Transportation Research Part C 126 (2021) 103007

Contents lists available at ScienceDirect

Transportation Research Part C

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

Visualizing, clustering, and characterizing activity-trip sequences via weighted sequence alignment and functional data analysis

Ying Song [a,*], Siyang Ren [b], Julian Wolfson [c], Yaxuan Zhang [a], Roland Brown [c], Yingling Fan [d]

[a] Department of Geography, Environment and Society, University of Minnesota, USA
[b] Statistician I, Dana-Farber Cancer Institute, USA
[c] Division of Biostatistics, School of Public Health, University of Minnesota, USA
[d] Humphrey School of Public Affairs, University of Minnesota, USA

ARTICLE INFO

Keywords:
Sequence alignment
Functional data analysis
Pattern recognition
Activity-travel behaviors
Smartphone

ABSTRACT

Smartphone-based activity-travel surveys have enabled the collection of continuous, multi-day data on individuals' trips and activities with high spatial and temporal resolution. However, the multi-dimensional nature of these data makes it challenging to compare activity-travel patterns and identify clusters of individuals with similar patterns and use them to study behaviors and forecast travel demands. To address this challenge, we adopt a discrete, step-based view on time and transform the episodic-based diary into a sequence of states observed at a sample interval. The resulting sequences can visually characterize variations in activity-travel patterns across days of a week and among different individuals. Motivated by techniques in genomics and data science, we apply sequence alignment methods to measure the pairwise similarity between these activity-trip sequences. To address its practical implementation in transportation studies, we define and compare four weighting schemas: (1) the unit-cost distance, which assigns equal weights to all substitution operations between states; (2) the fixed-flexible weighted distance based on the time geography framework, where costs differ for substitutions involving fixed and flexible activities; (3) the trip-activity weighted distance considering travel as a derived demand, where costs differ for substitutions between trips and activities; and (4) transition-based weighted distance, where costs are based on the global or time-varying activity and trip transition rates estimated from the data. Then, we calculate the pairwise distances between individuals' sequences and use them as inputs to a hierarchical clustering algorithm to group individuals with similar sequences. We visualize the state distributions of the identified clusters to infer and compare behavior patterns, and use functional data regression methods to estimate the time-dependent probabilities of engaging in various activities and trips. To demonstrate our methods, we analyze a smartphone-based survey dataset collected in the Minneapolis-St. Paul metropolitan area. We also conduct sensitivity analysis on the selection of cost metrics and sample intervals to ensure the robustness of our methods and discuss their implications in practice. By identifying population subgroups with distinct daily activity-travel patterns and explaining how these patterns vary over one day and depend on user profiles, our weighted

* Corresponding author at: Department of Geography, Environment and Society, University of Minnesota – Twin Cities, 414 Social Science, 267-19th Avenue South, Minneapolis, MN 55455, USA.

E-mail addresses: yingsong@umn.edu (Y. Song), sren@ds.dfci.harvard.edu (S. Ren), julianw@umn.edu (J. Wolfson), zhan6322@umn.edu (Y. Zhang), brow4288@umn.edu (R. Brown), yingling@umn.edu (Y. Fan).

https://doi.org/10.1016/j.trc.2021.103007
Received 9 June 2020; Received in revised form 29 January 2021; Accepted 30 January 2021
Available online 19 March 2021

sequence alignment approach provides an intuitive and flexible method for extracting and characterizing individuals' activity-travel behaviors for use in transportation planning.

1. Introduction

Understanding human mobility and activity patterns is essential in transportation and land use planning, public health, and many other fields. Recent advances in location-aware technologies such as Global Positioning System (GPS) allow us to track individuals' movements and activity participation. Many studies have used such tracking data to understand how individuals allocate their time and resources among activities and trips (see e.g. Kung et al., 2014, Simini et al., 2012). For instance, Song et al. (2010) and Simini et al. (2012) applied and refined stochastic models to investigate the durations of activities using cellphone detailed records (CDRs). Becker et al. (2013) also used CDRs and characterized mobility patterns at both the individual and community level. Das and Winter (2016) and Meng et al. (2017) integrated GPS data with geographic contexts to infer travel purposes as well as detect activity locations and times. While these studies provide valuable insights into individuals' mobility patterns, they often focus on the locations, durations, and purposes of a particular type of activities and trips, but not the sequence and interconnections among activities and trips.

Activities and trips are parts of an individual's entire schedule and they should not be treated as episodes independent from each other (Hägerstrand, 1970, McNally and Rindt, 2007, Miller, 2017, Miller and Roorda, 2003). To address this, recent projects have integrated GPS tracking devices with the traditional travel diary to get continuous information about participants' daily activities and trips throughout the day (Murakami and Wagner, 1999, Ohmori et al., 2006, Prelipcean et al., 2018, Seo et al., 2019). Compared to the traditional telephone and paper survey, GPS-based surveys allow participants to enter information in near real-time using mobile phones (or, in older studies, Personal Digital Assistants (PDAs)) and therefore can potentially reduce recall bias. Studies have collected field data and proved that GPS-based surveys can capture short-duration activities and trips, record more accurate times and location, and collect timely information such as users' physical activity, trip satisfaction, and emotion levels (Berger and Platzer, 2015, Nahmias-Biran et al., 2018). Hence, it is possible to enrich activity-based models in transport studies using GPS-based survey data (Zhao et al., 2015). However, the multi-dimensional nature of such data makes it quite challenging to identify behavior patterns that are distinct and representative. In activity-based approaches to travel analysis, these patterns would be the basis for simulating daily schedules and forecasting travel and access demands (Bhat and Koppelman, 2006, Ettema and Timmermans, 1997).

Given the travel survey data, one way to examine activities and trips within the context of the entire daily schedule is to use a *sequential state representation*, created by dividing the day into distinct, temporally ordered segments of equal duration (e.g., 5 min) and labeling each segment with the activity or trip that is occurring during that time period. With this sequential state representation, similar sequences indicate similar activity-travel scheduling. To quantitatively measure the dissimilarity between these sequences, we can apply the sequence alignment method (SAM), which has been widely applied to compare DNA sequences since the 1980s (Corpet, 1988) and nucleic acid and protein sequences more recently (Kumar et al., 2004, Li, 2018). The basis of SAM is a cost metric that explicitly defines the distance between two possible states at specific positions in the sequences. This metric can be either defined using domain knowledge or derived from the input data. Given this cost metric, the pairwise distance between two sequences is defined as the minimum cost to transform one sequence to the other via three operations: substitution (changing segment labels), insertion (adding a segment), and deletion (removing a segment).

This paper contributes to the emerging literature on using SAMs in behavior analysis and pattern recognition. The discovered patterns can provide inputs for the activity-based models to forecast travel demands and patterns. We address two gaps in sequential-based models. First, instead of applying the same unit cost in SAMs, the paper defines and compares different cost metrics, examines their impacts on the discovered patterns, and discusses practical implications of these metrics in behavior analysis and pattern recognition. Second, to describe the behavior patterns, the paper goes beyond visual summaries of the clusters by conducting functional data analysis to investigate the heterogeneous patterns within and across the clusters. To illustrate our methods, the paper uses activity survey data collected by a smartphone application in the Twin Cities (Minneapolis-St. Paul) metropolitan area as a study case.

Section 2 provides a brief review of the activity-based models in travel behavior analysis with a focus on the sequential approach and SAMs. Section 3 presents the study area and smartphone-based survey data. Section 4 describes the methods to construct, compare, and cluster activity-trip sequences and provides details about different cost matrices used to measure dissimilarity. Section 5 discusses and compares the results using different settings. And Section 6 summarizes the paper and identifies future research.

2. Literature review

The activity-based approach to travel analysis recognizes the interdependency between activities and trips in our daily schedules and views travel as a derived demand that results from the needs to obtain resources and participate in activities distributed in space and times (see Axhausen and Gärling, 1992, Bhat and Koppelman, 2006, Ettema and Timmermans, 1997). Compared to the trip-based approach that uses individual trips between zones as the analysis units, the activity-based approach also considers tours in daily schedules (e.g. home-work-shopping-home). Therefore, it can capture decisions on the sequencing, time allocation, and spatial locations or routes of activities and trips (e.g. the timing, location, and mode of a shopping trip is determined by the working hours, the work and home locations, and also the mode of the previous home-to-work trip). In essence, the activity-based approach emphasizes different needs of individuals or households and examines how they plan their daily life to fulfill such needs under specific sets of constraints. Time geography provides a conceptual framework to model the needs, constraints, and consequent behavior patterns of

individuals (Hägerstrand, 1970). The framework categorizes constraints into three major types related to the needs: (1) capability constraints that reflect the biological needs such as sleeping and eating and the mobility levels such as the ability to drive and access to public transit; (2) coupling constraints that address the needs to be present at given locations during certain times to interact with other individuals or materials (e.g., fixed work and school hours and locations; a scheduled doctor appointment); and (3) authority constraints that delineate the locations and times that are accessible to an individual such as the opening hours and locations of grocery stores, libraries, and other services. The observed schedule of an individual under these constraints is represented as a space–time path in space across time that captures the sequence of activities and trips, their spatial locations and routes, and their temporal profiles. The potential alternative schedules are modeled using a space–time prism (STP) that delineates all locations and times that are accessible to the individual. The conceptual framework of time geography became the basis to design travel surveys, identify activity-travel patterns, simulate scheduling, and forecast travel and access demands in the activity-based approach.

Since the mid-1980s, studies have developed analytical and computational methods to simulate daily schedules that address the interdependency between trips and activities (Gärling et al., 1994). One of the earliest attempts is the destination choice model developed by Kitamura and Kermanshah (1984), which considers the relative locations to the home location from candidate destinations of non-home-based trips. Some models directly adopt the framework of time geography: they identify objective constraints, simulate all feasible activity schedules, and compare their utilities (Jones et al., 1983, Recker et al., 1986). Other models start to differentiate long-term choices (e.g. home location, work location, and auto ownership) with short-term, daily scheduling of activities and trips. For instance, Gärling et al. (1994) represent the long-term calendar using the priorities and durations of all activities and then use it as the basis to retrieve a subset of essential activities for scheduling and determine destination locations and travel times for these activities. Instead of selecting all activities at once, Ettema et al. (1993) emulate the real-world scheduling process by computing the utility to include, delete, substitute activities, and getting the optimal schedule step by step. Recent studies also develop prototypes to describe scheduling patterns beyond activities and, for example, account for primary activities, primary tours, and secondary tours (Bowman and Ben-Akiva, 2001) and auto, transit, and various travel modes (Kitamura et al., 2000). Lately, studies have developed more comprehensive activity-based (micro) simulation models that have finer spatial and temporal resolutions, capture variations across different times of a day, address interactions among household members, consider the built environment, and integrate with land-use simulation systems (e.g. Bhat et al., 2013, Bradley et al., 2010, Pendyala et al., 2012, Saarloos et al., 2009, Ziemke et al., 2016).

Underlying these activity-based simulation models are explicit representations of each person (or household), their interactions with other people and the environment, and their movements and behaviors (Goulias et al., 2013). These provide the foundation to create individual agents and set their properties and behavior rules in the simulation models. The household activity diaries along with time-use diaries, health surveys, and census data have been used to detect behavior patterns that are distinct and representative and generate synthetic population for demand forecasting (e.g. Müller and Axhausen, 2011, Saarloos et al., 2009). Some of the early studies group individuals with similar distributions of activity type, start time, duration, frequency, and a set of explanatory variables such as age, gender, income, and occupation (Auld and Mohammadian, 2012, Miller and Roorda, 2003). Recent studies apply data mining and statistical learning methods to identify behavior patterns; they avoid the arbitrary selections of explanatory variables and can address the interdependencies between activities (Hafezi et al., 2019, Saadi et al., 2016, Su et al., 2020).

One approach to view activities and trips as part of daily schedules is to transform them into a sequence of *states* and apply a sequential alignment method (SAM) to compare the sequences. The SAM was first applied in social science in the late 1990s (Wilson, 1998). Since then, it has been used to compare individuals' life trajectories (Abbott and Tsay, 2000, Bras et al., 2010, Studer and Ritschard, 2016, Widmer and Ritschard, 2009) and activity-travel dairies (Joh et al., 2001, 2010, Saneinejad and Roorda, 2009, Shoval et al., 2015, Wilson, 2006, Xianyu et al., 2017, Zhai et al., 2019). These studies represented life events and daily activities as sequences to record their order of occurrence. However, these event-based sequences often do not store the time of events, and attributes such as the geographic locations and travel modes are missing too. To address this gap, Joh et al. (2001, 2002) proposed a position-sensitive SAM that modifies the cost metric to consider the recurrence of activities across time and the interdependencies among their attributes. Joh et al. (2010, 2011) also developed a heuristic algorithm to make this SAM scalable for large datasets. Wilson (2008) focused on the spatial aspect and defined the cost metric using both the Euclidean distance between two activity locations and their activity types. Zhai et al. (2019) continued to modify the metric by considering the time difference besides the geometric distance between activities. They also modified the substitution costs between trip purposes and between trip modes based on the concept of *ontology* (see e.g. Fonseca et al., 2002, Cho et al., 2013). Instead of defining the cost metric, Liu et al. (2015) and Saadi et al. (2016) derived the position-specific cost metric from the input data using the profile Hidden Markov Models (pHMMs). Hence, the metric can capture the probability that an activity type occurs at each position of the alignment as well as the probabilities of insertion and deletion of activities. Despite the great efforts to capture the multi-dimensional nature of diary data, the flexibility of activity durations has not been fully addressed. Zhai et al. (2019) started to consider the differences in the timing of occurrence and use the time difference to modify the cost metric. However, it would be difficult to determine the weights for the spatial distance, time difference, and activity type because they are based on different metric systems. Besides, the insertion and deletion operations during the alignment may result in cases where the sum of activity durations may not add up to 24 h.

