# Longest common subsequences: Identifying the stability of individuals' travel patterns

Adrian C. Prelipcean*[1], Yusak O. Susilo[1] and Győző Gidófalvi[2]

[1] Department of Transport Science, KTH, Sweden
[2]Department of Urban Planning and Environment, KTH, Sweden

**Abstract**

There is a strong consensus in the travel behaviour research community that the one day travel diary collection is insufficient to understand the finer aspects of behaviour that transcend attributes such as average trip length, duration, travel modes, etc. While a large body research was done on exploring the spatial, temporal and spatio-temporal travel behavioural patterns, the sequential aspect of behaviour is seldom studied. The consensus of the few papers that have studied travel behaviour variability from a sequential perspective has been to use edit distance and compute the costs of transforming one day of travel activities into another. While useful, this approach generates difficult to understand metrics since it does not directly extract (sub)sequences but computes penalties. This paper provides an alternative for investigating the sequential aspect of travel behaviour that makes use of longest common subsequences to extract the activities that are common to multiple days and / or users. The proposed methodology provides indexes for measuring the inter- and intra-personal stability of a given user base and its usefulness is proved in a case study on travel diaries collected from 51 users for a period of 7 days.

## 1 Introduction

When it comes to the study of day to day variability of individuals, the pioneering researchers sought to understand the dimensions that can explain the variability and stability of scheduling with regards to activity, travel modes, destinations or sociodemographics (Hanson & Huff 1981, Huff & Hanson 1986, Pas 1987, Pas & Koppelman 1987, Hanson & Huff 1988, Jones & Clarke 1988). While the findings differ in nature due to the different focus, e.g., Hanson & Huff (1981) proposed similarity measures to measure daily variability, Pas & Koppelman (1987) augmented the traditional goodness of fit measure to include inter-

*Corresponding author acpr@kth.se

and intra-personal variability, and Jones & Clarke (1988) explored methods to better visualize and explain similarity between days, they identified one of the main problems of transportation science: the insufficiency of one-day travel diary collection efforts. This became the basis of future research that focused on understanding what information can be extracted from travel diaries collected over multiple days (Axhausen et al. 2002, Schlich & Axhausen 2003, Kitamura et al. 2006, Neutens et al. 2012, Järv et al. 2014, Minnen et al. 2015), and on identifying the desired periods over which travel diaries should be collected to extract the aforementioned information (Cherchi et al. 2017).

The research that continued the studies of Hansson and Pas has focused on different aspects of travel behaviour such as: 1) *spatial variability* to identify the relationship between activities and the locations where individuals perform them (Susilo & Axhausen 2014, Buliung et al. 2008), 2) *temporal variability* to identify distinct core behaviors that occur at regular time intervals (Hanson & Huff 1981, 1988, Pas 1987, Thøgersen 2006, Heinen & Chatterjee 2015, Minnen et al. 2015), 3) *spatio-temporal variability* to explore the relationship between activity spaces (Dijst 1999) and space time prisms (Hägerstrand 1970) in interpersonal and intra-personal settings (Srivastava & Schönfelder 2003, Susilo & Kitamura 2005, Kitamura et al. 2006, Järv et al. 2014, Dharmowijoyo et al. 2016, Manley et al. 2016), and 4) *sequential variability* to identify the similarity of the order in which activities are performed in between days (Wilson 1998, Joh et al. 2001*b*,*c*, Moiseeva et al. 2014, Allahviranloo et al. 2016).

While it is straightforward to grasp and explain temporal, spatial and spatio-temporal variability due to the relationship between the studied concepts (i.e., repetition across the dimensions of time and / or space) and the proposed variability measures (i.e., indexes based on frequency of occurrence), the research studying sequential variability is based on penalties derived from Edit Distance (ED) alignment, which is a concept that is clearer when applied to strings than to activities. The ED is the minimum number of single-character edits – i.e., insertions, deletions and substitutions – that are required to change one sequence into another (Levenshtein 1966) and is primarily used to measure similarity between sequences. As such, researchers used ED measures to generate similarity indexes on schedules represented as strings (Wilson 1998, Joh et al. 2001*b*,*c*, Moiseeva et al. 2014, Allahviranloo et al. 2016), which has constituted the overwhelming majority of studies on sequences in this field. However, there are different issues with using ED-based indexes such as the difficulty of understanding what the index values represent since they are not as straightforward or descriptive as the temporal, spatial and spatio-temporal variability measures (Schlich & Axhausen 2003) and the results are heavily influenced by the chosen penalties for ED operations (Allahviranloo et al. 2016).

This paper proposes a new approach to measure sequential variability of activity schedules that is straightforward to explain and implement. The new approach is based on the concept of Longest Common Subsequence (LCS), which extracts the elements within a group of sequences that occur in the same order, while allowing gaps in between elements (Hirschberg 1975). Based on these concepts, a new type of indexes to measure the sequential stability of schedules is

proposed. The index values are straightforward to understand, e.g., if the index between the schedule of individual A and individual B is 50%, then individual A performs half of her activities in the same order as individual B. The paper puts an emphasis on the properties of the proposed index measure and illustrates its applicability by analyzing the stability of schedules of 51 users who recorded 1250 trips between 2nd and 11th of November 2016.

The remainder of the paper is organized as follows: Section 2 provides a literature review on similarity measures and LCS usage, Section 3 provides the preliminaries needed to understand the remainder of the paper, compares ED with LCS and explains the proposed index measures, Section 4 provides a case study to illustrate the usefulness of the proposed methodology, Section 5 provides discussions about the current disadvantages of the proposed methodology, and finally, Section 6 provides future work directions and concludes.

# 2 Literature review

This section is split in two main parts. First, it discusses the index measures that have been previously proposed in the transportation literature to measure travel behaviour variability. Second, it introduces applications of LCS that are closely related to travel behaviour.

## 2.1 Measuring travel behaviour variability via indexes

The main corpus of research that investigates travel behaviour variability can be traced back to the initial visionary work of three main research groups: 1) Hanson and colleagues (Hanson & Huff 1981, 1988, Huff & Hanson 1986), who collected travel diaries for more than 35 consecutive days in Uppsala, Sweden and proposed an index value that measures repetition based on equivalence classes mapped on a contingency table, 2) Recker and colleagues (Recker et al. 1985, 1987), who extracted a subsample of 665 individuals from the 1976 California Department of Transportation Urban and Rural Travel Survey and applied data reduction techniques to extract pattern profiles from feature vectors that represent travel diaries of individuals, and 3) Pas and colleagues (Pas 1983, 1987, Pas & Koppelman 1987), who used data collected in the Reading Travel Survey of 1971 to augment the traditional goodness of fit value to include intrapersonal and interpersonal variation. The main methods used for travel behaviour variability analysis are summarized in Table 1.

### 2.1.1 Spatial, temporal and spatio-temporal variability in travel behaviour

Hanson & Huff (1981) proposed a repetition index that is based on five dimensions: 1) trip purpose / activity, 2) mode of travel, 3) time interval of arrival, 4) distance from last stop, and 5) location of destination. The authors then defined equivalence classes based on combinations of 3, 4 or 5 of the dimensions

Table 1: The main types of methods used for defining indexes to measure travel behaviour variability, the index definition, their disadvantage and whether they explicitly model sequences.

| Method | Index definition | Disadvantages | Seq. |
|---|---|---|---|
| Equivalence classes | Frequency of equivalence classes based on contingency table | Trips aggregated per day | No |
| Primary-secondary attributes | Similarity of attributes of trips matched based on order | Ignores trip start time and duration. | No |
| Feature vector similarity | Euclidean distance between travel diaries as feature vectors | Inconsistent travel behaviour clusters | No |
| Herfindahl-Hirschman | Frequency of identical trip spatio-temporal attributes combinations. | Subjective matching criterion. | No |
| Time budget similarity | Frequency of matched labeled activites at a 15 minute granularity | Night-time stability biases the index | No |
| Levenshtein distance | Measures amount of difference between two trips (as strings) | Subjective penalties for operations | Yes |

and used contingency tables to establish frequent patterns, i.e., repetitions of trips with similar characteristics. The authors later changed the focus of the index away from trips and towards days to measure the similarity across all days for each person (Huff & Hanson 1986). The index still made use of equivalence classes and contingency tables and it was normalized based on the number of trips contained in the comparison between days.

Pas (1983) uses a schema to define primary and secondary attributes, each of which having a complementary weight, and then compute a similarity index based on the number of matched primary and secondary attributes (a secondary attribute contributes to the index only if the primary attributes are equal) normalized by the number of trips performed during the compared days.

Recker et al. (1985, 1987) focused on reducing the dimensionality of the representation of a travel diary, which is then used to generate feature vectors representing travel diaries. The authors then proceed with computing a similarity index based on the Euclidean distance between the feature vectors, which

is also used for clustering similar patterns based on various threshold values.

