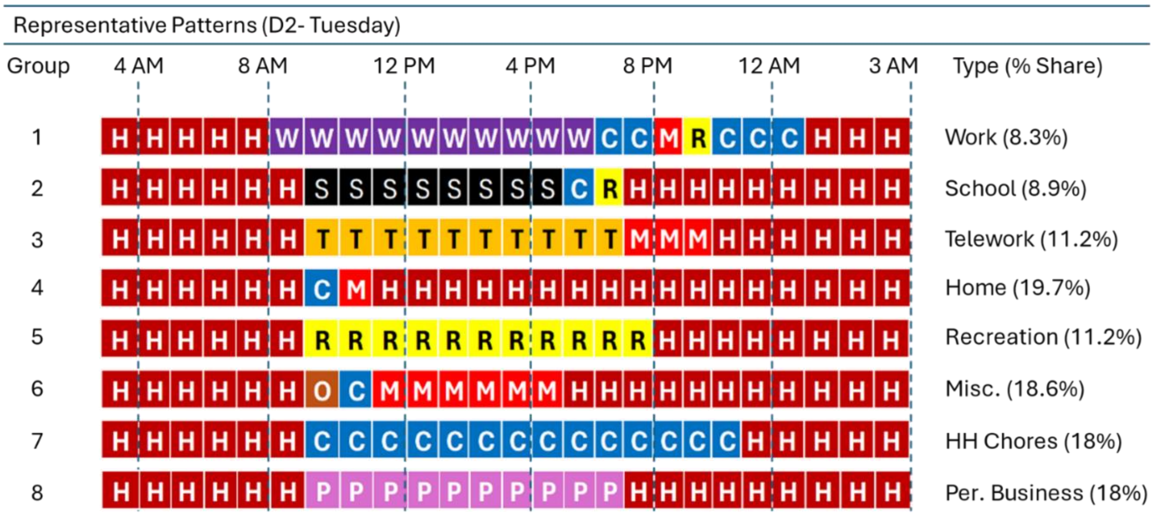


Figure 4: Dendrogram of clusters produced by hierarchical clustering algorithm



Representative Patterns (D2- Tuesday)

Group	4 AM				8 AM				12 PM				4 PM				8 PM				12 AM				3 AM	Type (% Share)
1	H	H	H	H	H	W	W	W	W	W	W	W	W	W	W	C	C	M	R	C	C	C	H	H	H	Work (8.3%)
2	H	H	H	H	H	H	S	S	S	S	S	S	S	S	C	R	H	H	H	H	H	H	H	H	H	School (8.9%)
3	H	H	H	H	H	H	T	T	T	T	T	T	T	T	T	T	M	M	M	H	H	H	H	H	H	Telework (11.2%)
4	H	H	H	H	H	H	C	M	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	Home (19.7%)
5	H	H	H	H	H	H	R	R	R	R	R	R	R	R	R	R	R	H	H	H	H	H	H	H	H	Recreation (11.2%)
6	H	H	H	H	H	H	O	C	M	M	M	M	M	M	H	H	H	H	H	H	H	H	H	H	H	Misc. (18.6%)
7	H	H	H	H	H	H	C	C	C	C	C	C	C	C	C	C	C	C	C	C	H	H	H	H	H	HH Chores (18%)
8	H	H	H	H	H	H	P	P	P	P	P	P	P	P	P	P	H	H	H	H	H	H	H	H	H	Per. Business (18%)

Figure 5: Representative activity patterns of the eight 'within-day similar' activity groups for Tuesday

Homogeneous Week-Group Formulation

To form homogeneous week-groups, we calculated both across-day and demographic similarities of group-level representative patterns, creating two score matrices. These matrices were combined with an 80/20 weightage to identify the most similar groups across the week. Figure 6 shows heatmaps of activity, demographic, and overall similarity matrices, where dark-red cells indicate high similarity. After applying a 60% similarity threshold, six final week-groups were identified. Figure 7 illustrates the composition of week-groups (WGs). Following this, demographic profiles

(Figure 8) and weekly activity patterns (Figure 9) were analyzed to explore time-use behaviour and day-to-day travel variations.