One possible solution is to segment one day (or a week) into equal-length intervals and adopt a step-based view to study individuals' activity-travel behaviors (see e.g. Kwan et al., 2014, Shoval and Isaacson, 2007). Then, each state in the sequence represents what a person is doing at a given time within a day and additional information of such status (e.g. at 12:16 eating lunch in a pizza restaurant near the office; at 12: 48 walking along a road from pizza restaurant to office). Compared to the duration-based approach discussed earlier, the step-based approach can generate sequences with equal length, that is, individuals' sequences for one day always have the same number of states. These sequences can be used as inputs for some advanced SAM algorithms that can only take equal-

length sequences as the input (e.g. Lesnard, 2009). Besides, the step-based approach allows us to consider the duration and timing of activities at the same time. This is because if and only if two episodes start and end at the same time, they are aligned in the sequences. Moreover, the step-based approach labels each time interval with an activity or a trip type. Hence, we can append the labels to include

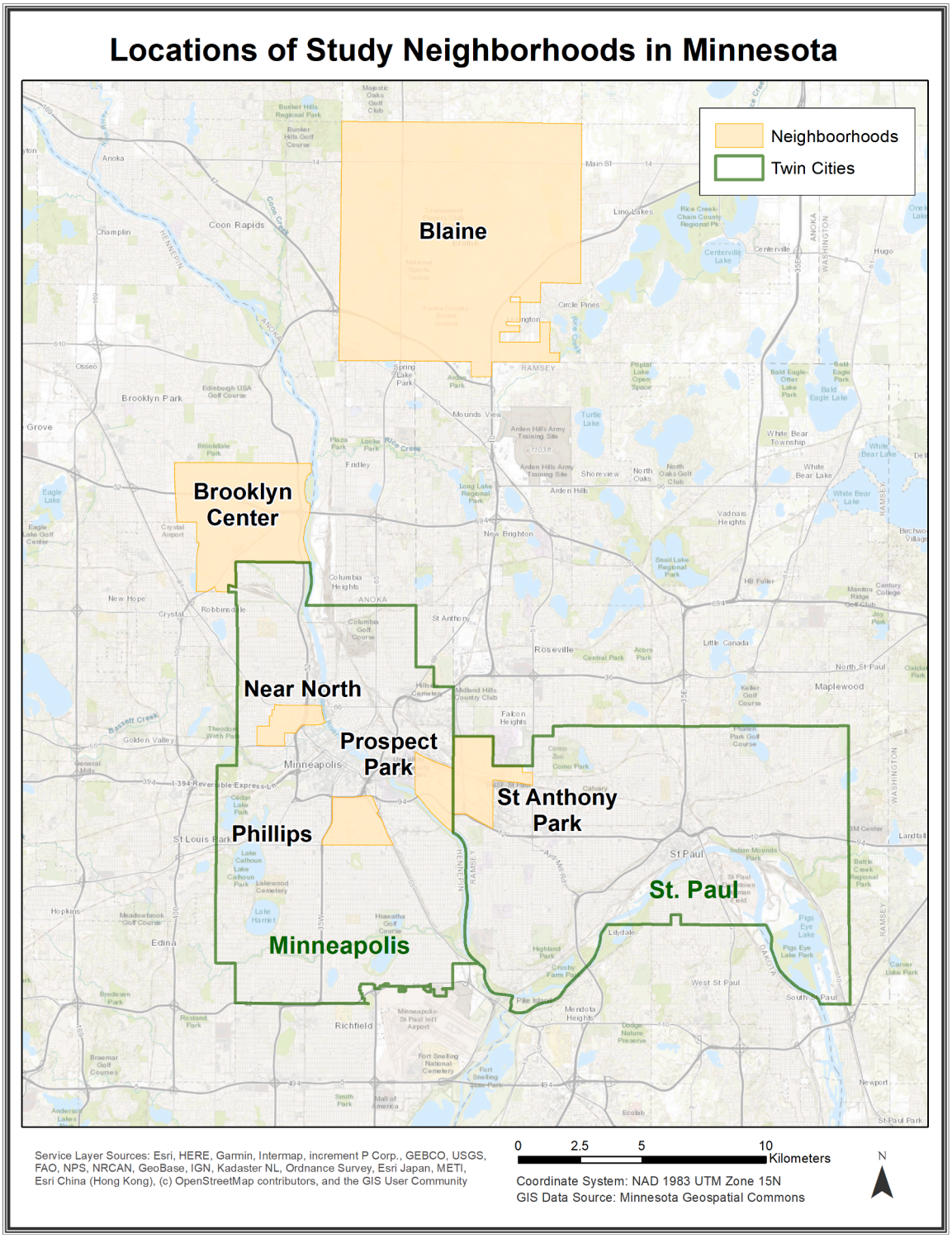


Fig. 1. Six survey neighborhoods in Twin Cities metro area.

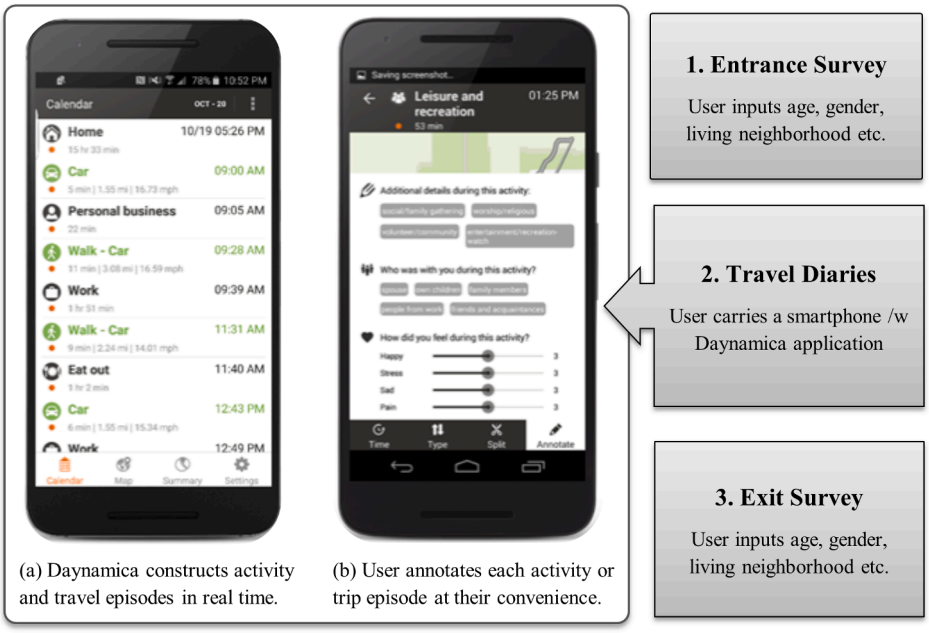
additional attributes of the activity or trip (see e.g. the spatial regions as well as the activity type in Kwan et al., 2014, Shoval and Isaacson, 2007).

The paper contributes to the emerging literature on using step-based SAMs to represent activities and trips and to identify groups of individuals with similar activity-travel behaviors. The primary focus of this paper is to define and compare different cost metrics and dissimilarity measures and examine their impacts on the alignment and clustering results. Existing studies have commonly assigned a unit cost for all operations. More recently, studies have shown that *ontology* weighted costs and dynamically derived costs may better reflect the likelihood to add, delete, or substitute an activity or trip than using the same unit cost for these operations (e.g. Zhai et al., 2019, Liu et al., 2015). However, the weighted and dynamic costs have been applied to the duration-based SAMs in the activity-travel behavior analysis and have not been studied in the step-based SAMs yet. Therefore, this paper defines different combinations of cost metrics and distance measures and relates them to their implications in the activity-based approach. The first set of measures adopt a *unit cost* for operations on all types of activities and trips and have been widely used in existing studies, which will serve as the baseline for result comparison in this paper. The second set of measures assign different costs for fixed and flexibility activities in the cost metrics. These cost metrics are based on the time geography framework and measure the *fixed-flexible weighted distance* between sequences. The third set of measures assigned different costs for activities and trips. These cost metrics consider travel as a demand derived from the need to participate in daily activities and measure the *trip-activity weighted distance* between sequences. Finally, the fourth set of measures are *transition-based distances* which use the likelihood that a trip or activity type is replaced by another type, and this paper considers both a global transition rate and dynamic transition rates. The unit cost, fixed-flexible weighted, and trip-activity weighted metrics are top-down and defined using domain knowledge, while the transition-based metrics are bottom-up and derived from the input data. Besides the dissimilarity measures, the paper also applies quantitative indicators of partition quality to determine the optimal number of clusters. Further, to characterize the discovered patterns of each cluster, the paper visualizes the state distributions and conducts functional data analysis to investigate the patterns within and across the clusters. Lastly, the paper examines how the choices of sample intervals and cost metrics may affect the analysis results and discuss the selection of them with examples. The following sections discuss the data, methods, and results in detail.

3. Data

The data used in this paper comes from an activity-travel survey conducted in the Twin Cities (Minneapolis-St. Paul) metropolitan area between October 17, 2016, and October 25, 2017. The survey successfully recruited 398 residents from six neighborhoods in the region (Fig. 1). The neighborhoods are selected to represent urban, suburban, and exurban areas considering their spatial locations, transit and highway access, and population demographics. Recruitment was based on geographic cluster sampling: random blocks were selected within each neighborhood, and efforts were made to recruit as many households as possible from each block. Each participant carried a smartphone equipped with GPS and the smartphone application that collected data during trips and activities and the participants were asked to report their activities and trips for seven consecutive days.

The smartphone application, Daynamica (previously called SmarTrAC, Fan et al., 2015), detected activities and trips in real-time and constructed sequential activity and trip episodes (Fig. 2a). The participants annotated each detected activity and trip with



(a) Daynamica constructs activity and travel episodes in real time.

(b) User annotates each activity or trip episode at their convenience.

Fig. 2. Daynamica main interface and survey flow (Fan et al., 2019).

Table 1
Descriptive statistics for the 2017 Daynamica survey data in the Twin Cities.

(A) Travel Diary (25,622 episodes)

(A1) Episode by Activity/Trip Type (unit: minutes)

Type	Subtype	Count	Mean	Median	S.D.	Max
Activity	Home	3,728	432	241	416	1435
	Work or Education	2,140	206	142	207	1377
	Personal Business	1,755	78	19	171	1123
	Shopping	1,668	31	17	49	639
	Leisure	1,949	103	56	162	987
	Eating out	729	52	39	48	338
	Other non-home activities	871	100	19	210	1291
Trip	Auto	7729	16	11	27	941
	Auto Passenger	432	12	9	11	99
	Bus	483	18	13	22	319
	Rail	163	22	18	21	197
	Bike	528	16	10	19	146
	Walk	3057	9	5	14	336
	Other	121	13	6	28	248
	Wait	269	6	3	10	97

(A2) Episode by Day of Week

Day of Week	Person-Day Count	Activity Count Per Person-Day				Trip Count Per Person-Day			
		Mean	Median	S.D.	Max	Mean	Median	S.D.	Max
Monday	364	4.9	4	3.0	20	5.3	5	3.3	24
Tuesday	402	4.7	4	3.2	18	5.1	4	3.3	17
Wednesday	410	4.8	4	3.3	20	5.4	5	3.3	18
Thursday	430	4.9	4	3.3	22	5.1	4	3.4	24
Friday	419	4.9	4	3.1	24	5.2	5	3.1	21
Saturday	362	4.9	4	3.2	18	5.2	4	3.3	22
Sunday	319	4.5	4	2.8	14	4.8	4	2.8	13

(*continued on next page*)

Table 1 (*continued*)

(b) User Profile (368 participants)

(b) User Profile (368 participants)

(b.1) Trips and Activities

Variable	Description	Mean	Median	S.D.	Min	Max
Day count	counts of survey days for each person	7	7	2.2	1	13
Activity count	counts of activities on each person-day	4.8	4	3.1	1	24
Trip count	counts of trips on each person-day	5.2	5	3.2	1	24

(b.2) Demographics

Variable	Description		Value	Percentage
Sex	Participants' sex		Male	33.43%
			Female	66.03%
			Not answer	0.54%
Age	Participants' age range		Under 18	0.00%
			18–25	6.52%
			26–35	18.48%
			36–45	16.03%
			46–55	16.03%
			56–65	22.56%
			66–75	15.76%
			Above 75	4.62%
Employment	Choose one or more that can describe the current employment status of participants	Employed	Employed full-time	42.12%
			Employed part-time	30.16%
		Student	Student full-time	8.97%
			Student part-time	5.16%
		Other	Homemaker	12.50%
			Retired	20.11%
			Unemployed	9.78%
Driver	Participants have valid driver's license or not		Yes	88.59%
			No	10.87%
			Not answer	0.54%
Neighborhood	Neighborhood the respondent lives in		Philips	19.02%
			Near North	15.22%
			Prospect Park	17.93%
			St. Anthony Park	19.57%
			Blaine	14.13%
			Brooklyn Center	14.13%

information such as activity and trip subtypes, companionship, and emotional experiences. Participants could choose to do annotation during each activity and trip or later at their convenience. Fig. 2b shows an example annotation questionnaire. The participants were also asked to finish one-time entrance and exit surveys to collect additional information about their social-demographics and other related information such as their overall life satisfaction and (at exit) experiences using the Daynamica app. Although each participant was asked to participate in the survey for seven consecutive days, many participants provided less (or more) than seven days of trip/activity data. Among the 398 participants, 369 provided at least one day of valid data as well as valid entrance and exit surveys. 29 participants were excluded for a variety of reasons including device failure (e.g. GPS tracking data missing), low response rate on episode annotations, or insufficient auxiliary data (e.g. missing entrance or exit survey).

The final dataset contains 12,840 activity episodes and 12,782 trip episodes in total. Each episode contains the start and end time, user annotation time, activity and trip type, movement trajectory, emotion status (e.g. happy and sad), and other optional attributes such as companionship. Table 1 presents descriptive statistics on the recorded activity and trip episodes and participants' socio-demographic profiles. On average, home activities took around 7.2 h but did not exceed one day (24 h = 1,440 min). Next are the work and education activities which took around 3.4 h on average. There are extremely long activities with durations longer than 18 h (1,080 min), which were most frequently observed for in-home activities (home episodes) and were also observed in work and personal business activities (10 and 7 episodes respectively). There might be unreported activities or trips during this long period (e.g. walking between office buildings), but they are often considered as part of a larger task and the interpolation of potential unreported episodes is outside the scope of this paper. For trip episodes, the average durations are much smaller. This is intuitive but also resulted from the fact that the smartphone application can break down a trip into small trip legs if there is more than one travel mode used (e.g. Walk-Car in Fig. 2a). Considering the relatively small sample size, the paper only distinguishes wait from other trip types to avoid over-fragmented trip legs and leaves the investigation of different travel modes for future research. We will discuss this in the conclusion (Section 6). Also, since participants may have more or less than 7 consecutive days of travel diaries, the paper will use activities and trips in one day from 12:00 am to 11:59 pm to construct daily (a.k.a. a person-day) sequences of activities and trips. In sum, there are more episodes recorded in the middle of the week from Tuesday to Friday and relatively fewer episodes recorded on Monday, Saturday, and especially Sunday (see Table 1A). The paper will compare the discovered patterns across different days of the week to address day-to-day variations. The number of person-day dairies also vary among participants as expected (see Table 1B). The total number of survey days for each person ranges from 1 to 13 with an average of 7 days. And on average, each person has 4.8 activity episodes and 5.2 trip episodes in one day. The summary statistics of participants' profiles suggest that there are more females than males, most of the participants are over 35 years old and have a valid driver's license, and more than 90% of them are employed. For spatial distributions, the participants are evenly distributed across the six neighborhoods. The paper will relate these user profiles to the behavior patterns of their belonged cluster in the result section (Section 5). Considering the small sample size and the aims of this paper, we will not conduct multivariable regression analysis.

4. Methodology

This section describes the methods we develop in this paper to represent, compare, and describe individuals' activity-travel patterns using a time-oriented, state-based framework, and weighted SAMs. The four major steps in our approach are (1) converting individual diary to a sequence consisting of equal-time steps; (2) defining or deriving the cost metric to measure dissimilarities between sequences; (3) grouping sequences and determining the optimal number of clusters; and (4) investigating the intra-group and inter-group variations using functional data analysis.