Jones & Clarke (1988) have an exploratory approach to computing variability for which they divide any two compared days in equal time intervals (15 minutes) labeled with activities. They then compute a similarity index based on the number of time intervals that have the same label normalized to the number of days and trips per day. Minnen et al. (2015) continues the methodology proposed by Jones & Clarke (1988) and further refines it including concepts such as tempo (e.g., a person travels daily) and regular timing (e.g., if a person travels, she always travels at 6 am).

The reader is directed towards Schlich & Axhausen (2003) for a comparison and an evaluation of the indexes proposed by Hanson & Huff (1981, 1988), Pas (1983) and Jones & Clarke (1988) on the same dataset.

Another approach to frequency based similarity index generation is using the Herfindahl-Hirschman index that measure the repetitiveness of identical combinations of individual's spatial-activity-travel mode choices within an observed period (Susilo & Axhausen 2014, Heinen & Chatterjee 2015).

Other research focused on generating explanatory models that embed the variability of travel behaviour as opposed to specifically studying the variability of travel behaviour use different modeling techniques such as: survival analysis (Schönfelder & Axhausen 2000), structural equation models (Dharmowijoyo et al. 2016), mixed logic models (Cherchi et al. 2017) or other types of models finely tuned to fit their needs (Thøgersen 2006, Zhong et al. 2015).

The aforementioned research proposed different indexes to measure spatial and / or temporal variability that are mostly frequency based and are usually computed as the number of matched elements divided by the number of total elements that could have been matched. While the indexes offer a good overview of spatial and / or temporal variability of the travel behaviour of individuals, they do not offer insights regarding the order in which activities are performed. This paper provides index measure that account for the sequential variability / order in which activities are performed using LCS.

### 2.1.2 Sequential variability in travel behaviour

While the focus of most research was primarily on investigating variability with regards to space and / or time (Axhausen et al. 2002, Schönfelder & Axhausen 2003, Axhausen et al. 2002, Susilo & Kitamura 2005, Kitamura et al. 2006, Buliung et al. 2008, Kang & Scott 2010, Neutens et al. 2012, Dharmowijoyo et al. 2016, Cherchi et al. 2017), relatively few research efforts have been invested in studying sequential variability (Wilson 1998, Joh et al. 2001*b*,*c*, Moiseeva et al. 2014, Allahviranloo et al. 2016). Furthermore the multiple ways to measure spatial, temporal and spatio-temporal (non-sequential) variability, contrasts the sequential variability research which is linear and primarily built on top of ED.

Wilson (1998) proposed the usage of ED to measure similarities between activities extracted from travel diaries. Joh et al. (2001*b*,*c*) followed on the research of Wilson (1998) and also proposed generating alphabets that embed multiple dimensions. The end result used the same ED metric, which was

also used by Moiseeva et al. (2014). While Wilson (1998) is more preoccupied with understanding the implications of using sequence alignments and its error measures to have a better grasp of sequential variability of individuals, the more recent studies have taken the methodology as given and focused on optimizing sequence alignment methods (Joh et al. 2001*a*, Kwan et al. 2014) or using a given sequence alignment method to generate coefficients that are subsequently used to cluster users (Joh et al. 2001*b*,*c*, 2002).

While the research progressed in this direction, it is worth pointing out some of the limitations of ED. Using ED is accompanied by the subjective choice of penalties for each of the ED operations, which is a sensitive operation that heavily influences the index and biases the output. Similarly, ED based indexes are non-unique, where an index value can be obtained by any combination of ED operation penalties, so there is no clear indication to what stability is, as the ED is a penalty based method.

This paper continues the initial work of Wilson (1998) and proposes an index to measure sequential stability that is easy to understand and generate. The proposed index is based on the widely used and accepted methodology named LCS extraction, which, in this case, extracts the activities that occur in the same order in between two compared entities (e.g., comparing the activity schedules of a user for two different days, or comparing the activity schedules of two users for the same day). While ED is a penalty based method, LCS is a sequence based metric that extracts the parts that are common between sequences, which has the advantage of extracting the activities that are stable between days and not just computing penalties. A thorough discussion on the difference between ED and LCS is provided in Section 3.2.

## 2.2 Applications of the Longest Common Subsequence

The need of identifying sequential patterns in different processes and phenomena led scientist to develop multiple methods for extracting common subsequences from within a set of sequences. LCS is one such algorithm that identifies the longest subsequence that is common to a set of sequences and it has been widely applied in different fields with successful outcomes. While initial forms of the LCS algorithms have been proposed to identify similarities in the amino acid properties of two proteins (Needleman & Wunsch 1970) and shortly after that for differential file comparison (Hunt & MacIlroy 1976), the proposed concept is still widely used in current research in various fields. For example, Lin & Och (2004) use LCS to measure sentence-to-sentence similarity between a candidate translation and a set of reference translations, Banerjee & Ghosh (2001) cluster web-users based on a function of the LCS of clickstreams that is based on the trajectory users took through a website and the time spent at each page, Guo & Siegelmann (2004) propose new methods for querying a database for songs based on users humming based on a version of the LCS that eliminates errors involving rhythmic distortions, and Frolova et al. (2013) propose a technique to dynamically identify the free-air hand gestures with a set of predefined gestures in a database based on LCS.

The aforementioned tasks are similar in nature to the analysis of travel behaviour from a sequential aspect, where the task is to extract which are the activities that occur in the same order for large periods of time, common sequences of travel mode usage, etc. As such, this paper proposes a new methodology based on LCS to investigate the sequential aspect of travel behaviour by proposing new index methods and by illustrating their use through a case study.

# 3 Methodology

This section contains a preliminaries section with the necessary conventions needed to follow the remainder of the paper, a brief comparison between ED and LCS in the frame of transportation science, and the proposed index measures. While it is common to talk about travel behaviour variability, because LCS extracts stable patterns, the proposed indexes are a measure of stability, as opposed to variability, and they are split in two main categories: *inter-personal* index measures, which indicate the stability of the collected user base (the indexes are based on the comparisons between every two users), and *intra-personal* index measures, which indicate the degree of sequential stability (the indexes are based on the comparison for the same user between multiple days).

## 3.1 Preliminaries

This section provides the necessary preliminaries, formalism, and terminology used in the remainder of this paper.

Let $P = \{P_1, P_2, ..., P_n\}$ be an exhaustive set of activity purposes that constitute the purpose schema of a collected travel diary (automatic or otherwise).

Let $p = \{p_1, p_2, ..., p_n\}$ be a set of single digit letters that constitute the alphabet needed to represent the purpose schema, and let $f_p : P \rightarrow p$ be the isomorphism used as a map function between the purpose set $P$ and the equivalent alphabet set $p$, with its inverse $g_P : p \rightarrow P$.

Similarly, let $M = \{M_1, M_2, ..., M_k\}$ be an exhaustive sets of travel modes that constitute the travel mode schema of a collected travel diary (automatic or otherwise).

Let $m = \{m_1, m_2, ..., m_k\}$ be a set of single digit letters that constitute the alphabet needed to represent the travel mode schema, and let $f_m : M \rightarrow m$ be the morphism used to map the travel mode set $M$ to its alphabet set equivalent $m$, with its inverse $g_M : m \rightarrow M$.

As this paper does not explicitly deal with the spatio-temporal dimensions of traveling and is strictly focusing on sequence, let $D^I$ represent a given day $I$ defined as a sequence of the trips performed during the day $< t_1, t_2, ..., t_n >$, and $d^I$ represents its alphabet, which is obtained by mapping the trips' purposes to alphabet letters $< p_1, p_2, ..., p_l >$.

By using the proposed notation, the LCS between two days, $d^I$ and $d^{II}$,

where $d_i^I$ denotes the $i$th letter of $d^I$, can be defined according to Equation 1:

$$LCS(d_i^I, d_j^{II}) = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0. \\ LCS(d_{i-1}^I, d_{j-1}^{II}) \cup d_i^I, & \text{if } d_i^I = d_j^{II}. \\ longest(LCS(d_i^I, d_{j-1}^{II}), LCS(d_{i-1}^I, d_j^{II})), & \text{if } d_i^I \neq d_j^{II}. \end{cases} \tag{1}$$

Similarly, using the same notations for two compared days, $d^I$ and $d^{II}$, the ED between days $d^I$ and $d^{II}$ is defined according to Equation 2:

$$ED(d_i^I, d_j^{II}) = \begin{cases} max(i,j), & \text{min(i,j)=0.} \\ min \begin{cases} ED(d_{i-1}^I, d_j^{II}) + cost_{del} \\ ED(d_i^I, d_{j-1}^{II}) + cost_{ins} \\ ED(d_{i-1}^I, d_{j-1}^{II}) + I_{d_i^I \neq d_j^{II}} * cost_{repl} \end{cases} & \text{otherwise} \end{cases} \tag{2}$$

The equivalent of Equations 1 and 2 holds for the case when $D^I$ represent a day $I$ defined as a sequence of triplegs $< tl_1, tl_2, ..., tl_m >$, and $d^I$ represents its alphabet, which is a sequence of mapped trip travel modes $< m_1, m_2, ..., m_k >$, or for any representation of a day as a sequence represented as alphabet letters.