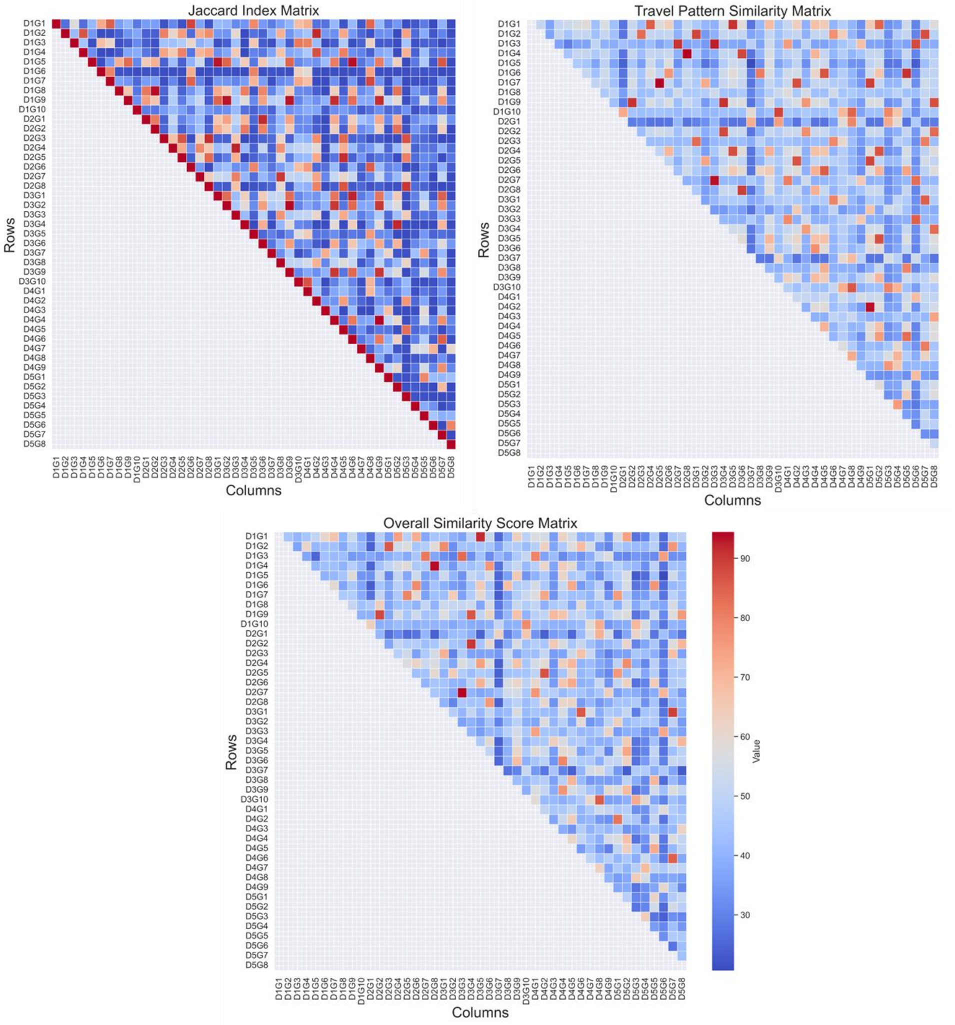


Figure 6: Demographic, Travel Pattern, and Overall