4.1. Creating activity-trip sequences

The activity-trip diary of an individual record a list of activity and trip episodes within one day (or a longer period). Table 2 shows selected attributes of an example diary for a weekday used in constructing the sequence. Each record stores attributes of one activity or trip episode, including the identifier, the user identifier, the start and end times, and the activity or trip type. We encode the seven types as letters: home (H), work or education (W), personal business (P), shopping (S), leisure (L), eating out (E) and other non-home

Table 2
Example data showing activity and trip episodes in one day for a single individual.

Record ID	User ID	Start time	End time	Type	Abbr.
8755	1001	11/2/2016 22:03	11/3/2016 8:35	HOME	H
8756	1001	11/3/2016 8:35	11/3/2016 9:15	TRIP	T
8757	1001	11/3/2016 9:15	11/3/2016 12:09	WORK	W
8758	1001	11/3/2016 12:09	11/3/2016 12:55	EAT OUT	E
8759	1001	11/3/2016 12:55	11/3/2016 16:55	WORK	W
8760	1001	11/3/2016 16:55	11/3/2016 17:12	TRIP	T
8761	1001	11/3/2016 17:12	11/3/2016 19:24	LEISURE	L
8762	1001	11/3/2016 19:24	11/3/2016 19:35	TRIP	T
8763	1001	11/3/2016 19:35	11/4/2016 6:45	HOME	H

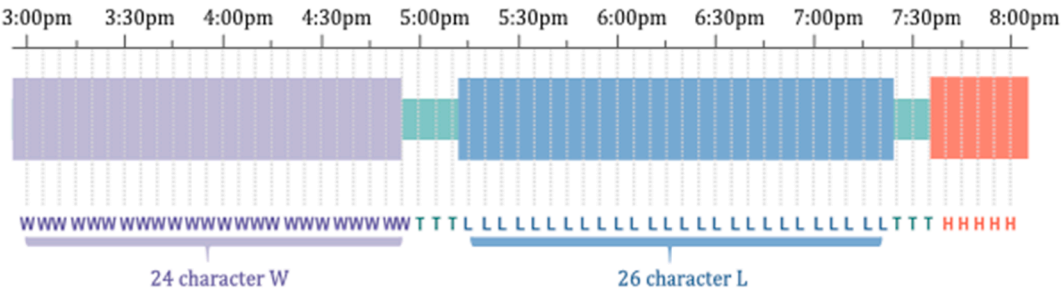


Fig. 3. An example sequence resulting from sampling activity-travel episodes in Table 1 between 3 pm and 8 pm at 5-minute intervals.

activities (O), and considered both trips (T), and waiting among these activities (N). Fig. 3 shows the resulted sequence for the records between 3 pm and 8 pm in Table 1 with 5 min sampling interval. Each character in the sequence represents the status of that individual at a specific time of a survey day. The paper will use 1, 3, 5, and 10 min in the result section (Section 5) to discuss the impacts of sample intervals and their implications. Note that we only use one letter for the activity and trip type, but the method is scalable to include additional attributes such as spatial region or emotion status. We will discuss such an extension in Section 6 as one future research direction.

4.2. Calculating distances between sequences

To carry out sequence alignments, we must first define a metric to describe the cost of operations on possible states in the sequences. The three basic operations are: (1) substituting one state with another state, (2) inserting one state into a sequence, and (3) deleting one state from a sequence. Insertions and deletions are often referred to as indels, and they can alternatively be defined as substitutions from and to the *Null* state, respectively. Therefore, the cost metric can be defined via a *substitution cost matrix*. For example, the states in Table 2 are {H, W, L, T} and if they are all possible states for all sequences, the cost metric can be represented as:

$$\begin{array}{c} \\ H \\ W \\ L \\ T \\ \phi \end{array} \begin{array}{c} \begin{array}{ccccc} H & W & L & T & \phi \end{array} \\ \begin{bmatrix} \gamma_{HH} & \gamma_{HW} & \gamma_{HL} & \gamma_{HT} & \gamma_{H\phi} \\ \gamma_{WH} & \gamma_{WW} & \gamma_{WL} & \gamma_{WT} & \gamma_{W\phi} \\ \gamma_{LH} & \gamma_{LW} & \gamma_{LL} & \gamma_{LT} & \gamma_{L\phi} \\ \gamma_{TH} & \gamma_{TW} & \gamma_{TL} & \gamma_{TT} & \gamma_{T\phi} \\ \gamma_{\phi H} & \gamma_{\phi W} & \gamma_{\phi L} & \gamma_{\phi T} & \gamma_{\phi\phi} \end{bmatrix} \end{array} \tag{1}$$

where ϕ represents the *Null* state and γ_{ij} corresponds to the cost of substituting state i with j. If the state i with j are the same, there will be no cost; otherwise, a positive cost will be assigned to reflect the difference between states (Studer and Ritschard, 2014). The paper applies four types of cost metrics: (1) the *unit-cost edit distances* with a unit cost for all activity and trip types; (2) the *fixed-flexible weighted distances* with higher costs for fixed activities; (3) the *trip-activity weighted distances* with costs differed by trips and activities; (4) the *transition-based distances* with costs derived by calculating transition rates between different activities and trips from input sequences.

Given the cost metric, we align two sequences and get the total distance (dissimilarity) between them. The most common family of distances is known as the *edit distances* in bioinformatics and computer science and as the *optimal matching distances* in social sciences (Abbott and Tsay, 2000). The total distance from a sequence $A = \{\alpha_i\}$ to another sequence $B = \{\beta_j\}$ is defined as the minimum total cost of transforming A to B via a series of basic operations on states α_i and β_j in these two sequences. A special type of edit distances is the *hamming distances* that compares sequences with equal length and do not allow indels (e.g. He, et al., 2004, Norouzi et al., 2012). The paper calculates both the hamming distance (HM) that does not allow indels and the optimal distance (OM) that allows indels. The HM compares the same position (time) in two sequences and counts the total number of cases when states are different for all positions (time). Therefore, if and only if two sequences are the same, the distance between them would be zero; and a small change in the order of activity episodes may result in a large distance. For instance, if one person did running in the morning and then went to work and another person did running after coming back from work in the afternoon, even the rest of these two persons' activities have the same order and durations, the HM between these two persons is very large because many timestamps would have different states due to the time offsets of these activities and trips. This suggests that the HM focuses on differences in *time allocations* of all activities and trips within 24 hours of a day. In comparison, the OM focuses primarily on the *sequential order* and are not as sensitive to the time allocations as HM. The OM allows indel operations. Therefore, in the previous example, the OM between those two persons would be much smaller because we only need to "move" the states for recreation (running) from morning to the afternoon to match their schedules. And the states for their other activities and trips would be the same and have no additional costs. The rest of this section discusses each *metric and distance combination* in detail.

Unit-Cost Edit Distance: The metric assigns a unit cost for operations between two different and no cost for operations between the same states (Studer and Ritschard, 2016):

$$\gamma_{\alpha\beta} = \begin{cases} 0 & \alpha = \beta \neq \phi \\ 1 & \alpha \neq \beta \\ 2 & \alpha = \phi \ or \ \beta = \phi \end{cases} \tag{2}$$

where α and β are two states in the state space and the *Null* state ϕ is introduced for insertions and deletions, and the cost $\gamma_{\alpha\varphi}$ and $\gamma_{\varphi\beta}$ often have a higher penalty, commonly set as 2 times as the substitution cost. The OM distance between two sequences is the minimum cost to convert one sequence into the other. And the HM distance does not allow indels by excluding the null state ϕ from the state space. So, the HM distance counts the total number of positions (times) that have different states.

The unit-cost edit distance is the most straightforward way to measure dissimilarity. For real-world behavior analysis, however, it imposes a somewhat unrealistic assumption that operations for all types of activities and trips have the same cost. Therefore, substituting a work activity for a shopping activity has the same cost as substituting a leisure activity for a shopping activity. But these two costs are commonly different in our daily scheduling practices as some tasks may have higher priorities and are less flexible than other tasks. To address the different priorities of tasks and demands, we include two weighted metrics below, namely, fixed-flexible weighted distances and trip-activity weighted distances.

Fixed-Flexible Weighted Distance: We first assign different operation costs according to whether the activities are fixed or flexible. We use the time geography framework, where home, work, and education activities are fixed $FIX = \{H, W\}$, and trips and other activity types are flexible $FLEX = FIX^C = \{L, T\}$ (see e.g. Miller, 1991, Schwanen et al., 2008). We set the substitution cost of *cross-category* and the indel costs as double that of *within-category*:

$$\gamma_{\alpha\beta=} \begin{cases} 0 & \alpha = \beta \neq \phi \\ 1 & \alpha \neq \beta \ and \ \alpha, \beta \in FIX \\ 1 & \alpha \neq \beta \ and \ \alpha, \beta \in FLEX \\ 2 & \alpha \neq \beta \ and \ \alpha \in FIX \ and \ \beta \in FLEX \\ 2 & \alpha \neq \beta \ and \ \beta \in FIX \ and \ \alpha \in FLEX \\ 2 & \alpha = \phi \ or \ \beta = \phi \end{cases} \tag{3}$$

Based on the unit-cost metric, the paper calculates both the OM distance between sequences, and the HM distance excludes the null state ϕ in the state space to avoid indels. In general, the fixed-flexible weighted distance focuses on how people schedule fixed activities besides considering the overall dissimilarities as in the unweighted (unit-cost) measures. So, the time allocation of fixed activities would have a greater impact on the measured distance between two sequences.

Note that these costs are not definite. The classification of fixed and flexible activities and their substitution costs can be modified according to the specific context of a problem or the input data. For instance, when we include the location of the home, work, and education to define a state at each position in future research, we could differentiate work at home from work in an office or a working trip (e.g. delivery pizza) and then set a higher substitution cost between fixed activities (e.g. 4 units) if the work locations are different from the home location for a person. For this paper, the primary focus is to initialize such a framework by examining the choice of different cost metric types.

Trip-Activity Weighted Distance: We then assign different operation costs to differentiate all trips from all activities. This recognizes that travel is a derived demand and activities are the demands and tasks in our daily schedule. So, the fixed set contains all activities $FIX = \{H, W, L\}$, and the flexible set contains all trips, $FLEX = FIX^C = \{T\}$. We use the same cost metric as in the fixed-flexible weighted measure to set the substitution costs (see Equation 3). And we calculate both the OM distance and HM distance (excluding null state ϕ). The trip-activity weighted distance focuses on how people schedule trips to connect activities, so it is expected to be more sensitive to the time allocation of trips.

Transition-Based Distance: Both the unweighted (unit-cost) metric and the two weighted metrics assign substitution costs using some prior knowledge such as the frameworks of activity-based approach and time geography. This cannot capture that different geographic regions and/or time of a day may have quite distinct preferences during activity and trip scheduling. For instance, for study areas with less connected road networks and/or time of the day with fewer transit services, trip scheduling becomes more important and should be assigned a higher weight. To address this potential variation, we adopt two transition-based cost metrics that calculate the transition costs using the input sequences. First, we calculate the global transition rates between each pair of activity/trip types and use them to define the substitution costs (Studer and Ritschard, 2016). The transition rate between two states, α and β, is the probability of substituting α with β (or β with α) between two successive positions (time) for all sequences:

$$p(\beta|\alpha) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(\alpha, \beta)}{\sum_{t=1}^{L-1} n_t(\alpha)} \tag{4}$$

where L is the total number of states in a sequence (or the sequence length), $n_t(\alpha)$ is the number of sequences with state α at position t, and $n_{t,t+1}(\alpha, \beta)$ is the number of sequences with state α at position t and state β at position $t + 1$. The symmetrical substitution cost is estimated as:

$$\gamma_{\alpha\beta} = 2 - p(\beta|\alpha) - p(\alpha|\beta) \tag{5}$$

The transition rate, in its simplest form, reflects the probabilities that α and β would occur next to each other along the sequence. A higher probability means it is more likely to occur and hence would infer that it is less obstacle to make such changes in schedule in

general.

However, the global transition rate assumes that substitution costs are constant over the positions of a sequence, i.e., that they do not depend on the time when they occur. This assumption may be violated in aligning activity sequences over the study period. For example, substituting a home activity for a car trip at a time segment from 2:00 am to 2:05 am should perhaps imply a higher cost between sequences than if that substitution occurs at a time segment from 8:30 am to 8:35 am because it is usually less likely to be observed such transition in our daily life (as recorded in the input sequences). One solution to address this temporal variation is the *Dynamic Hamming Distance (DHD)* proposed by Lesnard (2009). The time-dependent transition rate between two states, α and β, at position t is derived only using states at positions t and $t+1$ in sequences:

$$p_t(\beta|\alpha) = \frac{n_{t,t+1}(\alpha,\beta)}{n_t(\alpha)}, t = 1, 2, \cdots, L-1 \tag{6}$$

And the substitution costs between α and β at position t is the transition rates cross-sectionally observed between $t-1$ and t and between t and $t+1$, that is:

$$\gamma^t_{\alpha\beta} = 4 - p_{t-1}(\beta|\alpha) - p_{t-1}(\alpha|\beta) - p_t(\beta|\alpha) - p_t(\alpha|\beta), t = 2, \cdots, L \tag{7}$$

For the cost metric based on the global transition rate, we can apply both OM and HM distances. But for the dynamic, time-dependent distance metric, only HM distance can be applied since the cost is location-specific and require that two sequences have the same length. Note that the global and time-dependent cost metrics are symmetric matrices, $\gamma_{\alpha\beta} = \gamma_{\beta\alpha}$ and $\gamma^t_{\alpha\beta} = \gamma^t_{\beta\alpha}, \forall t$. This allows arbitrarily substituting α with β in the first sequence or β with α in the second sequence during sequence alignment and removes the impacts of the arbitrary ordering of the input sequences (determining which sequence is the first and which is the second). This means that the distance measured between two sequences always reflects the optimal solution. In the previous example of two persons who scheduled their exercise time differently (in the morning or the afternoon), the optimal solutions would be two people both substituting their recreation activities with home activities due to the high probability of being at home in the early morning and later afternoon. Therefore, to some degree, the transition-based distances are influenced by the global pattern or the local patterns at individual positions (time).

4.3. Sequence clustering

After applying the SAMs above to N sequences, we obtain an N by N symmetric distance matrix whose entries quantify the pairwise similarity between sequences. Based on this matrix, we then apply hierarchical clustering (Murtagh, 1983) to group similar sequences into clusters. We use an agglomerative hierarchical clustering method known as Ward's method (Ward, 1963), which is an example *bottom-up* approach whose objective is to minimize the variations within clusters.

To determine the optimal number of clusters for Ward's method, we calculate several indicators of partition quality (Studer, 2013).