### 3.2 A comparison between Longest Common Subsequence and Edit Distance

To better understand how LCS and ED differ from one another, it is important to compare why and how these two metrics are used. ED is used as a measure of similarity between two strings, since it returns the cost as the minimum number of edits (inserting, deleting and shifting) one has to perform to transform one string into another. On the other hand, LCS is used to extract the longest subsequence that is common to a set of (usually two) sequences.

Bluntly put, ED identifies the parts of a string that have to be modified to make it equal to the string it is compared with, and LCS identifies the parts that are already sequentially common for both strings. As such, the result of applying ED between two strings is a **number** that indicates the transformation cost / similarity measure, and the result of applying the LCS method on two strings is a **subset of the letters** that are common to both strings and appear in the same order in both strings, i.e., a subsequence.

Another difference between the two methods is the intuitive explanation associated with them in the context of activities. Whereas LCS has a simple explanation, i.e., it extracts the activities that were performed in the same order for both compared entities (e.g., days, users), the ED explanation is more complicated, i.e., it represents the minimum transformation cost between the two entities.

While there is clear potential in using ED to compute the similarity of travel behaviour exhibited between days, especially when the cost measures shift from units to spatio-temporal errors as proposed by Prelipcean et al. (2016), it serves to a different purpose from LCS: LCS extracts stable patterns of how people

Table 2: The main differences between ED and LCS.

| | Edit Distance | LCS |
|---|---|---|
| Input | S1, S2 | S1, S2 |
| Operations | Delete, Insert, Substitute | Delete, Insert |
| Main purpose | String comparison | Ordered elements extraction |
| Provides answer to | How many edits are needed to turn a string into another? | Which are the elements that occur in same order in both strings? |
| Result type | Sum of penalties | Subsequence of elements |
| TSc usage | Measure similarity | Extract stable patterns |

travel, ED measures how similar two patterns are. Although on a first look, these measures might seem that they are similar, they are different along multiple dimensions as shown in Table 2.

Consider the following case of two hypothetical days $D^I$ and $D^{II}$. During the first day, $D^I$, a traveler performed the following activities in the given order {Spent time at home, Worked, Ate at restaurant, Worked, Ate at restaurant, Did groceries, Spent time at home}. Similarly, during the second day, $D^{II}$, a traveler performed the following activities in the given order {Spent time at home, Worked, Ate at restaurant, Worked, Spent time at home, Worked}. Using the isomorphic mapping between activities and alphabet letters $f_p$ described in Section 3.1, the alphabets of the two days can be generated: $d^I = \{\text{HWRWRGH}\}$ and $d^{II} = \{\text{HWRWHW}\}$, where $f_p$ is given by Equation 3.

$$f_p = \{\text{Home} \rightarrow \text{H}, \text{Worked} \rightarrow \text{W}, \text{Restaurant} \rightarrow \text{R}, \text{Groceries} \rightarrow \text{G}\} \quad (3)$$

Applying ED on the alphabets of the two days $d^I$ and $d^{II}$, one obtains different values based on the cost assumption, as shown in Table 3. As such, ED is a good option for measuring similarity when different operations should be weighted differently (e.g., putting a lower cost on replacements when the emphasis is put on an algorithm's ability to detect the correct number of activities as opposed to detecting the correct activities).

Contrary to ED, applying LCS on the alphabets of the two days $d^I$ and $d^{II}$, one obtains a single result as there are no costs implied, i.e., $LCS(d_i^I, d_j^{II}) = \{\text{HWRWH}\}$. The result is also commutative, i.e., $LCS(d_i^I, d_j^{II}) = LCS(d_i^{II}, d_j^I)$, which is a property that makes the comparison of two days of activities independent of the order in which the days are compared. This property is useful when one wants to identify the activities that happen in the same order in either days, which is relevant for exploratory analysis.

Table 3: The difference of ED values for the strings $d^I$ = {HWRWRGH} and $d^{II}$ = {HWRWHW} based on the costs chosen for the deletion ($cost_{del}$), insertion ($cost_{ins}$) and replacement ($cost_{repl}$) operations. The only case of symmetric ED is obtained only when the following relationship between costs holds: $cost_{del}$=$cost_{ins}$=$2 * cost_{repl}$.

| | ED($d^I$,$d^{II}$) | ED($d^{II}$,$d^I$) |
|---|---|---|
| $cost_{del}$=1<br>$cost_{ins}$=1<br>$cost_{repl}$=2 | 3<br>(2*$cost_{del}$ + $cost_{ins}$) | 3<br>(2*$cost_{ins}$+$cost_{del}$) |
| $cost_{del}$=1<br>$cost_{ins}$=2<br>$cost_{repl}$=3 | 4<br>(2*$cost_{del}$+$cost_{ins}$) | 5<br>(2*$cost_{ins}$+$cost_{del}$) |
| $cost_{del}$=3<br>$cost_{ins}$=1<br>$cost_{repl}$=1 | 5<br>(2*$cost_{repl}$+$cost_{del}$) | 3<br>(2*$cost_{repl}$+$cost_{ins}$) |

There is the innate relationship in between LCS and ED shown in Equation 4, for the particular case of ED when the cost of insertion is equal to the cost of deletion, the cost is equal to 1, and the only type of substitution is done by deleting the mismatched character and inserting the missing character.

$$ED(d_i^I, d_j^{II}) = |d_i^I| + |d_j^{II}| - 2 * |LCS(d_i^I, d_j^{II})| \tag{4}$$

To summarize, both ED and LCS have their merits and there is a relationship in between the two. However, when the focus of the analysis is identifying activities and not solely on looking at indexes reflective of similarity, LCS is the only method capable of extracting those activities.

### 3.3 Measuring stability with LCS

This paper defines individual pattern stability by extracting patterns, which are subsequences of generic schedules, at given unit of time intervals (which are usually of one day or for the duration of a particular type of trip) and comparing the length of the subsequences with the length of the schedule. The generic formula for these types of indexes, which if further expanded for specific cases in this section, is given by Equation 5, where the length of LCS between the schedules $sch$ of users $u_k$ and $u_l$ on the time units $d_i$ and $d_j$ is divided by the length of the schedule of user $u_k$ at time unit $d_i$. The $P_{u_k,u_l}^{d_i,d_j}$ index represents the percentage of elements of the schedule $sch$ on the time unit $d_i$ that user $u_k$ performs in the same order as the schedule $sch$ of user $u_l$ on the time unit $d_j$.

$$P_{u_k,u_l}^{d_i,d_j} = |LCS(sch_{u_k}^{d_i}, sch_{u_l}^{d_j})|/|sch_{u_k}^{d_i}| \tag{5}$$

This generic explanation can then be easily put in the frame of travel behaviour, where $u_i$ and $u_k$ represent two users of a given user base, $d_i$ and $d_j$ can represent two different days, and the schedule $sch$ can be a sequence of activities / purposes, travel modes or any other type of entity that has an expected sequential behaviour, which is obtained via the isomorphic mapping function shown in Section 3.1. The further sections expand on Equation 5 to define *inter-personal* indexes that describe the stability of schedules by comparing different users for the same time unit, and *intra-personal* indexes that describe the stability of schedules for the same users for different time units.

It is important to note that the proposed index is not commutative, i.e., $P^{d_i,d_j}_{u_k,u_l} \neq P^{d_j,d_i}_{u_k,u_l}$ and $P^{d_i,d_j}_{u_k,u_l} \neq P^{d_i,d_j}_{u_k,u_l}$, due to the division to the length of the schedule (Equation 5). This is an important observation because it allows for the logical relationship described in Equation 6, which allows for the same index to be read both as $sch^{d_i}_{u_k}$ can **be explained** by $sch^{d_j}_{u_l}$ as well as $sch^{d_j}_{u_l}$ can **explain** $sch^{d_i}_{u_k}$. This implies that a schedule can have two properties: the ability of being explained by other schedules and the ability to explain other schedules, which is relevant when one wants to allocate differential weights based on schedule length.

$$High\ P^{d_i,d_j}_{u_k,u_l} \nRightarrow High\ P^{d_j,d_i}_{u_k,u_l} \nRightarrow High\ P^{d_i,d_j}_{u_l,u_k} \tag{6}$$

As a side note, while it is possible to have set comparison instead of element comparison (e.g., having a set of users $U$ instead of two users $u_k$ and $u_l$), due to the nature of LCS, the result is biased towards users that have a shorter schedule (e.g., a user staying home all day), which is the reason why this paper employs a paired schedule comparison instead of the full set of schedules comparison.