Similarity Matrices

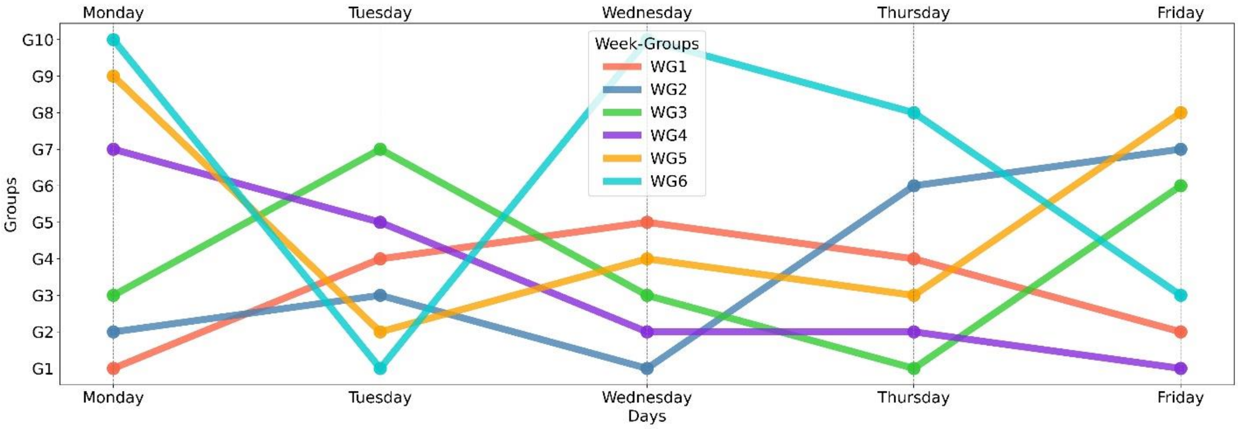
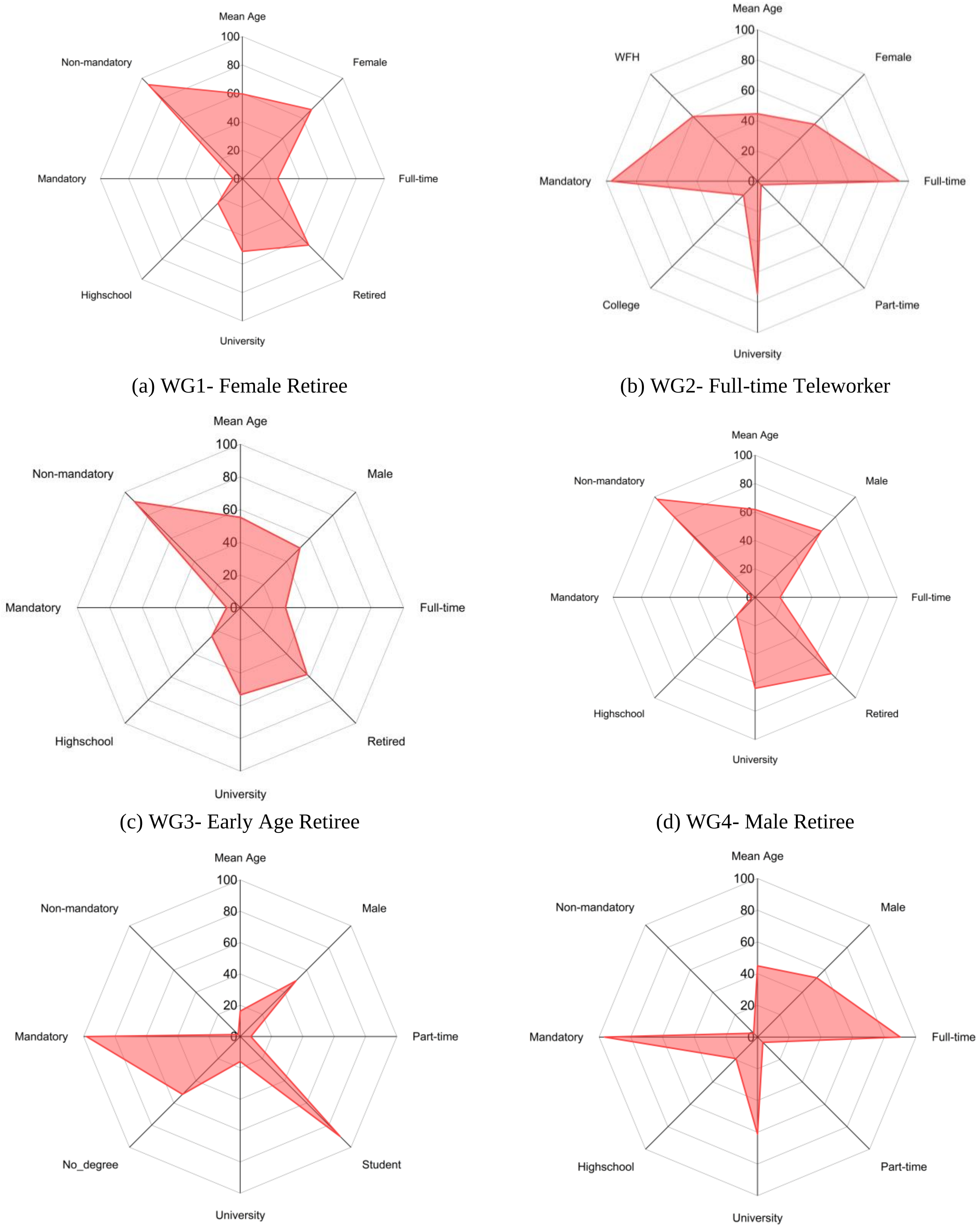