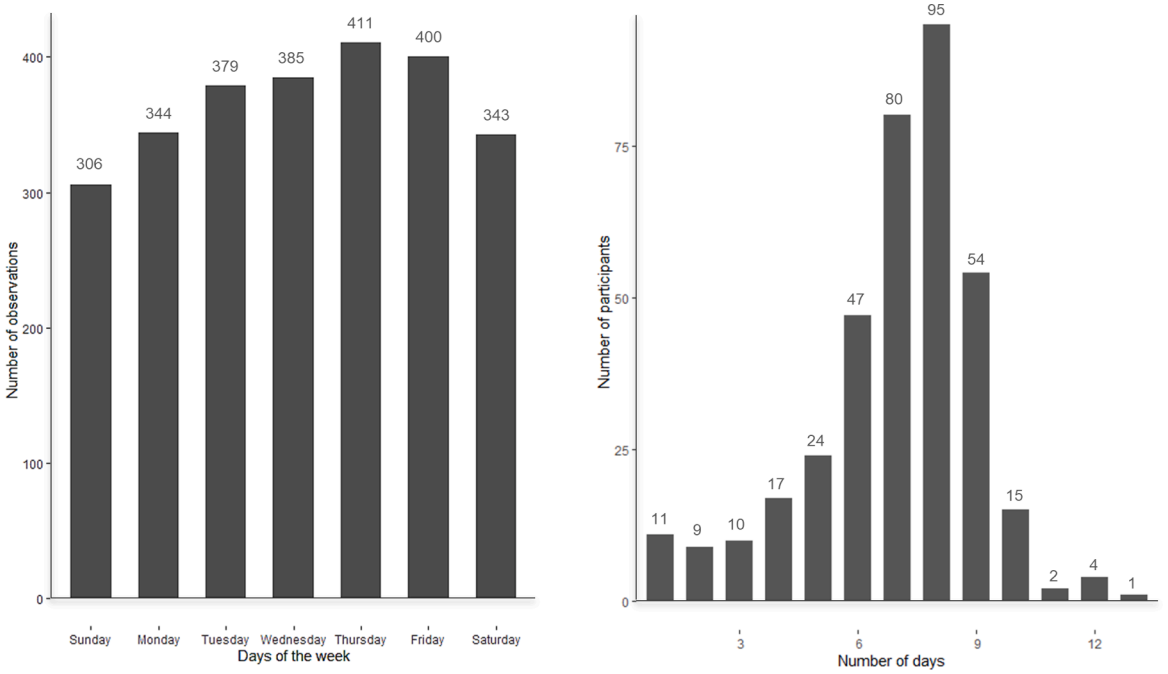


Fig. 4. Number of person-day sequences using 1 min as sample interval.

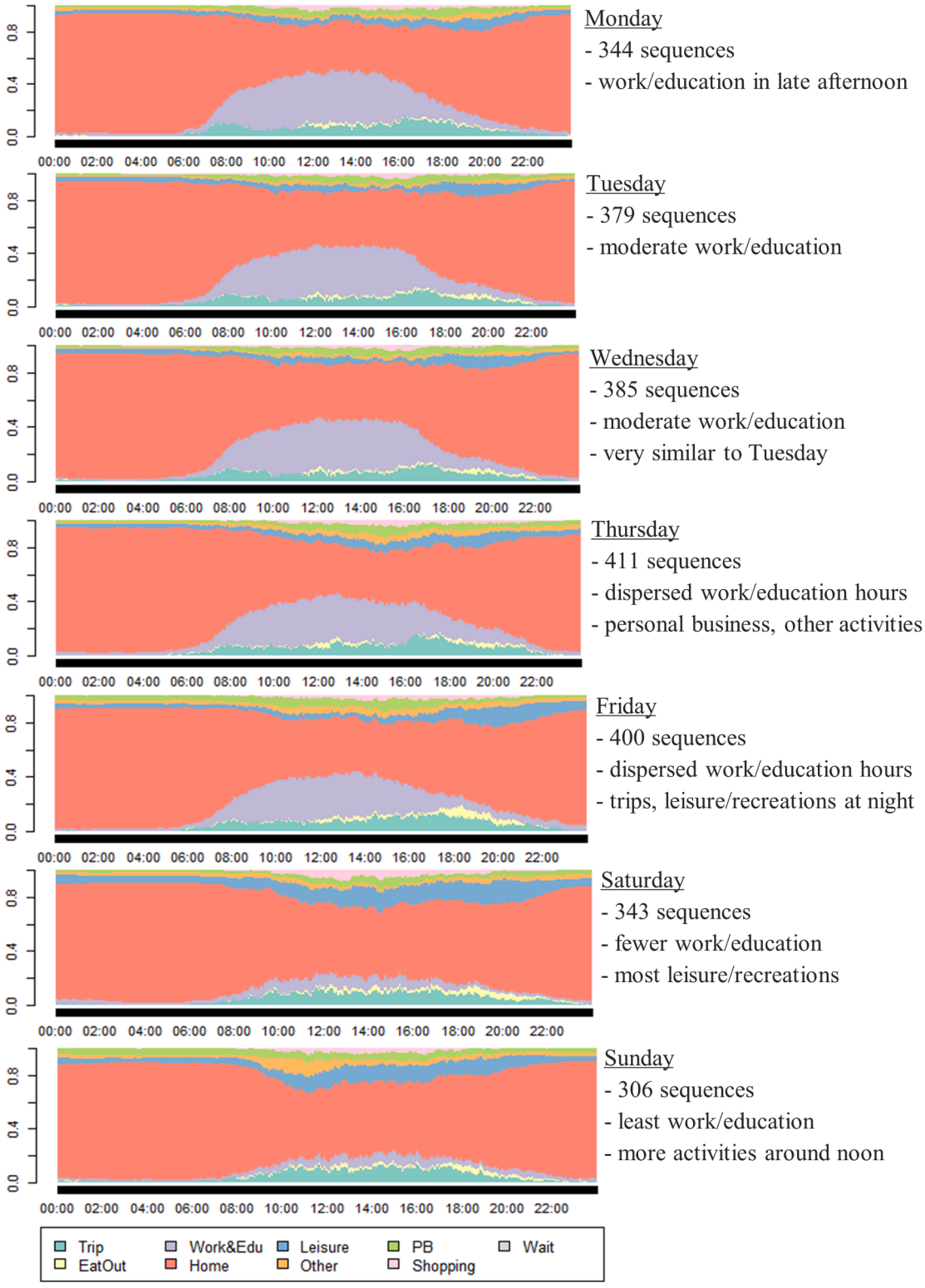


Fig. 5. State distribution for each day of the week with major patterns.

The first two indicators are Point Biserial Correlation (PBC) (Milligan and Cooper, 1985) and Hubert's Gamma (HG) (Hubert and Arabie, 1985) which consider a partition valid if the intra-cluster distances are smaller than the inter-cluster distances. The third indicator is weighted Average Silhouette Width (ASWs) (Kaufman and Rousseeuw, 1990), which examines the coherence of assignments. A high coherence indicates that individual observations in each cluster are homogeneous concerning their large distances to other clusters. The fourth is the Calinski-Harabasz index (CH) (Calinski and Harabasz, 1974) based on the F-statistics of the analysis of variance (ANOVA) that also considers variations within each cluster. The fifth is Hubert's C coefficient (HC) (Hubert and Levin, 1976). It uses the maximum distance between any two observations in the entire dataset to indicate the maximum inter-cluster distance that can be possibly achieved the best partition and compare it to the current partition. Therefore, unlike other indicators, a smaller HC value indicates a better partition. For easy comparison, we apply $1-HC$, namely, RHC in this paper. In sum, all these partition quality measures examine the heterogeneity across clusters and homogeneity within clusters (Studer, 2013). Unless there is a specific outcome for optimization given by a practical application, it is reasonable to determine the optimal number of clusters based on a *majority vote* across all different quality measures.

4.4. Functional data analysis

Applying hierarchical clustering to the distance metric of activity sequences allows us to identify clusters of similar sequences. Graphical methods are often used to visualize and compare these identified clusters and their inferred patterns. A typical method is density plot, which shows the percentage of sequences with state α at position t given all sequences' states at position t. Even though these graphical methods provide informative visualizations, they discovered patterns and comparisons are descriptive and somewhat informal. Moreover, the state-based SAMs are based on a selection of time intervals that best suits the application needs or available data (e.g. every 5 min at 12:00am, 12:05am, 12:10am, and so on). Therefore, these plots can only show values at those discrete time stamps and may not capture possible variables in between. Therefore, this paper uses a more principled statistical approach to comparing key features of clusters based on functional data analysis (FDA) (Ramsay, 2006).

In the FDA framework, the goal is to model a functional response $Y(t)$, where the argument t is viewed as continuous. (Ullah and Finch, 2013). In our context, the activity-travel patterns in a day are $Y(t)$ where t can be anytime within that day, and $Y(t_i)$ is the patterns at time t_i. The patterns can be represented as a set of functional responses, $\{Y_1(t), Y_2(t), \cdots, Y_K(t)\}$, where $Y_k(t)$ is one of the indicators to describe activity-travel patterns, such as being in a specific activity/trip state α at time t across a day. The data used to model $Y_k(t)$ are observations taken at discreate time:

$$\widehat{Y}_k^i(t) = \left\{ \widehat{y}_k^i(t) = \left(y_k^i(t_1), y_k^i(t_2), \cdots, y_k^i(t_L)\right) \right\}, i = 1, 2, \cdots, N \tag{8}$$

where $y_k^i(t_j)$ is the observed value for pattern indicator k of the ith individual at time t_j, L is the length of state-based sequences, and K is the number of pattern indicators. For instance, if we are interested in the probability of an individual being at home over a day, $Y_H(t)$, we create a binary sequence based on whether the state at time t_j for individual i, α_j^i, is a home activity or not:

$$y_H^i(t_j) = \begin{cases} 1 & \alpha_j^i = H \\ 0 & \alpha_j^i \neq H \end{cases} \tag{9}$$

We are also interested in knowing how $Y_k(t)$ may differ, depending on the cluster membership. If $\vartheta$is the cluster membership indicator, we can apply a technique known as function-on-scalar regression (FOSR) (Goldsmith et al., 2015). The FOSR can assess whether the probability curves $p_k^i(t;\vartheta) = E(y_k^i(t_j)|\vartheta)$ are significantly differing across clusters ϑ at any time t. For simplicity, let ϑ be a

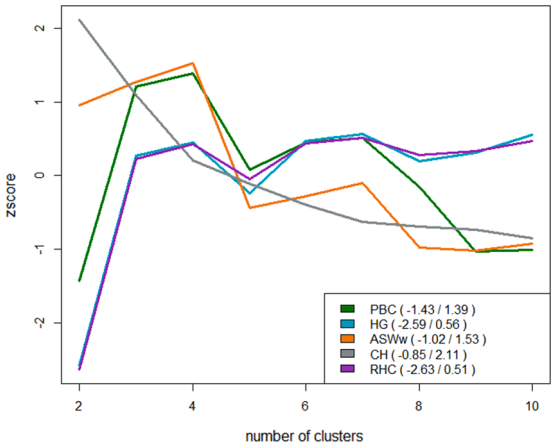


Fig. 6. Measures of partition quality for Monday (unit-cost HAM, 1 min).

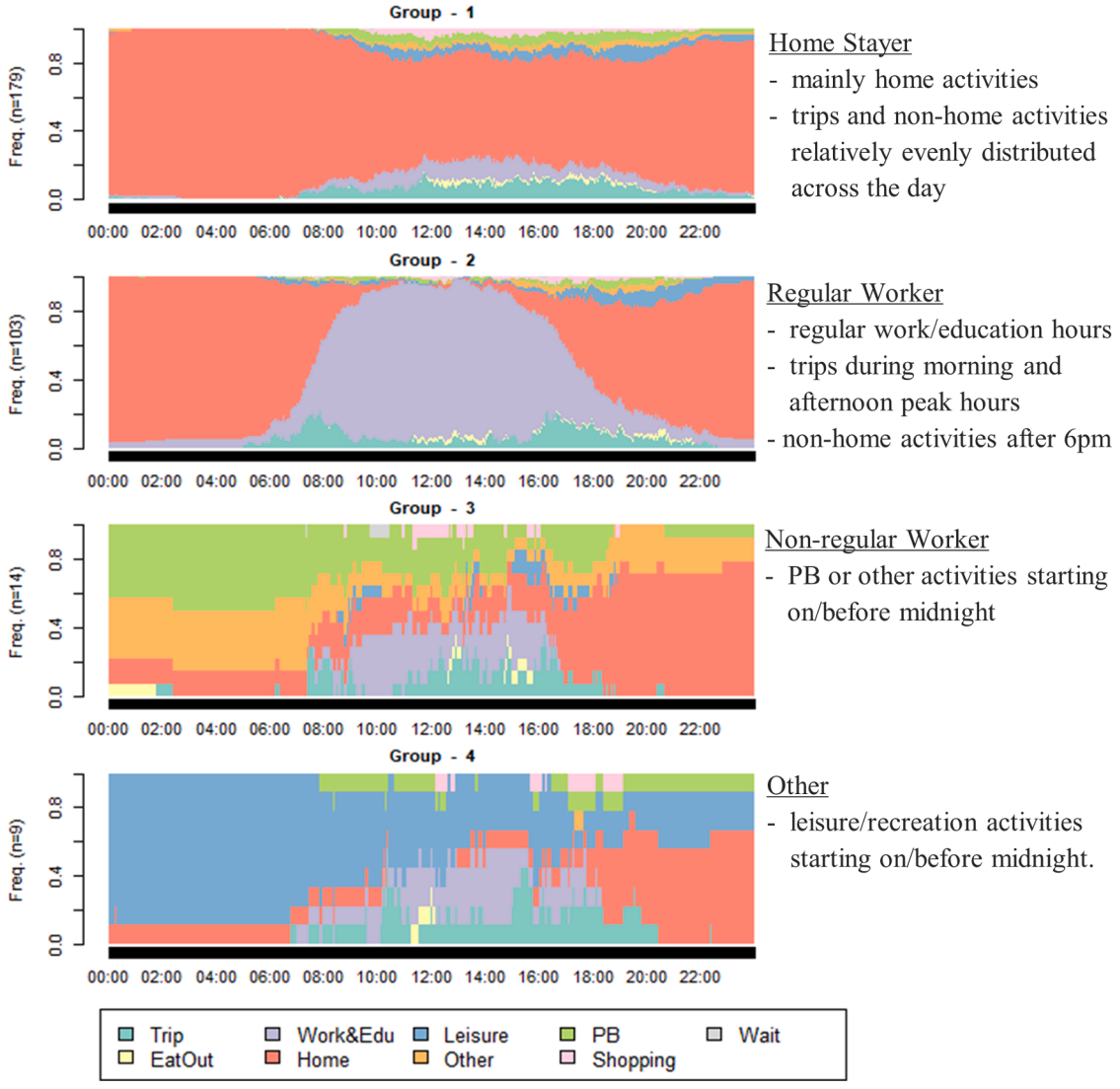


Fig. 7. State distributions of daily sequences on Monday (unit-cost Hamming).

binary variable indicating membership in one of two clusters; then, the FOSR model in this problem takes the form:

$$p_k^i(t) = \mu_k(t) + \vartheta_k^i f\nu_k(t) \tag{8}$$

where the ϑ_k^i indicating the cluster membership of individual i for the pattern indicator k.