### 3.3.1 Measuring inter-personal stability with LCS

The first type of proposed index, measures the degree of stability in a given user base and is an indicator for how much variability can one explore in a given data set. While these measures can apply to multiple attributes or dimensions that have a sequential aspect, this section only proposes two: inter-personal purpose schedule stability and inter-personal travel mode schedule stability.

These methods compare the schedules of every two users within the unit of expected sequential variability (which is a day in most cases) and can be obtained as a special case of Equation 5. As such, considering the sequential variability unit of a day $d$, its schedule $sch^d$ and two given users $u_i$ and $u_j$, each of which having the schedule $sch^d_{u_i}$, and $sch^d_{u_j}$, respectively, the general formula for computing the inter-personal stability (OP) between $u_i$ and $u_j$ is given by Equation 7:

$$OP^d_{u_i,u_j} = |LCS(sch^d_{u_i}, sch^d_{u_j})|/|sch^d_{u_i}| \tag{7}$$

The relation between the definition of the general stability index and the proposed inter-personal stability index is given by Equation 8. In this case, the unit of time is constant and the comparison is done across different users.

$$OP^{d}_{u_l,u_k} = P^{d_i=d_j=d}_{u_k,u_l}, \forall k \neq l \tag{8}$$

The index value defined in Equation 7 signifies the percentage of elements from $sch^{d}_{u_i}$ that are present in the same order in $sch^{d}_{u_j}$. A high index value indicates that most elements in $sch^{d}_{u_i}$ are performed in the same order as in $sch^{d}_{u_j}$, whereas a low index value indicates that either most elements in $sch^{d}_{u_i}$ are performed in a drastically different order than in $sch^{d}_{u_j}$, or that $sch^{d}_{u_i}$ contain elements that are not present in $sch^{d}_{u_i}$. One way to avoid the uncertainty of low index values is to perform subset comparison operations to identify if different activities are performed in both days (Prelipcean et al. 2015).

Equation 7 can be rewritten for the case when the schedule $sch$ represents a sequence of trips $t$ as shown in Equation 9, or as a sequence of triplegs $tl$ as shown in Equation 10. These measures can help scientists explore and test new hypothesis regarding the sequential stability of the whole user base with regards to day of week or any other time unit that is expected to exhibit noticeable seasonality.

$$OPP^{d}_{u_i,u_j} = |LCS(t^{d}_{u_i}, t^{d}_{u_j})|/|t^{d}_{u_i}| \tag{9}$$

$$OPT^{d}_{u_i,u_j} = |LCS(tl^{d}_{u_i}, tl^{d}_{u_j})|/|tl^{d}_{u_i}| \tag{10}$$

### 3.3.2 Measuring intra-personal stability with LCS

Contrasting the inter-personal stability indexes (IP) where the comparison is done between users in the considered unit of time and gives an indication regarding the degree of stability in a user base, the intra-personal stability measures are intrinsic and give an indication regarding the degree of stability for a given user. Considering a user $u$ who has performed activities for two days $d_i$ and $d_j$, with schedules $sch^{d_i}_{u}$ and $sch^{d_j}_{u}$, the general formula for computing the intra-personal stability between days $d_i$ and $d_j$ is given by Equation 11.

$$IP^{d_i,d_j}_{u} = |LCS(sch^{d_u}_{u}, sch^{d_j}_{u})|/|sch^{d_i}_{u}| \tag{11}$$

The relation between the definition of the general stability index and the proposed intra-personal stability index is given by Equation 12. In this case, the user is constant and the comparison is done across different units of time.

$$IP^{d_i,d_j}_{u} = P^{d_i,d_j}_{u_k=u_l=u}, \forall i \neq j \tag{12}$$

Similarly to previous index measures, the equivalent equations for intra-personal activity stability index (Equation 13) and intra-personal travel mode stability index (Equation 14) can be defined to measure the percentage of activities, and travel modes, that are performed in the same order in day $d_i$ as in $d_j$ by user $u$.

$$IPP^{d_i,d_j}_{u} = |LCS(t^{d_u}_{u}, t^{d_j}_{u})|/|t^{d_i}_{u}| \tag{13}$$

$$IPT_u^{d_i,d_j} = |LCS(tl_u^{d_u}, tl_u^{d_j})|/|tl_u^{d_i}| \quad (14)$$

The proposed index measures can be aggregated for the whole user base to describe the expected stability at the user level, which still contrasts the inter-personal index measures since the aggregation unit is daily stability for each user in the case of intra-personal aggregation and daily stability between every two users in the case of inter-personal aggregation.

## 3.4 Extensibility of proposed measures

If one wants to augment the proposed indexes, it is a straightforward process, since the indexes are easily extensible to cover any type of attribute combination that involves sequences. For example, consider the case of travel mode sequence stability when traveling for a particular purpose. In this case, the unit of time for the stability comparison switches from day to trip.

For this task, one can rewrite Equations 9 and 10 from an inter-personal perspective to measure the travel mode sequence stability between all trips with the same purpose as performed by different users. In this case, the schedule is represented by the sequence of mapped triplegs $tl$ that users $u_k$ and $u_l$ employ on trips $t_i$ and $t_j$ with the same purpose $p$, as shown in Equation 15.

$$OPPT_{u_k,u_l}^{t_i^p,t_j^p} = |LCS(tl_{u_k}^{t_i^p}, tl_{u_l}^{t_j^p})|/|tl_{u_k}^{t_i^p}| \quad (15)$$

Similarly, one can rewrite Equations 13 and 14 from an intra-personal perspective to measure the travel mode sequence stability between all trips with the same purpose for one user only. In this case, the schedule is represented by the sequence of mapped triplegs $tl$ that user $u$ employs on trips $t_i$ and $t_j$ with the same purpose $p$, as shown in Equation 16.

$$IPPT_u^{t_i^p,t_j^p} = |LCS(tl_u^{t_i^p}, tl_u^{t_j^p})|/|tl_u^{t_i^p}| \quad (16)$$

While the indexes are easily extensible, one has to pay attention to what they represent to prevent explanation obscurity. In this case, Equation 15 measures the percentage of travel modes that users $u_k$ performs in the same order as user $u_l$ when traveling for the same purpose $p$ trips $t_i$ and $t_j$. Similarly, Equation 16 measures the percentage of travel modes that user $u$ performs in the same order when traveling for the same purpose $p$ trips $t_i$ and $t_j$. In the case of sequences, it is key to propose indexes that can be easily understood to aid scientists in further exploring ideas among these directions or in forming hypotheses regarding the usefulness and representativeness of these indexes.

Finally, the authors want to emphasize the key aspect of using LCS for measuring pattern stability. While the proposed indexes in this section only make use of the length of the LCS, the subsequences are still available for further analysis and exploration (as shown further in Section 4). This contrasts the existing methodology behind sequential variability analysis that define similarity based on penalty cost for aligning sequences.

Table 4: Isomorphic mapping between purposes and their alphabet, and between travel modes and their alphabet

(a) Alphabet mapping for purposes

| Letter | Purpose |
|---|---|
| B | Business |
| F | Grocery Shopping |
| O | Hobby |
| H | Home |
| L | Leisure |
| z | Other |
| N | Other Shopping |
| P | Personal |
| D | Pickup / Dropoff |
| R | Restaurant/Caf |
| S | School |
| V | Visit |
| W | Work |

(b) Alphabet mapping for travel modes

| Letter | Travel Mode |
|---|---|
| B | Bicycle |
| S | Bus |
| D | Car as driver |
| P | Car as passenger |
| Y | Commuter train |
| Z | Ferryboat |
| O | Flight |
| M | Moped / Motorcycle |
| T | Subway |
| z | Taxi |
| F | Train |
| X | Tram |
| W | Walk |

# 4 Case study

The present case study is an example of applying the proposed methodology to better understand an available dataset. The authors do not provide any inferences / models any phenomenon in this case study, the purpose of the case study is to exemplify the use of the suggested methodology by complementing the descriptive statistics that stand at the basics of most hypotheses.

## 4.1 Data description

The original dataset contains 2142 trips from 171 users gathered in Stockholm, Sweden between 2nd and 9th of November 2015 with MEILI, a semi-automated travel diary collection system (Prelipcean et al. 2018). The authors filtered out all users who did not collect data for a period of at least one week, which mainly consisted of eliminating trips of users with an early dropoff rate. The remainder of the dataset that is used in this case study consists of 1250 trips collected from 51 users for a period of at least one week. The purpose and travel modes schema used for collecting data consist of 13 different purposes and 14 different travel modes, which are summarized in Table 4 together with the isomorphic mapping between purposes / travel modes and their associated alphabet letters.

The MEILI travel diary collection system automatically infers travel modes and purposes and presents to the users a pre-ordered list of choices for travel modes and purposes in a web annotation user interface, where users can correct any inference mistakes.