Figure 7: Identified Week-Groups with Highest Across-Day Similarity



(a) WG1- Female Retiree

(b) WG2- Full-time Teleworker

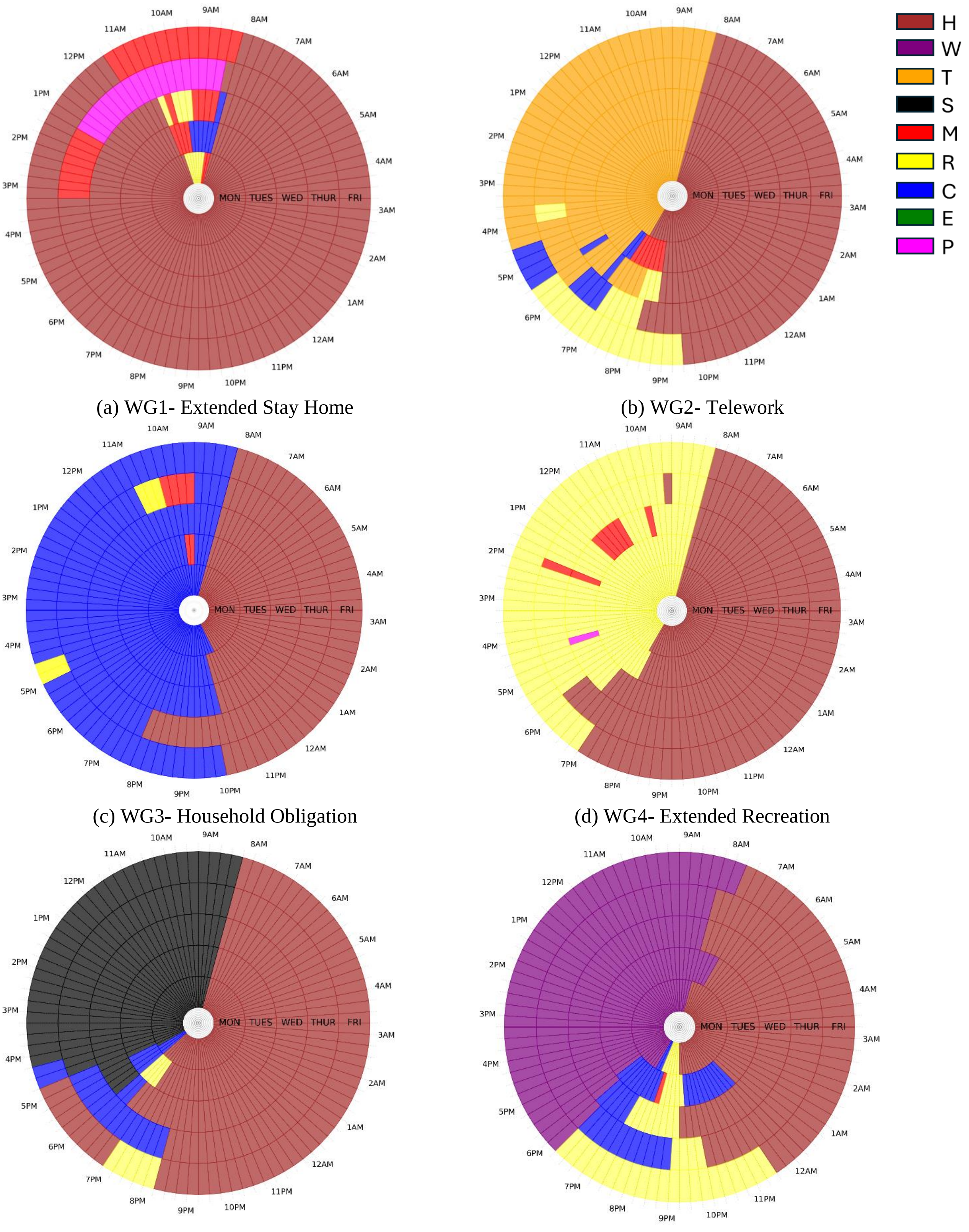
(c) WG3- Early Age Retiree

(d) WG4- Male Retiree

(e) WG5- Young Student

(f) WG6- Full-time Workplace Worker

Figure 8: Synthesis of Week-Groups' Demographic Profiles



(a) WG1- Extended Stay Home

(b) WG2- Telework

(c) WG3- Household Obligation

(d) WG4- Extended Recreation

(e) WG5- School

(f) WG6- Workplace Work

Figure 9: Weekly Activity and Time Use Pattern of the Week Groups

Demographic profiles and weekly activity patterns revealed key insights. Figure 8 shows simplified profiles, highlighting dominant attributes within each group. Figure 9 details 15-minute interval activity patterns. Week-Group 1 consists mainly of older females spending most of their time at home, while Week-Group 2 includes predominantly female remote workers with shorter work durations on Fridays. Week-Group 3 comprises early retirees engaging in non-mandatory activities. Week-Group 4 consists of male retirees focused on recreation. Week-Group 5 includes young students with varied school schedules, and Week-Group 6 is composed of full-time workers with consistent work durations. This framework provides a basis for multi-day travel demand modelling.

5. CONCLUSIONS

This study provides valuable insights into weekly travel behaviour by leveraging limited single-day data through a multi-module pattern recognition approach. The methodology includes hierarchical clustering, pairwise and progressive alignments, and incorporates socio-demographic similarities using the Jaccard Similarity Index. The analysis identified diverse travel behaviours across different week-groups, such as remote workers and retirees. Despite limitations, including the lack of original multi-day data, the framework reflects common travel patterns observed in recent studies. Future work will address variability in activities, integrating non-mandatory activities and weekly participation data to refine representative activity patterns for more accurate modelling.

ACKNOWLEDGEMENTS

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, Climate Action Awareness Fund (CAAF), and Halifax Regional Municipality (HRM) for supporting this research.

REFERENCES

Bhat, C. R. (1996). A generalized multiple durations proportional hazard model with an application to activity behavior during the evening work-to-home commute. *Transportation Research Part B: Methodological*, *30*(6), 465–480. https://doi.org/https://doi.org/10.1016/0191-2615(96)00007-0