Note that we use a functional linear regression model here even though the outcome is binary. This is partly because the techniques for generalizing FOSR (e.g., function-on-scalar logistic regression) have not been well-established in the literature or implemented in existing software packages yet. The targets of estimation and inference are the curves $\mu(t)$ and $\nu(t)$ that can show variations across time across clusters, which are represented using a set of smoothing functions (e.g. B-spline, exponential) or non-parametric method (e.g. Kernel smoothing). The coefficients of these smoothing functions are estimated under a smoothness constraint which penalizes large fluctuations in $\mu(t)$ and $\nu(t)$ locally within a small time window. This paper uses B-spline as the basis smoothing function and uses the optimal number of functions to generate trend lines.

5. Result

We use the programming language R and three supporting packages to implement our methods. First, we use the package *TraMineR* (Gabadinho et al., 2011) to construct activity-trip sequences, derive pair-wise distances, generate clustering results, and visualize clusters. Second, we convert individuals' activity-trip sequences into functional data using the package *fda* (Febrero-Bande and Oviedo de la Fuente, 2011). Third, we use these functional data and the cluster membership as input for the function-on-scale regression

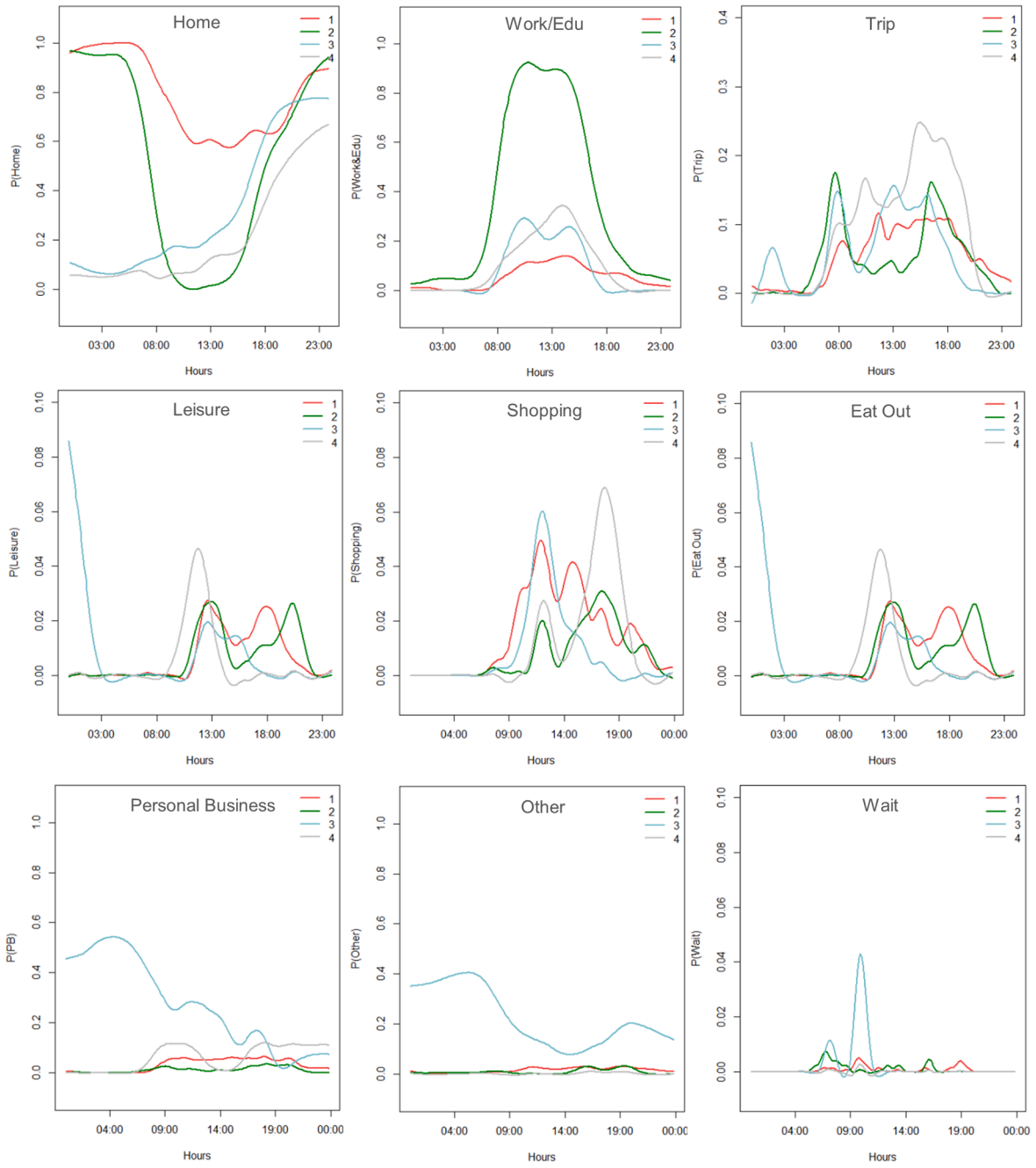


Fig. 8. State probabilities over time estimated through FDA (Monday, HAM, 1 min).

models available in the package *refund* (Goldsmith et al., 2015). Access to our GitHub repository with scripts and example results can be provided upon request to the corresponding author.

This section presents and discusses the analysis results of survey data collected by Daynamica in the Twin Cities (see Section 3). First, we provide an example analysis of behaviors on Monday to illustrate the steps and results, using 1 min as the time interval and the unit-cost HAM for SAMs. Second, we repeat the analysis for seven days of the week, using four sample intervals to construct sequences and applying the four types of distances for SAMs. We compare the analysis results and discuss the performances of different methods and settings via graphical methods and FDA. Third, we relate the findings to user profiles and discuss the practical implications.

Table 3
The optimal number of clusters for different sample interval and distance matrices.

	Time (min)	Ham Unit	Ham FixFlex	Ham TripAct	Ham TRate	Ham DHD	OM Unit	OM FixFlex	OM TripAct	OM TRate
Monday	01	4	2	4	4	4	4	2	4	5
	03	4	2	4	4	4	6	2	4	4
	05	4	2	4	4	4	5	2	4	5
	10	4	2	3	6	5	4	2	5	4
Tuesday	01	9	2	7	9	9	5	2	5	8
	03	7	2	7	8	7	5	2	9	5
	05	7	2	5	6	3	5	2	5	4
	10	6	2	7	6	5	4	2	5	4
Wednesda	01	4	2	4	4	3	3	2	4	4
	03	5	2	4	3	3	3	2	4	4
	05	3	2	4	2	2	3	2	4	4
	10	4	2	5	5	5	3	2	5	5
Thursday	01	2	2	5	2	2	3	2	5	2
	03	8	2	4	4	4	3	2	7	3
	05	5	5	2	4	5	2	2	8	2
	10	8	2	2	6	4	6	2	7	3
Friday	01	4	2	5	4	4	5	2	6	4
	03	4	2	4	4	4	7	2	5	5
	05	4	2	5	4	4	4	2	4	2
	10	2	2	2	2	2	6	2	4	4
Saturday	01	2	2	2	2	2	4	2	3	4
	03	2	2	2	2	2	7	2	3	4
	05	2	2	2	2	2	7	2	3	8
	10	2	2	2	2	2	4	2	8	8
Sunday	01	3	2	2	3	3	2	2	2	2
	03	3	2	2	3	3	2	2	2	2
	05	2	2	3	3	2	2	2	2	5
	10	6	2	2	6	6	9	2	2	9

5.1. Example analysis

We apply the methods in Section 4.1 to transform activity diaries to activity-trip sequences using 1 min as the sample interval. Fig. 4 shows the number of sequences by day of the week and for each participant. The result is consistent with the distribution described in Section 3, Table 1. A sequence is invalid the participant started to record a non-home activity in the middle of a day (e.g. the first record of that day is a work/education episode at 10 am) or stopped recording in the early afternoon with a non-home activity (e.g. the last record is a shopping episode at 4 pm).

Fig. 5 shows the state distributions for each of a week. The x-axis corresponds to the time of a day the y-axis shows the state distributions among all possible activity and trip types.

The second step is to calculate pairwise distances between sequences using the unit-cost HM in Section 4.2. Since there are 344 sequences for Monday, the result here is a 344 by 344 symmetric matrix with each cell indicating the distance between two sequences. Since one person may have two or more sequences in one day of a week, we the average distances of all his/her sequences to measure his/her overall distance to other persons. If two persons both have multiple sequences of the same day of a week, we have a m by $n$matrix of the pairwise distance between them, and we use the mean value of all cells in that matrix to measure the overall distance between the two persons. Therefore, the result cost matrix will be a N by N matrix with each cell representing the distance between two individuals and N determined by the number of participants ($N = 369$).

The third step is to determine the optimal number of clusters and group participants into clusters with "similar" sequences. Fig. 6 plots partition quality versus the number of clusters for the five partition quality metrics described in Section 4.3. We use z-score because indicators may have quite different scales. Since four of the five measures suggest that four clusters provide the best performance, the optimal number of clusters is four and we use four clusters for the rest of our analysis in this example. If there is a special preference for partition quality, we can focus on the selected indicator(s). For instance, if we focus on the coherence of assignments and

Table 4
Global and dynamic cost matrices derived for Monday (1-minute interval).

Global Transition Rate	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	1.9773	0.0000							
	W	1.9842	1.9997	0.0000						
	H	1.9800	2.0000	1.9998	0.0000					
	L	1.9842	1.9997	1.9998	1.9992	0.0000				
	O	1.9878	2.0000	1.9999	1.9989	1.9999	0.0000			
	P	1.9797	1.9993	1.9999	1.9996	1.9979	1.9991	0.0000		
	S	1.9595	1.9998	2.0000	1.9989	1.9999	1.9998	1.9998	0.0000	
	N	1.9029	2.0000	1.9883	1.9884	1.9941	1.9908	1.9911	1.9997	0.0000

DHD Rate (9 AM)	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	3.0000	0.0000							
	W	3.9553	4.0000	0.0000						
	H	3.9778	4.0000	3.9889	0.0000					
	L	4.0000	4.0000	4.0000	4.0000	0.0000				
	O	4.0000	4.0000	4.0000	3.9945	4.0000	0.0000			
	P	3.8975	4.0000	4.0000	3.9945	4.0000	4.0000	0.0000		
	S	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000	
	N	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000

DHD Rate (9:01 AM)	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	4.0000	0.0000							
	W	3.8989	3.9944	0.0000						
	H	3.9888	4.0000	4.0000	0.0000					
	L	4.0000	4.0000	4.0000	4.0000	0.0000				
	O	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000			
	P	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000		
	S	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000	
	N	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000

DHD Rate (10:00 PM)	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	0.0000	0.0000							
	W	4.0000	4.0000	0.0000						
	H	3.8750	4.0000	3.9889	0.0000					
	L	4.0000	4.0000	4.0000	4.0000	0.0000				
	O	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000			
	P	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000		
	S	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000	
	N	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000

consider a cluster is homogeneous if most participants have large distances to other clusters, we should weigh more on ASWw than other indicators.

After determining the optimal number of clusters, we apply Ward's method to generate clusters of participants, each has one or more activity-travel sequences on Monday. Fig. 7 visualizes the four clusters of daily sequences and their state distributions. Note that the *N* by *N* matrix with pairwise distances between participants is used as input for clustering. But since the states are all activity and trip types, their values are categorical and do not have an average value. Hence, we plot all sequences of participants in a cluster to show behavior patterns. The behavior patterns of each cluster differ from each other. However, for other less frequent trip or activity types, it is hard to visually identify patterns and compare across different clusters.

Hence, we conduct the FDA to show the probabilities of a state happening across time in a day as described in Section 4.4. Fig. 8 shows the estimated probability distributions for each type of activities and trips. Note that the y-axis has different scales for each type to reflect its highest probability: (1) for home, work/education, personal business, and other non-home activities, the scale is [0,1]; (2) for trips, the scale is [0, 0.4], (3) for all other types, the scale is [0, 0.1]. This allows us to compare different patterns concerning those less frequent activities or trips. Take the type Trip as an example, the morning and afternoon peak hours for group 2 (regular workers) become more obvious, and group 3 (non-regular workers) has some trips at midnight which would be hardly identified in the state distribution (see Fig. 7). And for leisure activities, the ending time for late-night activities is around 3 am for group 3 (other), which cannot be detected in the state distribution. This is because the state distribution can only show observed frequency at each position

Table 5
Global and dynamic cost matrices derived for Monday (10-minute interval).

Global Transition Rate	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	1.8483	0.0000							
	W	1.8856	1.9840	0.0000						
	H	1.4866	1.9939	1.9948	0.0000					
	L	1.8884	1.9866	1.9966	1.9852	0.0000				
	O	1.9184	1.9922	1.9972	1.9869	1.9968	0.0000			
	P	1.8689	1.9863	1.9936	1.9882	1.9935	1.9948	0.0000		
	S	1.7661	1.9892	1.9866	1.9519	1.9934	1.9925	1.9930	0.0000	
	N	1.6741	1.9970	1.9113	1.9117	1.9701	1.9385	1.9396	1.9968	0.0000

DHD Rate (10:00 PM)	$\gamma_{\alpha\beta}$	Trip	Eat Out	Work	Home	Leisure	Other	PB	Shop	Wait
	T	0.0000								
	E	3.3333	0.0000							
	W	4.0000	4.0000	0.0000						
	H	3.2222	4.0000	4.0000	0.0000					
	L	3.7836	3.0000	4.0000	4.0000	0.0000				
	O	3.7460	4.0000	4.0000	3.7500	4.0000	0.0000			
	P	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000		
	S	3.4667	4.0000	4.0000	4.0000	4.0000	3.8000	4.0000	0.0000	
	N	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	4.0000	0.0000

(time) derived from a small number of observations ($n = 9$), but the FDA also accounts for other positions and applies a smooth function to estimate the probability using the observations in all clusters. Therefore, these estimated probabilities can quantitatively and more objectively describe the behavior patterns; they could also be used to simulate behaviors of each group in the future and therefore forecast travel and access demands over a day.

5.2. Parameter and metric selection

The analysis results are affected by the *sample interval* while creating activity-travel sequences and the choice of the *distance metrics* while measuring pairwise distances between sequences. To provide an overview of it, Table 3 list the optimal number of clusters for different settings across seven days of a week. Although the number of optimal numbers alone cannot infer any clustering patterns, they can prove that the settings do have an impact on the result. Besides, they can also provide some preliminary evaluation of the pros and cons of these settings. For instance, fixed-flexible-based OM and Ham distance tend to generate more generalized groups and are therefore less sensitive to the sample interval. Also, since Saturday have no peak hours for any activity or trip type (see Fig. 7), the Hamming distance tends to be less sensitive to the cost metrics.