## 4.2 Preparing data for the LCS stability analysis

As mentioned previously, the first step the authors took was to filter out the trips of users who did not collect data for at least one week because the proposed analysis relies on having data for the period of a week.

The second step was to define the isomorphic mapping functions needed to generate a purpose alphabet, as well as a travel mode alphabet. Since the function is very simple, a one to one association table was build for trip purposes and travel modes each, which contain the information both for the direct and inverse alphabet mapping.

These are the only two steps needed in the pre-processing stages to apply the proposed methodology.

## 4.3 Inter-personal indexes

This section mainly discusses the applications of inter-personal purpose index (Equation 9) and travel mode index (Equation 10).

### 4.3.1 Inter-personal purpose index

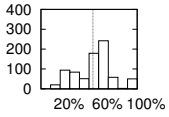
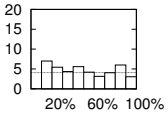
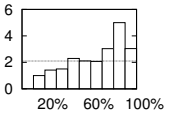
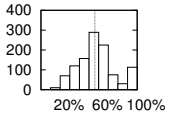
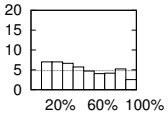
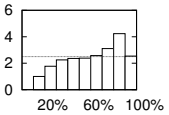
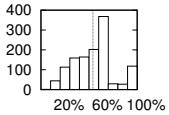
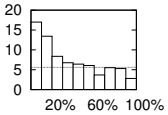
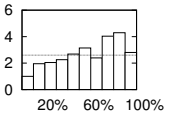
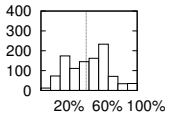
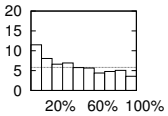
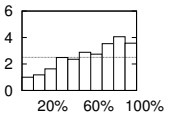
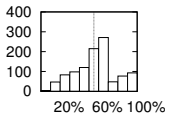
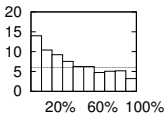
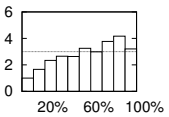
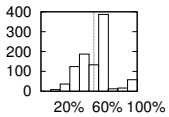
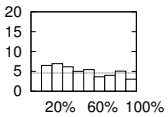
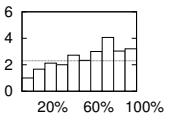
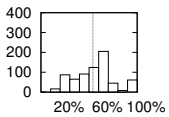
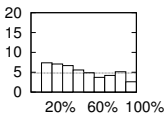
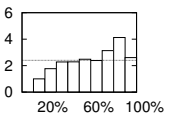
As mentioned in Section 3.3.1, the inter-personal purpose index is a metric applied on the whole user base. The method extracts the LCS values and computes the index between every two users in the user base for the same day (as shown in Equation 9). Table 5 presents an overview of the aggregated index value obtained for each day using this method.

Table 5 shows that the average degree of similarity between how users scheduled their activities is around 50% for every day, which implies that about half of the activities performed by a user are performed, on average, in the same order by any other user. While this is a good metric to start the exploration of how users schedule their days with regards to activity, it is by no mean complete. To further examine the similarity between users, Table 5 also contains the length of the schedule as well as the LCS, which shows that the average number of activities each user performs every day is between 5 and 6 and the number of activities that are performed in the same order by another user is on average between 2 and 3.

To get a more complete picture of the similarity, Table 5 also contains the distributions of the index value with regards to the number of user pairs (column 5), average length of schedule (column 6) and average length of LCS (column 7). For all days, the highest frequency index value is between 60% and 70%, the exception being Tuesday where the highest frequency index value is between 50% and 60%. If one compares the average value of Tuesday and Thursday, in which case most index values for Thursday are between 60% and 70%, the frequency of low value indexes (under 50%) brings the average to 47%, which is lower than the average for Tuesday, i.e., 56%, even though most index values for Tuesday are inside a lower distribution bucket (between 50% and 60%).

When analyzing the distribution of the index value as compared to the schedule length as well as the LCS length, it is interesting to note that the average

Table 5: Inter-personal analysis for purpose, where "Idx" represents the inter-personal purpose index value given by Equation 9, "Sch length" represents the average length of an *activity schedule* for the given day, "LCS length" represents the average length of the *activity LCSes* extracted for the given day. The table also contains the distribution of the index plotted against the number of compared user pairs in "Idx freq", the average schedule length "Sch length distribution" and the average LCS length "LCS length distribution". The dashed lines in the distribution represent the average values in the first columns.

| Day | Idx | Sch length | LCS length | Idx freq | Sch length distribution | LCS length distribution |
|---|---|---|---|---|---|---|
| Mo. | 54% | 4.1 | 2.4 |  |  |  |
| Tu. | 56% | 4.8 | 2.1 |  |  |  |
| We. | 54% | 5.6 | 2.5 |  |  |  |
| Th. | 47% | 5.8 | 2.6 |  |  |  |
| Fr. | 55% | 6 | 2.5 |  |  |  |
| Sa. | 55% | 4.6 | 3 |  |  |  |
| Su. | 54% | 4.9 | 2.3 |  |  |  |

value for index between 90% and 100% is around 3 activities, which implies that the matched schedules are particularly short. In general, the distribution of the schedule length with regards to the index value has a specific trend of descending length towards index value at around 70%, followed by an ascent to 90%, followed by another descent for 100%. While the correlation between a decrease in length and an increase in index is logical, it is also affected by the regression to the mean phenomenon since the index values lower than 20% occur seldom in the user base. This phenomenon is most visible for Wednesday, Thursday and Friday, where very few index values between 0% and 10% introduce a spike in the length of the schedule (more than 10 activities).

Finally, the LCS length analysis reveals an expected trend: an increase in the LCS length is usually correlated with an increase in the index value until 90%. Then there is a slight decrease from 90% to 100%, which is mostly due that the schedules that have the same activities in the same order tend to be shorter. The only irregularity in this case is Saturday, where the LCS length decrease starts at 80% instead of 90%, which can be explained by the low number of 70%-80% index values, which again points at the failure to regress to the mean due to lack of data.

### 4.3.2 Inter-personal travel mode index

As briefly mentioned in Section 3.3.1, the inter-personal travel mode index is a metric applied on the whole user base. The method extracts the LCS values and computes the index between every two users in the user base for the same day (as shown in Equation 10). Table 6 presents an overview of the aggregated index value obtained for each day using this method.

Contrasting the inter-personal purpose index similarity, the inter-personal travel mode index similarity has most values between 0% and 10% for all days, and an average between 30% and 40% for all days. This implies that while the user base perform on average around half of the activities in the same order, the used travel modes en-route to the activities differ more than the activities they perform. Another interesting aspect is the length of the travel mode schedule is greater than the length of activity schedule, which captures multi-travel mode trips, with the largest length differences on Tuesday, Wednesday and Friday. This implies that on those days, it is more common for users to travel with multiple travel modes while performing a trip. Another interesting finding is that while the users have a schedule on Thursday with as many activities or more than Tuesday, Wednesday and Friday, the difference between the purpose schedule's length and the travel mode schedule's length is small, which implies that even though the users perform more activities than on average, they seldom use travel mode chains for Thursday. This might mean that the users are more susceptible to using non-chainable travel modes (such as driving) when performing Thursday activities.

The distributions of schedule length and LCS length are as expected, the travel mode schedule length distribution is relatively flat, with a slight decrease after 50% and a strong decrease after 90%. The LCS length follows an interesting

Table 6: Inter-personal analysis for travel mode, where "Idx" represents the inter-personal travel mode index value given by Equation 10, "Sch length" represents the average length of a *travel mode schedule* for the given day, "LCS length" represents the average length of the *travel mode LCSes* for the given day. The table contains the distribution of the index plotted against the number of compared user pairs in "Idx freq", the average schedule length "Sch length distribution" and the average LCS length "LCS length distribution". The dashed lines in the distribution represent the average values in the first columns.