Hanson, S. (1982). The determinants of daily travel-activity patterns: relative location and socio-demographic factors. *Urban Geography*, *3*(3), 179–202.

Hensher, D. A., & Rose, J. M. (2007). Development of commuter and non-commuter mode choice models for the assessment of new public transport infrastructure projects: A case study. *Transportation Research Part A: Policy and Practice*, *41*(5), 428–443. https://doi.org/https://doi.org/10.1016/j.tra.2006.09.006

Jackson, B. N., & Aluru, S. (2005). Pairwise sequence alignment. *Handbook of Computational Molecular Biology*, 1.

Meloni, I., Bez, M., & Spissu, E. (2009). Activity-Based Model of Women's Activity–Travel Patterns. *Transportation Research Record*, *2125*(1), 26–35. https://doi.org/10.3141/2125-04

Pas, E. I. (1984). The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environment and Planning A*, *16*(5), 571–581.

Pas, E. I., & Sundar, S. (1995). Intrapersonal variability in daily urban travel behavior: some additional evidence. *Transportation*, *22*, 135–150.

Saneinejad, S., & Roorda, M. (2009). Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. *Transportation Letters*, *1*(3), 197–211. https://doi.org/10.3328/TL.2009.01.03.197-211

Stopher, P. R., Kockelman, K., Greaves, S. P., & Clifford, E. (2008). Reducing Burden and Sample Sizes in Multiday Household Travel Surveys. *Transportation Research Record*, *2064*(1), 12–18. https://doi.org/10.3141/2064-03

Stopher, P. R., & Zhang, Y. (2011). Repetitiveness of daily travel. *Transportation Research Record*, *2230*(1), 75–84.

Strathman, J. G., Dueker, K. J., & Davis, J. S. (1994). Effects of household structure and selected travel characteristics on trip chaining. *Transportation*, *21*, 23–45.

Verreault, H., & Morency, C. (2011). Transcending the Typical Weekday with Large-Scale Single-Day Survey Samples. *Transportation Research Record*, *2230*(1), 38–47. https://doi.org/10.3141/2230-05

Veterník, M., & Gogola, M. (2017). Examining of correlation between demographic development of population and their travel behaviour. *Procedia Engineering*, *192*, 929–934.

Wilson, C., Harvey, A., & Thompson, J. (1999). ClustalG: Software for analysis of activities and sequential events. *IATUR Conference Proceedings*.

Xianyu, J., Rasouli, S., & Timmermans, H. (2017). Analysis of variability in multi-day GPS imputed activity-travel diaries using multi-dimensional sequence alignment and panel effects regression models. *Transportation*, *44*(3), 533–553. https://doi.org/10.1007/s11116-015-9666-2