Then, we use Monday and a 1-minute sample interval as an example to show how the transition rate affects the cost metric. Table 4 shows the substitution cost matrices based on the global transition rate and the three selected dynamic transition rates at 9:00am ($t = 540$), 9:01am ($t = 541$) and 10:00 pm ($t = 1320$). Since matrices are always symmetric, we only show half of it.

The global substitution matrix has 2 as the highest cost, indicating there is no transition happens between those two types of activities or trips at any sample times. For instance, home activities and eating outside never happened at two consecutive timestamps (minutes) in our data, and neither did work and shopping, eat out and other activities, and wait and eat out. And the smaller costs result from higher frequencies of two activity or trip types happening one after another. For example, if a trip happened at time t, waiting is the most frequent type that happened at $t-1$ or $t+1$, and shopping is the second. Note that these costs are relative: if person A is shopping and person B is working, their behaviors are less similar than, if person B is on a trip.

The dynamic metric calculates the substitution cost locally at each position t. Since the transition rates are calculated using conditional probabilities, they are very sensitive to the state distribution locally at time t. For example, although trip and wait has the lowest global substitution cost, for locally at all three selected time, their substitution costs are 4.0, meaning no transition happened within three minutes And the costs at 9:00am and 9:01am suggest that some trips start (or end at) that time. This becomes even more obvious at 10:00 pm, when most people are at home and the only transition that happened was between trip and home.

Increasing the sample interval may better capture the global transition between activity and trip episodes but at risk of misrepresentation. Table 5 shows cost matrices based on the 10-minute interval. The global costs become smaller in general, and the lowest cost for trips shift from wait to home. Similarly, the dynamic costs at 10:00 pm have a few smaller than 4.0. This suggests that using 10-minute seemingly have a better performance because it can generate cost matrices with higher variations. However, it may provide false messages such as it is possible to go shopping directly from work without any connecting trips, which is not possible (in our data) based on the global matrix in Table 4. Therefore, using a finer resolution can more objectively reflect the data.

After examining the impacts of the sample interval, we investigate how different distance metrics affect the clustering results and identified patterns. First, as shown in Table 3, different metrics would result in different numbers of optimal clustering. We choose Tuesday as an example given its various cluster optimal numbers. Fig. 9 shows that the increase in cluster number always increases the

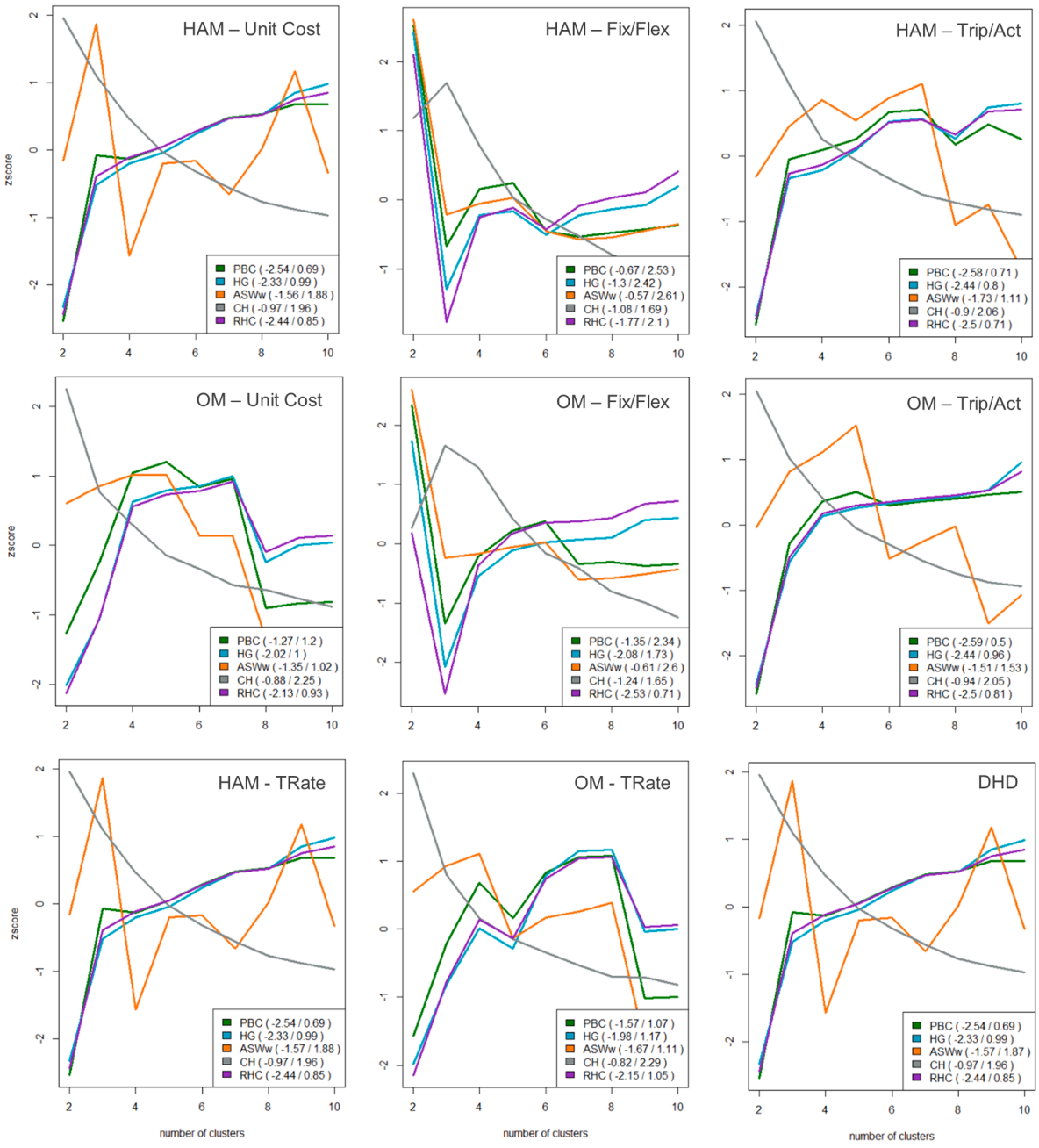


Fig. 9. Partition quality for Tuesday for nine distance measures (1 min).

maximum inter-cluster distance (RHC); the PBC, HG, and CH have similar patterns; and the coherence indicator ASWw has the most variations and could potentially be used as the basis for our example dataset. Using unit-cost HAM or OM as the base, the fixed-flexible metric has the most significant impact on the patterns; and the choice of HAM and OM does not affect the overall pattern much.

Occasionally, there is no absolute optimal choice. For instance, both 4 and 8 clusters have higher scores than others for the transition-based OM distance. So, we compare the state distributions of 4 and 8 cluster options (Fig. 10). The *regular working* (Group 2) out of the 4 clusters are divided into two clusters if we have 8 clusters: Group 3 has regular working hours from 8 am to 6 pm, travels during peak hours, and stays at home at night; and Group 2 that have more work/leisure activities before going home. Similarly, the *home stayer* (Group 1) in 4-cluster are divided into two clusters in the 8-cluster case (Group 1 and Group 6), depending on whether there are activities other than staying at home in the afternoon. Moreover, the *others* (Group 4) in 4-cluster are divided into three sub-clusters in 8-cluster, depending on whether and when personal business is done. Therefore, 4 clusters provide general patterns such as worker and non-workers, while 8 clusters can provide more detailed behavior patterns.

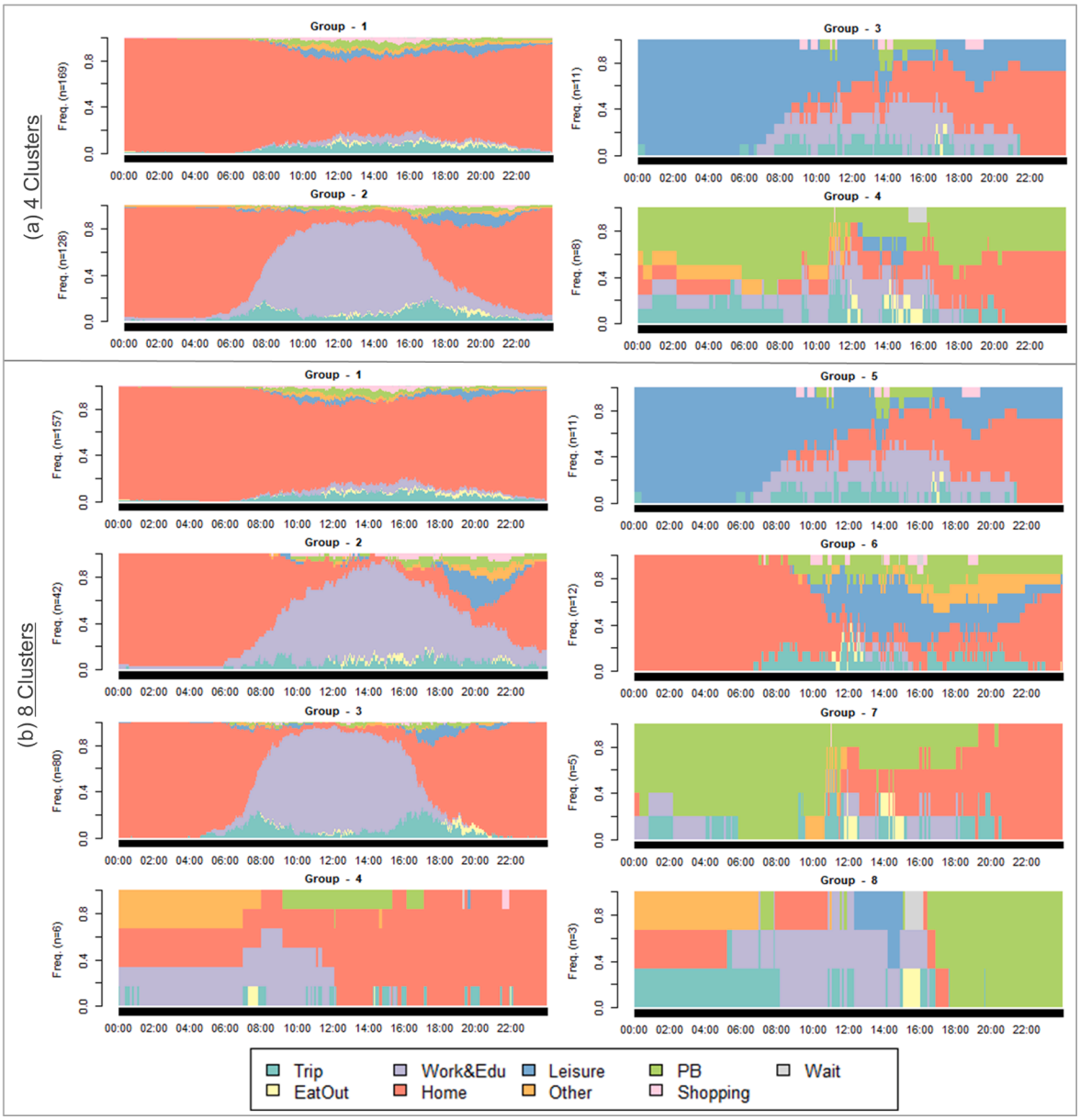


Fig. 10. State distributions of sequences for (a) 4 clusters and (b) 8 clusters.

5.3. Behavior analysis

Based on the sensitivity analysis, we choose the hamming distance family and use 1 min as the sample interval to examine the behavior patterns in the study area. To start, we use Monday to demonstrate the pros and cons of different cost metrics. First, we visualize and compare the state distributions based on *unit-cost*, *fixed-flexible*, *trip-activity*, *transition-based*, and *DHD* cost metrics. We choose Monday because four metrics have 4 clusters, while the fixed-flexible metric always has 2 as its optimal number; and we can mitigate the impacts of using the same number of clusters instead of the optimal one for visual comparison. We find that the state distributions for trip-activity, transition-based, and DHD are almost identical, and the counts of sequences in each cluster are the same too (except that trip-activity has 1 more sequence in Group 2 than the other two). So, we only present the unit-cost, DHD, and fixed-flexible in Fig. 11. Besides, we also include the original fixed-flexible, 2-cluster results for comparison purposes.

Fig. 11 shows that the unit-cost and DHD metrics generate similar group patterns. The number of sequences in Group 1 and 2 is slightly different; we find that this is caused by 10 participants move between these two groups; Group 2 based on the DHD metric has a lower percentage of home activities from 10 am to noon; Group 3 and Group 4 have the same group of participants using this two metrics. By checking the participants in each group, we confirm that the unit-cost and DHD metrics generate very similar results. Then,

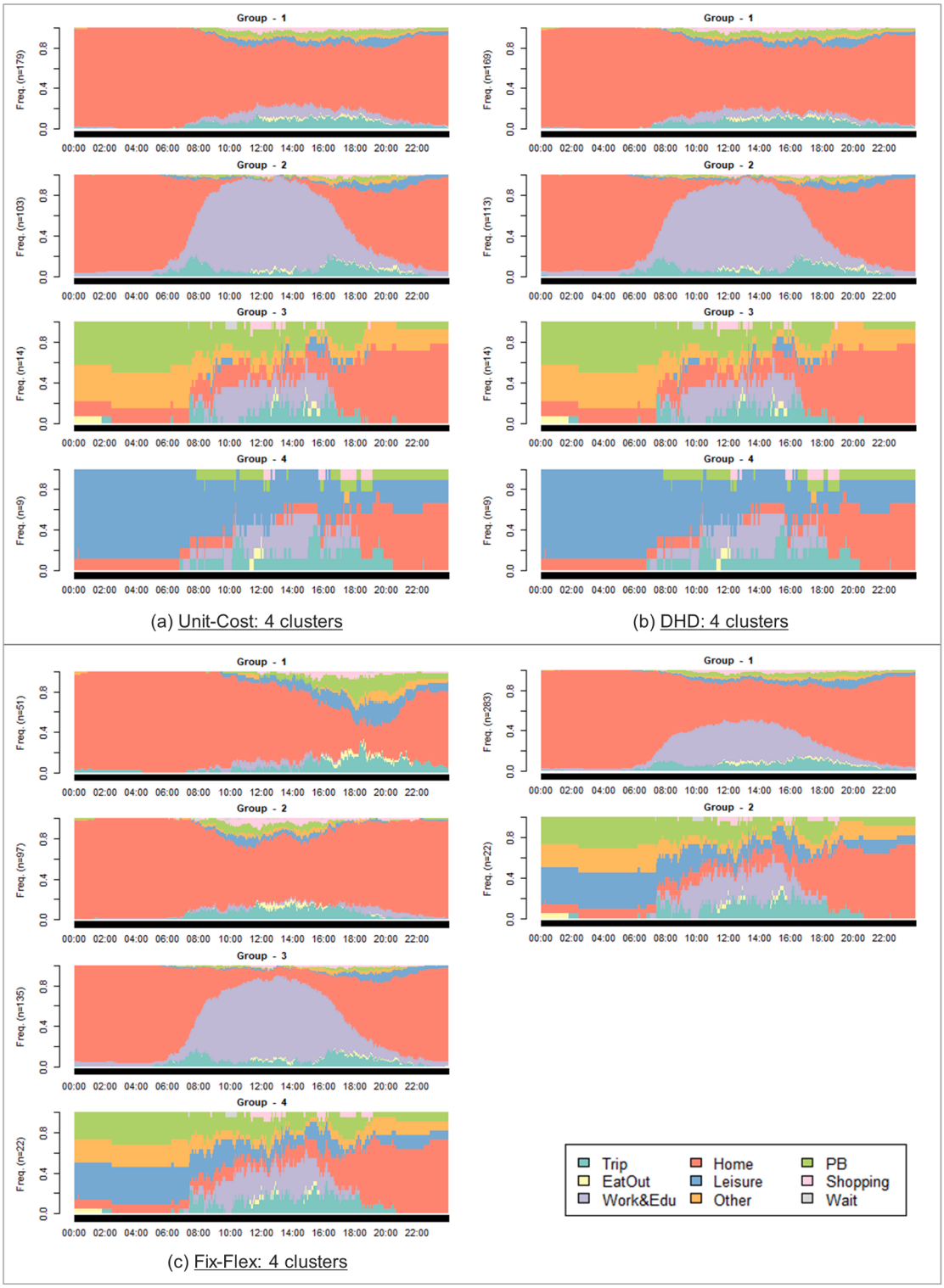


Fig. 11. Different group patterns using unit-cost, fixed-flexible and DHD metrics.

we compare unit-cost and fixed-flexible metrics (for both 2 and 4 clusters). We find that the 22 participants in Group 4 of the 4 clusters and Group 2 of the 2 clusters are the same group of people. These 22 participants are belonging to either Group 3 or Group 4 in both the unit-cost and DHD results. This suggests that the fixed-flexible metric does not further distinguish non-typical patterns that contain

Table 6
Percentage of participants in each cluster related to their profiles.