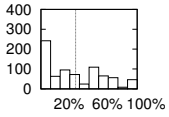
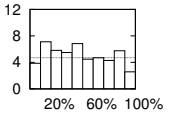
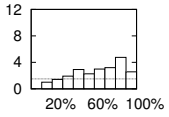
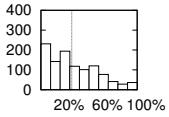
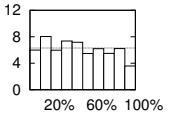
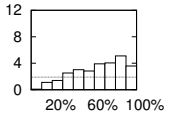
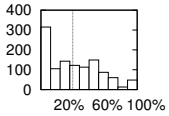
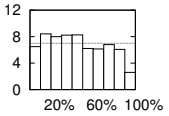
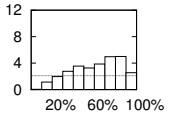
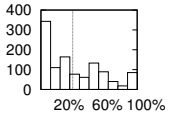
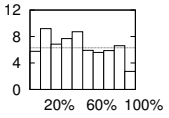
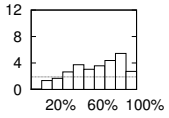
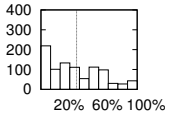
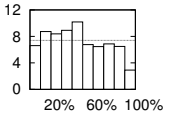
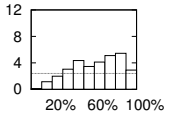
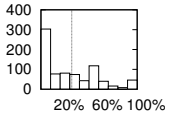
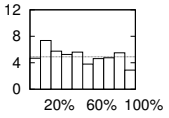
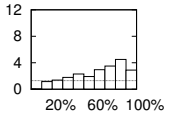
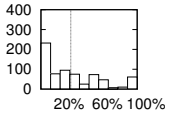
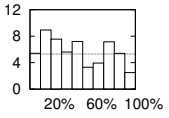
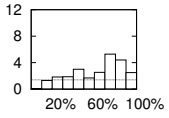
| Day | Idx | Sch length | LCS length | Idx freq | Sch length distribution | LCS length distribution |
|---|---|---|---|---|---|---|
| Mo. | 36% | 4.7 | 1.5 |  |  |  |
| Tu. | 32% | 6.3 | 1.9 |  |  |  |
| We. | 33% | 7 | 2.1 |  |  |  |
| Th. | 33% | 6.3 | 1.9 |  |  |  |
| Fr. | 37% | 7.4 | 2.4 |  |  |  |
| Sa. | 32% | 4.9 | 1.3 |  |  |  |
| Su. | 31% | 5.6 | 1.4 |  |  |  |

Table 7: The intra-personal purpose index given by Equation 13 plotted for every day of week combination. The vertical column summarizes the ability of the day in the table header of *being explained*: the percentage of activities that are performed in the day in the header in the same order as in the days on the side of the table. The horizontal column summarizes the ability of the day on the side of the table of *explaining* other days: the percentage of activities that are performed by the days in the header of the table in the same order as in the day on the side of the table. Friday can explain a high percentage of the activity sequences that occur during the other days, but no other day can explain a high percentage of the activity sequences that occur during Friday.

| | Mo. | Tu. | We. | Th. | Fr. | Sa. | Su. |
|---|---|---|---|---|---|---|---|
| Mo. | N/A | 54.4% | 49.2% | 54% | **51.3%** | 53.8% | 59.7% |
| Tu. | 68.7% | N/A | 60.4% | 65.4% | **53.9%** | 55.2% | 62.9% |
| We. | 72.1% | 66.2% | N/A | 67.5% | **60.4%** | 63.4% | 62.7% |
| Th. | 65.6% | 67.1% | 60.9% | N/A | **57.8%** | 53.5% | 56.9% |
| Fr. | **72.4%** | **65.6%** | **63.4%** | **70.7%** | N/A | **56.7%** | **63.2%** |
| Sa. | 70.1% | 56% | 55.3% | 58.3% | **49.3%** | N/A | 63.4% |
| Su. | 69.7% | 58% | 51.4% | 56.6% | **50.1%** | 58.3% | N/A |

pattern, with a relatively flat distribution and a slight increase after 50% and a strong decrease after 90%. The observed decrease / increased of the mention distribution can be due to the regression of the mean phenomenon since there are few index value user pairs with a value between 60% and 90%. The 90% to 100% decrease is again linked to the very short travel mode schedules, with an average length of 3 travel modes.

## 4.4 Intra-personal indexes

This section discusses the applications of the intra-personal purpose index (Equation 13) and travel mode index (Equation 14).

### 4.4.1 Intra-personal purpose index

As mentioned in Section 3.3.2, the intra-personal purpose index is an intrinsic metric that shows the degree of stability for the same user across different days. The method extract the LCS values and computes the index between every two days for the same user with regards to purposes. Table 7 presents an overview for the aggregated index values obtained between every two days for every user.

The columns of Table 7 represent the percentage of activities performed on the given day (top label) that are performed in the same order as in the corresponding day (left label). As such, 68.7% of the activities performed by an user on Monday are, on average, performed in the same order on Tuesday,

and 72.1% of the activities performed by an user on Monday are, on average, performed in the same order on Wednesday. When read vertically, the table gives a good overview of how "explainable" a day is in terms of activity schedule ordering.

By analogy, the rows of Table 7 represent the percentage of activities performed in the corresponding day (top label) in the same order as in the given day (side label). As such, 54.4% of the activities performed by an user on Tuesday are, on average, performed in the same order on Monday, and 49.2% of the activities performed by an user on Wednesday are, on average, performed in the same order on Monday. When read horizontally, the table gives a good overview of how much "explanation power" a day has in terms of activity schedule ordering.

Considering the "explainable" and "explanation power" concepts, one can investigate Table 7 to identify which days have a highly regular pattern as well as a highly distinctive pattern. By analyzing the table vertically, the workdays have a highly regular pattern, with index values between 60% and 73%. On average, users have a more irregular pattern on Wednesday and Friday, since the "explainable" ability of Wednesday and Friday is low when compared to other weekdays. Conversely, these days have a high "explanation power", since the rows in Table 7 for Friday and Wednesday have consistently larger values than other rows. At the same time, Monday has the lowest "explanation power", which is expected since, as shown in Table 5, Monday contains, on average, the lowest number of trips. Similarly, Wednesday and Friday have a higher number of trips, which can assess to their "explanation power" as explained in this section. One interesting finding is that although Thursday has a high number of trips (Table 5), it has a low "explanation power", which is consistent with the previous observation that users use non-chainable travel modes on Thursday, which can suggest highly specific trips that have a degree of seasonality that is longer than one week and it is not captured in the available dataset. Another surprising finding is that users make, on average, over 60% of Sunday's activities in the same order on Tuesday, Wednesday, Friday and Saturday, which can suggest that the days with the most irregular patterns are Wednesday, Friday and Saturday.

### 4.4.2 Travel mode

As mentioned in Section 3.3.2, the intra-personal travel mode index is an intrinsic metric that shows the degree of stability for the same user across different days. The method extract the LCS values and computes the index between every two days for the same user with regards to travel modes, which is shown in Table 8.

An observation that is inline with the inter-personal index analysis, the travel modes schedules are more irregular than the purpose schedules. One pattern that can be easily observed in Table 8 is the high index values between any working day combination, the high index values between any weekend day combination, and the low index value for any combination between working days and weekend days. While Wednesday and Friday still have a high "explanation

Table 8: The intra-personal travel mode index given by Equation 14 plotted for every day of week combination. The vertical column summarizes the ability of the day in the table header of *being explained*: the percentage of travel modes that are used in the day in the header in the same order as in the days on the side of the table. The horizontal column summarizes the ability of the day on the side of the table of *explaining* other days: the percentage of travel modes that are used during the days in the header of the table in the same order as in the day on the side of the table. Users employ patterns during weekend days that are dissimilar to weekdays and similar to other weekend days.

| | Mo. | Tu. | We. | Th. | Fr. | Sa. | Su. |
|---|---|---|---|---|---|---|---|
| Mo. | N/A | **52.5%** | **47.3%** | **46.6%** | **46.5%** | 46.4% | 43.7% |
| Tu. | **58.4%** | N/A | **54.8%** | **61.4%** | **54%** | 49.6% | 44% |
| We. | **59.4%** | **53.7%** | N/A | **60.7%** | **51.7%** | 50.8% | 50.6% |
| Th. | **57.2%** | **56.7%** | **55.2%** | N/A | **50.4%** | 47% | 39.9% |
| Fr. | **58.1%** | **60.9%** | **55.7%** | **59.9%** | N/A | 55.1% | 45% |
| Sa. | 39.1% | 31.6% | 36.7% | 37.5% | 38.1% | N/A | 54.2% |
| Su. | 40.1% | 41.4% | 41.9% | 40.1% | 37.5% | 56% | N/A |

power", Monday can only be easily explained by other working days, and not by weekend days. This suggests that even if the users might do same activities in the same order on Monday and Sunday, they travel differently when performing them.

## 4.5 Comparing inter-personal and intra-personal index for the travel modes of same purpose trips

As this paper claims that the indexes can be easily extended for any sequential phenomenon (Section 3.4), this section analyzes the similarity of how users travel when performing same purpose trips. The analysis compares the inter-personal (Equation 15) and intra-personal (Equation 16) index values side by side in Table 9.

As expected, the inter-personal index values are lower than the intra-personal index value due to the various travel modes available for each user as well as due to the different opportunities each user has when traveling for a given purpose. The amount of variation for each of the indexes is roughly similar per purpose in both inter- and intra-personal settings, but the average and median values are higher for the intra-personal indexes. The highest inter-personal similarity is for travel modes while traveling for sport / hobby (44%), restaurants (41%) and non-food shopping (42%), and the highest inter-personal dissimilarity is for travel modes while traveling for visiting friends and family (18%), business (26%) and returning home (26%). Interestingly enough, the

Table 9: The inter-personal (Equation 15) and intra-personal (Equation 16) index values for the travel mode similarity of same purpose trips. As expected, there is low inter-personal similarity for most purposes (with the exception of non-food shopping, restaurant and sport trips, all of which have an index value more than 40%) while the intra-personal similarity is considerably higher with some index value of more than 60% (business, leisure, sport and school trips).

| | Inter-personal | | | Intra-personal | | |
|---|---|---|---|---|---|---|
| | Avg. | S.D. | Med. | Avg. | S.D. | Med. |
| Business travel | **26.0%** | 41.9% | 0.0% | 62.7% | 46.7% | 100.0% |
| Food/grocery shopping | 33.0% | 44.2% | 0.0% | 59.2% | 44.0% | 66.7% |
| Leisure travel | 38.4% | 41.5% | 25.0% | 62.8% | 38.6% | 50.0% |
| Non-food shopping | **41.7%** | 47.0% | 0.0% | **57.1%** | 46.2% | 100.0% |
| Other | 26.7% | 41.7% | 0.0% | 49.9% | 48.9% | 41.7% |
| Personal business | 31.3% | 41.8% | 0.0% | 35.1% | 40.5% | 25.0% |
| Pick-up / drop-off | 33.5% | 44.6% | 0.0% | 56.7% | 46.5% | 100.0% |
| Restaurant / Café | **41.1%** | 46.2% | 0.0% | 52.5% | 45.8% | 50.0% |
| Return home | **26.3%** | 39.7% | 0.0% | **41.2%** | 45.2% | 20.0% |
| Sport / hobby | **44.7%** | 47.5% | 0.0% | **78.6%** | 39.8% | 100.0% |
| School | 36.8% | 36.9% | 33.3% | 60.7% | 38.7% | 66.7% |
| Work | 33.0% | 41.4% | 0.0% | 50.9% | 44.8% | 50.0% |
| Visit relatives and friends | **18.0%** | 34.3% | 0.0% | **49.8%** | 40.9% | 50.0% |

higher than average similarity holds for intra-personal indexes in the case of sport / hobby (78.6%) and non-food shopping (57.1%), as well as the lower than average similarity holds for visiting-friends and family (49.8%) and returning home (41.2%). The travel mode patterns for business traveling has a high value for the intra-personal index and a low value for the inter-personal index. The centric activities of work and home have a low intra-personal index because they can be considered as the central parts of days, where it is common to perform a secondary activity (such as shopping) and return to the locations associated with primary activities. The reverse is true for secondary activities such as sport / hobby and leisure travel that have a high intra-personal index, which suggests that they are highly specific activities and users travel via the same modes when performing them. Finally, the differences between the intra-personal and inter-personal index values suggest that while the user group travels differently when performing same purpose trips, in most cases, the users themselves use a similar sequence of travel modes when performing the same purpose again.

It is important to note that there is no direct association between inter-personal and intra-personal indexes, i.e., a high value for an inter-personal index

Table 10: Percentage of instances each purpose is present in the LCS.

| | Inter-personal (%) | Intra-personal (%) |
|---|---|---|
| Business | 3.1 | 3.5 |
| Grocery Shopping | 5.0 | 7.7 |
| Hobby | 2.3 | 5.3 |
| Home | 100.0 | 100.0 |
| Leisure | 0.3 | 0.2 |
| Other | 1.7 | 1.8 |
| Other Shopping | 1.7 | 3.1 |
| Personal | 0.2 | 1.1 |
| Pickup / Dropoff | 4.5 | 10.0 |
| Restaurant/Caf | 4.8 | 8.5 |
| School | 1.1 | 3.3 |
| Visit | 1.7 | 1.8 |
| Work | 16.0 | 13.1 |

does not come at the cost of the associated intra-personal value, since they are independently generated.

## 4.6 LCS advantages for sequence comparison

Whereas the analysis performed in the previous section could be achieved by using a modified version of ED (see Equation 4) instead of LCS, the ED-based analysis would not reveal further information regarding which purposes and travel modes are common in between days or users. This section briefly investigates which purposes and travel modes are sequentially stable, and which subsequences of purposes and travel modes are common throughout the analyzed data set.

### 4.6.1 Purposes

While investigating the stability of purposes between days and between users offers interesting results, one can deepen that analysis to reveal which of the purposes are most sequentially stable. Table 10 summarizes the percentage of instances when a purpose is part of the LCS for both inter- and intra-personal cases.

It is interesting to notice that all LCSes include home, which is intuitive since most travelers start and end their day at home and, in the performed case study, all travelers started their days from home. Furthermore, the activities related to work, pickup/dropoffs, restaurants and grocery shopping are present in more LCSes than the other types of activities, in both inter- and intra-personal cases. On average, with the exception of work and leisure, activities are part of LCSes more often for the intra-personal case than for the inter-personal one, which

Table 11: The three most common purpose LCS for an LCS length of 3, 4 or more. The alphabet mapping is presented in Table 4a.

| Length | Inter-personal Sequence | # | Intra-personal Sequence | # |
|---|---|---|---|---|
| | H→W→H | 646 | H→W→H | 87 |
| 3 | H→H→H | 262 | H→D→H | 51 |
| | H→F→H | 246 | H→S→H | 37 |
| | H→W→W→H | 93 | H→O→O→H | 20 |
| 4 | H→W→R→H | 91 | H→R→F→H | 11 |
| | H→W→H→H | 60 | H→R→D→H | 10 |
| | H→W→R→W→H | 15 | H→W→D→W→H | 6 |
| >4 | H→W→R→H→H | 11 | H→W→R→W→H | 6 |
| | H→W→W→R→H | 10 | H→R→N→F→H | 4 |

is consistent with the previous findings and can be explained by the fact that once the travelers repeat activities, they tend to repeat activities in the same sequential order.

For a deeper insight into which purpose LCSes occur most often, Table 11 shows the three most common purpose LCSes when the LCS contains 3, 4 or more activities. As expected, the inter-personal purpose LCSes contain common activities such as work, home, grocery shopping and restaurants, which is partially intuitive because home and work occur in the same order since most days start at home, the trips to restaurants and coffee shops are either for a work break or on the way back home, and the groceries are usually done on the way back home. For the intra-personal LCSes, new common activities are present, such as: pickup and dropoffs of kids, school, hobbies and shopping other than groceries. An explanation for these new activities is the fact that while constraints such as dropoff kids before going to work are not present across the whole population, they are a user constraint and, as such, present across multiple days. This reveals a finding which can have important implications for activity modeling: the sequences of activities that occur in the same order for the entire user base are different from the sequences of activities that occur in the same order for each individual across different days. The extent of the implications of this observation in modeling is outside of this paper's scope.

#### 4.6.2 Travel modes

When extending the previous type of analysis to travel mode LCSes, one can observe expected patterns (Table 12). First, walking is part of half the travel mode LCSes both for inter- and intra-personal cases, which is expected because walking is both a standalone travel mode as well as a connection travel mode between higher speed modes. Similarly, personal travel modes such as bicycles and cars, as well as public transportation modes, are part of more intra-personal

Table 12: Percentage of instances each travel mode is present in the LCS.

| | Inter-personal (%) | Intra-personal (%) |
|---|---|---|
| Bicycle | 1.2 | 5.5 |
| Bus | 13.1 | 20.2 |
| Car as driver | 14.2 | 29.4 |
| Car as passenger | 2.4 | 2.9 |
| Commuter train | 0.7 | 4.9 |
| Moped / Motorcycle | 0.1 | 0.6 |
| Subway | 7.7 | 13.2 |
| Taxi | 0.5 | 2.1 |
| Train | 0.1 | 0.5 |
| Walk | 52.6 | 55.5 |

LCSes than inter-personal ones, which can be explained by economic factors (e.g., the cost of each tripleg is smaller with public transportation than with other travel modes after a public travel card is acquired) and constraints (e.g., if the user owns a car / bicycle and the tripleg away from home is via car, subsequent triplegs are more likely to also be by car / bicycle). Other travel modes such as ferryboat, flight and tram are not present in any LCS.

For a deeper insight into which purpose LCSes occur most often, Table 13 shows the three most common travel modes LCSes when the LCS contains 3, 4 or more travel modes. One interesting observation is the fact that the 3 most common LCSes of length 3 and 4 are the same for both inter- and intra-personal travel mode LCSes. Furthermore, the common LCSes include either the same mode (e.g., walking, driving) or a combination of walking and public transport, i.e., taking a bus combined with walking. Another interesting observation is that there are no symmetrical LCSes that one might expect for users that make use of public transport, which implies that for the studied user base it was more common for travelers to come back home from work using a different sequence of travel modes than when going from home to work. Finally, the intra-personal travel mode LCSes consisting of 4 or more triplegs include the sequences of travel mode one might expect (e.g, walking to the bus, taking the bus, walking to the subway station, taking the subway, etc.) but the relatively small number of occurrences of such LCSes implies that the studied user base contained considerably fewer users that travel by public transport than users traveling by car and other private modes.