Variable	Group 1 N = 51	Group 2 N = 97	Group 3 N = 135	Group 4 N = 22	Base percentage
General					
Female	60.78	65.98	67.41	54.55	66.03
Senior (over 65)	29.41	30.93	14.07	27.27	20.38
Driver's License	90.2	82.47	96.3	90.91	88.59
Working Status					
Employed full-time	17.65	18.56	73.33	31.82	42.12
Employed part-time	29.41	38.14	28.89	22.73	30.16
Student full-time	5.88	6.19	14.07	9.09	8.97
Student part-time	7.84	2.06	3.7	13.64	5.16
Homemaker	27.45	19.59	2.22	4.55	12.50
Retired	31.37	30.93	4.44	22.73	20.11
Unemployed	15.69	12.37	0.74	22.73	9.78
Spatial Variation					
Philips	17.65	22.68	13.33	22.73	19.02
Near North	25.49	13.4	7.41	22.73	15.22
Prospect Park	13.73	22.68	21.48	9.09	17.93
St. Anthony Park	23.53	17.53	20.74	22.73	19.57
Blaine	13.73	9.28	17.04	18.18	14.13
Brooklyn Center	5.88	14.43	20	4.55	14.13

Table 7
Average time spent on out-home activities/trips on Monday for each group.

Activity/trip types	Group 1 N = 51	Group 2 N = 97	Group 3 N = 135	Group 4 N = 22	Base avg. minutes
Work/education	124.1	157.8	445.8	218.7	367.3
Personal business	182.7	82.8	46.7	268.1	115.3
Leisure/recreation	236.0	93.9	112.5	336.0	153.9
Eat out	67.2	47.9	38.8	23.3	45.5
Shopping	53.1	45.2	30.9	35.8	41.1
Other	196.6	56.1	51.5	213.7	107.1
Trip	105.4	65.8	67.7	77.8	73.9
Wait	5.6	15.9	6.6	25.5	10.8

non-home activities extending from or toward midnight. Instead, it further divides the *home-stayer* group (Group 1) pattern for unit-cost based on whether there are out-home activities or trips or not.

To quantitatively compare 4 clusters based on the unit-cost and fixed-flexible weighted cost, we conduct the FDA for fixed-flexible cost metric. In general, fixed-flexible cost generates more unique groups: (1) it identify a group of participants (n = 51) who stay at home during the day but go out for trip and leisure (among other non-home) activities during the early night; (2) it has a generalized *late-night movers* who have leisure (among other non-home) activities and return home mostly in the afternoon; and (3) it can still identify the *regular workers* (n = 135) who leave home in the morning and return home at night, mostly travel during morning and afternoon peak hours, and have leisure activities mainly after work at night. This suggests that the fixed-flexible weighted cost metric works the best to detect distinct behavior patterns on Monday.

Therefore, we use the Monday clustering results based on fixed-flexible weighted cost and relate the group behavior patterns to the user profiles collected by the entrance survey (see Section 2, Table 1). Table 6 shows the percentage of participants in each clustering group, and we include a global percentage as a reference to show how each cluster may have a different profile from the entire sample. For Group 3, the most significant variation, not surprisingly, is between full-time employers and people with another employment status. Spatially, the Near North neighborhood has lower than usual *regular workers* and has leisure activities in the afternoon. Another interesting finding is that student part-time workers are more likely to have a *late-night movers* type, which is intuitive. This suggests that people prefer not to travel too early in the morning or too late at night.

To further describe the behavior patterns of each group, we examine the time allocations among different out-home activities or trips on Monday. Table 7 shows that each group has its obvious and unique patterns of time allocation. For instance, Group 3 spent an average of 445.8 min (7.43 h) on working or education, while the other three groups spend less than 3.65 hours on average, Group 2 has the least time spent on out-home activities, and Group 1 and 4 have a decent amount of time spent among various out-home trips and activities. This proves that our methods can capture different time allocation patterns like existing studies. Compared to those studies, our methods can also capture the sequence of these allocated times and identify unique patterns such as Group 4 whose out-home activities and trips started from midnight (as shown in Figs. 11 and 12). Due to the relatively small sample, we did not conduct statistical analysis using (e.g. multivariate regression).

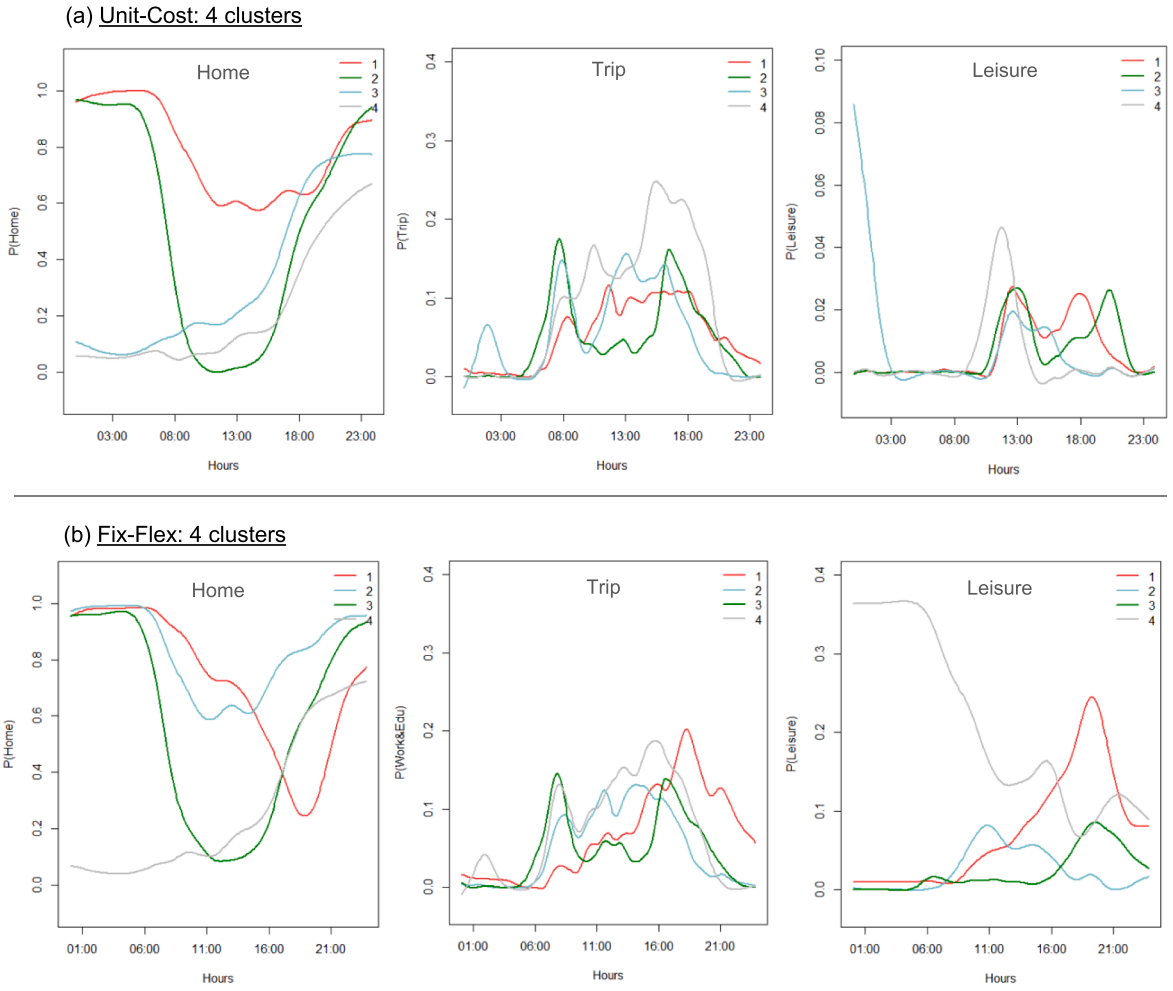


Fig. 12. State probabilities estimated through FDA for Home, Trip, and Leisure.

6. Conclusion and discussion

We introduced a sequence alignment-based approach to characterize the distance between daily activity-travel behavior patterns. We define unit-cost, fixed-flexible, trip-activity, and transition-based distance measures to address key theories in time geography and activity-based transport research. Using a GPS-based travel survey collected in the Twin Cities areas, we demonstrate our methods in detail. And we conduct functional data analysis in addition to graphical methods to describe and compare clusters of participants with similar activity-travel sequences. While the visualization of state distributions can provide a direct view of the general patterns, the FDA can statistically describe behavior patterns of a cluster relate to all other clusters. The FDA allows us to estimate probabilities of a trip or activity occur at a given time, and plots of such probability distributions allow us to describe patterns for the less frequent trip/activity types like shopping and leisure recreations statistically. Finally, we also relate each identified cluster and its behavior patterns to the participants' characters to further explain the identified patterns. The percentage of participants in each group are not always the same as that for all participants; the patterns are consistent with our intuitive expectation (e.g. full-time employee and student works are likely in the *regular worker* group); some novel insights may arise (e.g. two neighborhoods have fewer *nighttime movers*, likely due to the built environment such as safety). These findings, taking together, can support further statistical analysis (given a larger sample) to test the significance of each character on the behavior pattern outcomes.

To ensure rigor in the analysis results, we include a thorough investigation of the impact of two-parameter settings: the sample interval used to create activity-travel sequences, and the choice of cost metrics for sequence alignment. We find that the optimal number of clusters based on some common partition quality indices may vary a lot when we apply different settings, and the most *stable* one is the fixed-flexible weighted cost that focuses on distinguishing how we schedule our flexible activities and trips between home and work/education. Then, we further investigate the transition-based substitution costs derived from the data. We find the global cost metric can better capture the global patterns on changing from one activity/trip type to another, and the local cost metrics are more sensitive to the local state distribution within a small time window and can better represent the cost of schedule change locally at a

given time. For instance, even though the cost between Trip and Work is lower than the other, at 10:00 pm, the cost would be high because there are few trips from or to work during the night. We also find that larger sample intervals (10 min) may better reflect the general trend but may ignore small trips/activities, and it is better to use a smaller interval (1 min) to mitigate the impacts of discrete-time sampling and capture the actual transition rate in the original data. Then we examine the impacts of the four weighting schemas (unit-cost, fixed-flexible, trip-activity, transition-based) and two alignment methods (hamming distance and optimal matching). The comparison of partition quality suggests that the selection of weighting schemas plays a more important role than the choice of HAM and OM. And when two cluster numbers have similar performance, clusters based on a larger number are often the subsets of clusters based on a smaller number. Therefore, based on the specific aims of a project, we could choose fewer clusters to identify generalized patterns and choose more clusters to extract and study the non-typical behaviors.

Our weighted sequence alignment approach, which adopts a discrete view of the time dimension, can be extended in several directions. First, by encoding daily activity-travel patterns as current-state sequences, it does not explicitly incorporate spatial information which may better explain the behavior patterns. One possible strategy for integrating spatial data would be to enrich the set of state indicators to incorporate spatial information (e.g., by defining activity types "H1", "H2", "H3" to distinguish between urban, suburban, and rural home locations, or defining "C1" and "C2" as car trips along local roads or highways). However, this strategy would rapidly become infeasible as the number of distinct location types encoded increases. In ongoing work, we are investigating alternative approaches to integrating spatial information into distance metrics, and for comparing the resulting clusters.

Similarly, besides activity and trip types, our approach could also be applied to other dimensions of activity-travel sequences. For example, in the Twin Cities travel survey we used in this paper, participants were asked to self-report their emotional experiences (e.g. happy, stressful) on a 1–6 scale for each trip and activity. These self-reported emotion data could be used to generate mood sequences. which could be aligned and clustered using our approach. Indeed, another interesting avenue of future research is how to visualize, cluster, and describe multi-dimensional sequences defined by activity patterns plus other sensor-derived or self-reported measurements such as the physical activity measurements generated by accelerometers and the subjective wellbeing status mentioned earlier).

In addition to the thematic enrichment, we can also characterize shorter- or longer-term patterns than daily patterns using the same basic techniques and procedures. In this paper, one notable challenge that arises for capturing longer-term (e.g., weekly) patterns is the *missing data*. For example, in our dataset, each participant should have between 5 and 10 days of activity dairy data available. But participants often started collecting data on different days that are not consecutive and had 1 to 3 days of incomplete/invalid data. A possible solution is to apply methods for partially overlapping activity sequence alignment with a substantial amount of missing information. This would allow us to examine the weekly patterns (or other periods) beyond individual days, and probably gain new insights into the activity-travel scheduling.

Finally, we can apply the framework to other study areas where detailed activity-travel episodic data are available. This would allow us to validate and further refine our current four weighting schema, clustering algorithms, and functional data analysis settings. For instance, in the field of time geography, both home and work/education are considered as fixed activities. However, in the field of transportation planning that examines trip chaining patterns, home activities are often not considered as fixed, but instead the most flexible because substituting a home activity with out-of-home activities would be easier than rescheduling ongoing out-of-home activities such as shopping or recreation. Hence, it would be worth refining our current weighting schema to adopt a length-based transition rate to account for the previous and later activity episodes and its type to define mesoscale transition-based cost metrics. Another solution is to use the tour as the analysis unit and define a hierarchical way to define a tour-based transition cost metric (e.g. the work by Su et al., 2020 used a similar approach). This may allow us to also study the multi-modal trip scheduling by further consider the trip legs and encode each differently according to their type. For instance, for a trip from home to work, it will be labeled differently based on primary mode (e.g. transit) and supporting modes (e.g. walk-transit-walk and bike-transit-bike would be encoded differently). Again, by including more details, we are at risk of introducing additional uncertainty on parameter settings, and therefore we need to conduct the sensitivity analysis as in this paper for future research.

CRediT authorship contribution statement

Ying Song: Formal analysis, Methodology, Software, Conceptualization, Visualization. **Siyang Ren:** Software, Methodology, Visualization, Formal analysis, Conceptualization. **Julian Wolfson:** Conceptualization, Supervision, Methodology, Formal analysis. **Yaxuan Zhang:** Data curation, Visualization, Methodology. **Roland Brown:** Conceptualization, Visualization, Methodology. **Yingling Fan:** Conceptualization, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbott, A., Tsay, A., 2000. Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociol. Methods Res. https://doi.org/10.1177/0049124100029001001.

Auld, J., Mohammadian, A.K., 2012. Activity planning processes in the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. Transp. Res. Part A: Policy and Practice. https://doi.org/10.1016/j.tra.2012.05.017.

Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems: Foreign summaries. Transport Rev. https://doi.org/10.1080/01441649208716826.
Becker, R., Ramón, C., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C., 2013. Human mobility characterization from cellular network data. Commun. ACM. https://doi.org/10.1145/2398356.2398375.
Berger, M., Platzer, M., 2015. Field evaluation of the smartphone-based travel behaviour data collection app "smartMo". Transp. Res. Procedia. https://doi.org/10.1016/j.trpro.2015.12.023.
Bhat, C.R., Goulias, K.G., Pendyala, R.M., Paleti, R., Sidharthan, R., Schmitt, L., Hu, H.H., 2013. A household-level activity pattern generation model with an application for Southern California. Transportation. https://doi.org/10.1007/s11116-013-9452-y.
Bhat, C.R., Koppelman, F.S., 2006. Activity-based modeling of travel demand. Handbook Transp. Sci. https://doi.org/10.1007/0-306-48058-1_3.
Bowman, J.L., Ben-Akiva, M.E., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transp. Res. Part A: Policy Practice. https://doi.org/10.1016/S0965-8564(99)00043-9.
Bradley, M., Bowman, J.L., Griesenbeck, B., 2010. SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. J. Choice Modelling. https://doi.org/10.1016/S1755-5345(13)70027-7.
Bras, H., Liefbroer, A.C., Elzinga, C.H., 2010. Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. Demography. https://doi.org/10.1007/BF03213737.
Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Statistics - Simulation Computation. https://doi.org/10.1080/03610917408548446.
Cho, S., et al., 2013. An activity-based carpooling microsimulation using ontology. Proc. Comput. Sci. https://doi.org/10.1016/j.procs.2013.06.012.
Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. https://doi.org/10.1093/nar/16.22.10881.
Das, R.D., Winter, S., 2016. Detecting Urban transport modes using a hybrid knowledge driven framework from GPS trajectory. ISPRS Int. J. Geo-Inf. https://doi.org/10.3390/ijgi5110207.
Ettema, D.F., Timmermans, H.J.P., 1997. Activity-based approaches to travel analysis. In: Activity-based approaches to travel analysis.
Ettema, D., Borgers, A., Timmermans, H., 1993. Simulation model of activity scheduling behavior. Transport. Res. Rec.
Fan, Y., Wolfson, J., Adomavicius, G., Vardhan Das, K., Khandelwal, Y., Kang, J., 2015. SmarTrAC: A Smartphone Solution for Context-Aware Travel and Activity Capturing. Access at: https://conservancy.umn.edu/handle/11299/173005 (on March 7, 2021).
Fan, Y., Brown, R., Das, K., Wolfson, J., 2019. Understanding trip happiness using smartphone-based data: the effects of trip- and person-level characteristics. Transport Findings. https://doi.org/10.32866/7124.
Febrero-Bande, M., Oviedo de la Fuente, M., 2011. Refrence manual of R package: functional data analysis and utilities for statistical computing (fda.usc). The Comprehensive R Archive Network (CRAN).
Fonseca, F.T., Egenhofer, M.J., Agouris, P., Cmara, G., 2002. Using ontologies for integrated geographic information systems. Trans. GIS. https://doi.org/10.1111/1467-9671.00109.
Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M., 2011. Analyzing and visualizing state sequences in R with TraMineR. J. Stat. Softw. https://doi.org/10.18637/jss.v040.i04.
Gärling, T., Kwan, M.P., Golledge, R.G., 1994. Computational-process modelling of household activity scheduling. Transp. Res. Part B. https://doi.org/10.1016/0191-2615(94)90034-5.
Goldsmith, J., Zipunnikov, V., Schrack, J., 2015. Generalized multilevel function-on-scalar regression and principal component analysis. Biometrics. https://doi.org/10.1111/biom.12278.
Goulias, K.G., Pendyala, R.M., Bhat, C.R., 2013. Keynote — total design data needs for the new generation large-scale activity microsimulation models. In: Transport Survey Methods: Best Practice for Decision Making. https://doi.org/10.1108/9781781902882-002.
Hafezi, M.H., Liu, L., Millward, H., 2019. A time-use activity-pattern recognition model for activity-based travel demand modeling. Transportation. https://doi.org/10.1007/s11116-017-9840-9.
Hägerstrand, T., 1970. What about people in regional science? Papers of the Regional Science Association. https://doi.org/10.1007/BF01936872.
He, M.X., Petoukhov, S.V., Ricci, P.E., 2004. Genetic code, hamming distance and stochastic matrices. Bull. Math. Biol. https://doi.org/10.1016/j.bulm.2004.01.002.
Hubert, L.J., Levin, J.R., 1976. A general statistical framework for assessing categorical clustering in free recall. Psychol. Bull. https://doi.org/10.1037/0033-2909.83.6.1072.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. https://doi.org/10.1007/BF01908075.
Joh, C.H., Arentze, T.A., Timmermans, H.J.P., 2001. A position-senstive sequence-alignment method illustrated for space-time activity-diary data. Environ. Planning A. https://doi.org/10.1068/a3323.
Joh, C.H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity pattern similarity: A multidimensional sequence alignment method. Transp. Res. Part B: Methodol. https://doi.org/10.1016/S0191-2615(01)00009-1.
Joh, C.H., Timmermans, H., 2011. Applying sequence alignment methods to large activity-travel data sets: Heuristic approach. Transp. Res. Rec. https://doi.org/10.3141/2231-02.
Joh, C.-H., Arentze, T.A., Timmermans, H.J.P., 2010. Multidimensional sequence alignment methods for activity-travel pattern analysis: a comparison of dynamic programming and genetic algorithms. Geographical Anal. https://doi.org/10.1111/j.1538-4632.2001.tb00447.x.
Jones, P.M., Dix, M.C., Clarke, M.I., Heggie, I.G., 1983. Understanding Travel Behaviour. Gower, Aldershot, Hants.
Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data: an introduction to cluster analysis (wiley series in probability and statistics). In: Eepe.Ethz.Ch. https://doi.org/10.1007/s13398-014-0173-7.2.
Kitamura, R., Chen, C., Pendyala, R.M., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation. https://doi.org/10.1023/A:1005259324588.
Kitamura, R., Kermanshan, M., 1984. Sequential model of interdependent activity and destination choices. Transp. Res. Rec. 987, 81–89.
Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings Bioinf. https://doi.org/10.1093/bib/5.2.150.
Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE. https://doi.org/10.1371/journal.pone.0096180.
Kwan, M.P., Xiao, N., Ding, G., 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. Geographical Anal. https://doi.org/10.1111/gean.12040.
Lesnard, L., 2009. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociol. Methods Res. https://doi.org/10.1177/0049124110362526.
Li, H., 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty191.
Liu, F., Janssens, D., Cui, J., Wets, G., Cools, M., 2015. Characterizing activity sequences using profile hidden markov models. Expert Syst. Appl. https://doi.org/10.1016/j.eswa.2015.02.057.
McNally, M.G., Rindt, C.R., 2007. The Activity-Based Approach. https://doi.org/10.1108/9780857245670-004.
Meng, C., Cui, Y., He, Q., Su, L., Gao, J., 2017. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. https://doi.org/10.1109/BigData.2017.8258062.
Miller, E.J., Roorda, M.J., 2003. Prototype model of household activity-travel scheduling. Transp. Res. Rec. https://doi.org/10.3141/1831-13.
Miller, H.J., 1991. Modelling accessibility using space-time prism concepts within geographical information systems. Int. J. Geograph. Inf. Syst. https://doi.org/10.1080/02693799108927856.
Miller, H.J., 2017. Time geography and space-time prism. Int. Encyclopedia Geography. https://doi.org/10.1002/9781118786352.wbieg0431.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika. https://doi.org/10.1007/BF02294245.
Müller, K., Axhausen, K.W., 2011. Population synthesis for microsimulation : State of the art. 90th Annual Meeting of the Transportation Research Board.
Murakami, E., & Wagner, D. P. (1999). Can using Global Positioning System (GPS) improve trip reporting? Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/S0968-090X(99)00017-0.
Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. https://doi.org/10.1093/comjnl/26.4.354.
Nahmias-Biran, B.H., Han, Y., Bekhor, S., Zhao, F., Zegras, C., Ben-Akiva, M., 2018. Enriching activity-based models using smartphone-based travel surveys. Transp. Res. Rec. https://doi.org/10.1177/0361198118798475.
Norouzi, M., Fleet, D.J., Salakhutdinov, R., 2012. Hamming distance metric learning. Advances in Neural Information Processing Systems.
Ohmori, N.T., Nakazato, M.T., Harata, N.T., Kuniaki, S.Y., Kazuo, N.Y., 2006. Activity diary surveys using GPS mobile phones and PDA. TRB Annual Meeting.
Pendyala, R., Konduri, K., Chiu, Y.C., Hickman, M., Noh, H., Waddell, P., Wang, L., You, D., Gardner, B., 2012. Integrated land use-transport model system with dynamic time-dependent activity-travel microsimulation. Transp. Res. Rec. https://doi.org/10.3141/2303-03.
Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O., 2018. MEILI: A travel diary collection, annotation and automation system. Comput. Environ. Urban Syst. https://doi.org/10.1016/j.compenvurbsys.2018.01.011.
Ramsay, J.O., 2006. Functional data analysis. Encyclopedia of Statistical Sciences. https://doi.org/10.1002/0471667196.ess3138.
Recker, W.W., McNally, M.G., Root, G.S., 1986. A model of complex travel behavior: Part I-Theoretical development. Transp. Res. Part A: General. https://doi.org/10.1016/0191-2607(86)90089-0.
Saadi, I., Mustafa, A., Teller, J., Cools, M., 2016. Forecasting travel behavior using Markov Chains-based approaches. Transp. Res. Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2016.06.020.
Saarloos, D., Kim, J.E., Timmermans, H., 2009. The built environment and health: Introducing individual space-time behavior. Int. J. Environ. Res. Public Health. https://doi.org/10.3390/ijerph6061724.
Saneinejad, S., Roorda, M.J., 2009. Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. Transp. Lett. https://doi.org/10.3328/TL.2009.01.03.197-211.
Schwanen, T., Kwan, M.P., Ren, F., 2008. How fixed is fixed? Gendered rigidity of space-time constraints and geographies of everyday activities. Geoforum. https://doi.org/10.1016/j.geoforum.2008.09.002.
Seo, T., Kusakabe, T., Gotoh, H., Asakura, Y., 2019. Interactive online machine learning approach for activity-travel survey. Transp. Res. Part B: Methodol. https://doi.org/10.1016/j.trb.2017.11.009.
Shoval, N., Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. Ann. Assoc. Am. Geogr. https://doi.org/10.1111/j.1467-8306.2007.00536.x.
Shoval, N., McKercher, B., Birenboim, A., Ng, E., 2015. The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. Environ. Planning B: Planning Des. https://doi.org/10.1068/b38065.
Simini, F., González, M.C., Maritan, A., Barabási, A.L., 2012. A universal model for mobility and migration patterns. Nature. https://doi.org/10.1038/nature10856.
Song, C., Koren, T., Wang, P., Barabási, A.L., 2010. Modelling the scaling properties of human mobility. Nature Physics. https://doi.org/10.1038/nphys1760.
Studer, M., 2013. WeightedCluster Library Manual. *Cran*. https://doi.org/10.12682/lives.2296-1658.2013.24.
Studer, M., Ritschard, G., 2016. What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. J. Royal Stat. Soc. Series A: Statistics Soc. https://doi.org/10.1111/rssa.12125.
Studer, M., Ritschard, G., 2014. A comparative review of sequence dissimilarity measures. LIVES Working Papers.
Su, R., McBride, E.C., Goulias, K.G., 2020. Pattern recognition of daily activity patterns using human mobility motifs and sequence analysis. Transp. Res. Part C: Emerging Technol. https://doi.org/10.1016/j.trc.2020.102796.
Ullah, S., Finch, C.F., 2013. Applications of functional data analysis: A systematic review. BMC Med. Res. Method. https://doi.org/10.1186/1471-2288-13-43.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.1963.10500845.
Widmer, E.D., Ritschard, G., 2009. The de-standardization of the life course: Are men and women equal? Adv. Life Course Res. https://doi.org/10.1016/j.alcr.2009.04.001.
Wilson, C., 2006. Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. Environ. Planning A. https://doi.org/10.1068/a3722.
Wilson, C., 2008. Activity patterns in space and time: Calculating representative Hagerstrand trajectories. Transportation. https://doi.org/10.1007/s11116-008-9162-z.
Wilson, W.C., 1998. Activity pattern analysis by means of sequence-alignment methods. Environ. Planning A. https://doi.org/10.1068/a301017.
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. https://doi.org/10.1007/s11116-015-9666-2.
Zhai, W., Bai, X., Peng, Z., ren, Gu, C., 2019. From edit distance to augmented space-time-weighted edit distance: Detecting and clustering patterns of human activities in Puget Sound region. J. Transp. Geogr. https://doi.org/10.1016/j.jtrangeo.2019.05.003.
Zhao, F., Pereira, F.C., Ball, R., Kim, Y., Han, Y., Zegras, C., Ben-Akiva, M., 2015. Exploratory analysis of a smartphone-based travel survey in Singapore. Transp. Res. Rec. https://doi.org/10.3141/2494-06.
Ziemke, D., Nagel, K., Moeckel, R., 2016. Towards an agent-based, integrated land-use transport modeling system. Procedia Comput. Sci. https://doi.org/10.1016/j.procs.2016.04.192.