# 5 Discussion

This paper avoids making any general claims due to the relatively modest user set and explorative nature of the methodology. We advise scientists to take these observations as an extension of the usual descriptive statistics associated

Table 13: Top 3 travel mode LCS for an LCS length of 3, 4 or more. The alphabet mapping is presented in Table 4b.

| Length | Inter-personal Sequence | # | Intra-personal Sequence | # |
|---|---|---|---|---|
| 3 | W→W→W | 341 | D→D→D | 126 |
| | D→D→D | 183 | W→W→W | 59 |
| | S→W→W | 67 | S→W→W | 26 |
| 4 | W→W→W→W | 111 | D→D→D→D | 19 |
| | W→S→W→W | 36 | W→S→W→W | 16 |
| | D→D→D→D | 32 | W→W→W→W | 13 |
| >4 | W→W→W→W→W | 46 | D→W→W→D→D | 12 |
| | W→W→W→W→W→W | 20 | W→S→W→T→W→W→W | 9 |
| | W→S→W→W→W | 19 | S→W→W→W→W | 8 |

with a dataset, since the the focus of the paper is on proposing a new way to explore sequences and their stability.

One of the most pertinent disadvantages of the proposed methodology is the purely sequential nature of the analysis. As such, the proposed indexes do not take into account any spatial or temporal dimensions. While this is a correct observation, the authors avoided introducing any other dimensions in the analysis to prevent feature induced obscurity. It is possible to add penalties that substitute the length of the LCS with spatio-temporal segmentation errors as shown by Prelipcean et al. (2016), as well as modifying the isomorphism functions that provide the mapping between purposes / travel modes into the used letters of the alphabet so that instead of equality they allow for more fuzzy concepts, such as mapping purposes to letters based on time of day, distance to home or any other discrete penalty. A continuous penalty can also be introduced, which would be based on time / route differences or any type of continuous attributes, but it becomes difficult to provide an isomorphic transformation that allows lossless direct and reversed mapping. This paper does not go in the direction of complicating the used mapping functions since the main focus is proposing a new way of analyzing sequential aspects of travel, whose main feature should be simplicity. It is easier to explain an index of 50% that means that half of the elements of sequence A occur in the same order in sequence B, but it is not trivial to explain an index of 50% that is built on spatio-temporal penalties.

It is also important to note that this paper does not challenge other methods to investigate similarities. One of the main drivers behind proposing this methodology is having many different space and / or time research work that yielded interesting results and moved travel behaviour modelling and understanding forward. The intent of this paper is to give a new set of tools for researchers to test their hypotheses with regards to sequences.

Finally, the authors are aware that the analysis performed in this paper is purely exploratory and, as such, it is not definitive. One of the main explanation

behind this deficit is the data collection strategy used for the project on which this analysis is based on. The project had two scopes: 1) identify the feasibility of replacing or complementing traditional travel diary collection methods with methods based on smartphone data collection accompanied by machine learning for travel entities inferences, and 2) identify how one can explore the potential dataset collected over a longer period of time from the same users. As such, the data was not collected to prove similarity of travel behaviour for users, which is the reason why this is a methodological paper with a short case study to illustrate how the methodology can be used for new research questions.

# 6 Conclusions and future work

This paper proposed a new methodology of integrating sequence analysis in the study of travel behaviour. The methodology is robust enough to allow for the definition of new indexes based on any sequential phenomenon related to travel, as well as to allow for the modification of the proposed indexes with penalties / errors that researchers desire to study or use in their models. Some of the applications of the proposed methodology were illustrated via a case study that consists of data collected from 51 users for a period of one week, where two types of similarities were explored: inter-personal similarity – for every two user combination for every day– and intra-personal similarity – for every user for every two day combination –, which are commonly used similarity measure types. These indexes were then computed for purpose, travel modes, and travel modes used for same purpose trips and were thoroughly discussed. The analysis shows that using LCS to generate indexes has considerable potential and that the interpretation of the results is not complicated, which is mainly due to the simplicity of the proposed indexes. As such, the paper proposes a starting point for embedding common subsequences into the analysis of travel behaviour.

In terms of future work, one promising direction is what was also discussed in the previous section, namely, introduce grammar rules in the isomorphism that deal with more than the obvious equality operation, which in turn would allow for dealing with continuous variables and phenomenon as well that are more complex in nature than the equality of activities. Scientists could use these measures to define ambiguous concepts such as near and far, short and long, which would enrich the similarity semantics.

Another research work direction is proposing new error measures that can bridge LCS and ED methodology to include the most common patterns which can be further examined for their distribution, as well as a measure of dissimilarity that embeds spatio-temporal penalties. One starting point for this would constitute a combination of the current paper, the ED measure proposed by Wilson (1998) and Joh et al. (2002) and the spatio-temporal penalties proposed by Prelipcean et al. (2016). This work should also be expanded to include improved ways to visualize these patterns and penalties so that they are easier to grasp and explore than by investigating tables and histograms.

Finally, this methodology can contribute to the generation of more rele-

vant travel mode and purpose schemas that are commonly used when collecting travel diaries. Since the paper's results show that there is a degree of similarity and dissimilarity both between days and users, it is important to capture the phenomenon of interest and understand the implications of using a more aggregated schema (e.g., considering car as driver, car as passenger and taxi as one travel mode instead of three) or adding more granularity to the schema (e.g., removing semantics from the non-food shopping by adding separate categories for electronics shopping and furniture shopping).

To conclude, this paper proposes a new method for the study and computation of traveler similarity that focuses on the sequential aspect of travel. The method allows for the extraction of regular patterns, as well as the computation of sequential similarity measures, which complements the existing similarity measures widely used in the travel behaviour research literature.

# References


Allahviranloo, M., Regue, R. & Recker, W. (2016), 'Modeling the activity profiles of a population', *Transportmetrica B: Transport Dynamics* pp. 1–24.

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G. & Haupt, T. (2002), 'Observing the rhythms of daily life: A six-week travel diary', *Transportation* **29**(2), 95–124.

Banerjee, A. & Ghosh, J. (2001), Clickstream clustering using weighted longest common subsequences, *in* 'Proceedings of the web mining workshop at the 1st SIAM conference on data mining', Vol. 143, Citeseer, p. 144.

Buliung, R. N., Roorda, M. J. & Remmel, T. K. (2008), 'Exploring spatial variety in patterns of activity-travel behaviour: initial results from the toronto travel-activity panel survey (ttaps)', *Transportation* **35**(6), 697.

Cherchi, E., Cirillo, C. & de Dios Ortúzar, J. (2017), 'Modelling correlation patterns in mode choice models estimated on multiday travel data', *Transportation Research Part A: Policy and Practice* **96**, 146–153.

Dharmowijoyo, D. B. E., Susilo, Y. O. & Karlström, A. (2016), 'Day-to-day variability in travellers' activity-travel patterns in the jakarta metropolitan area', *Transportation* **43**(4), 601–621.

Dijst, M. (1999), 'Two-earner families and their action spaces: A case study of two dutch communities', *GeoJournal* **48**(3), 195–206.

Frolova, D., Stern, H. & Berman, S. (2013), 'Most probable longest common subsequence for recognition of gesture character input', *IEEE transactions on cybernetics* **43**(3), 871–880.

Guo, A. & Siegelmann, H. (2004), 'Time-warped longest common subsequence algorithm for music retrieval'.

Hägerstrand, T. (1970), 'What about people in regional science?', *Papers in regional science* **24**(1), 7–24.

Hanson, S. & Huff, J. O. (1981), 'Assessing day-to-day variability in complex travel patterns'.

Hanson, S. & Huff, O. J. (1988), 'Systematic variability in repetitious travel', *Transportation* **15**(1), 111–135.

Heinen, E. & Chatterjee, K. (2015), 'The same mode again? an exploration of mode choice variability in great britain using the national travel survey', *Transportation Research Part A: Policy and Practice* **78**, 266–282.

Hirschberg, D. S. (1975), 'A linear space algorithm for computing maximal common subsequences', *Communications of the ACM* **18**(6), 341–343.

Huff, J. O. & Hanson, S. (1986), 'Repetition and variability in urban travel', *Geographical Analysis* **18**(2), 97–114.

Hunt, J. W. & MacIlroy, M. (1976), *An algorithm for differential file comparison*, Bell Laboratories New Jersey.

Järv, O., Ahas, R. & Witlox, F. (2014), 'Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records', *Transportation Research Part C: Emerging Technologies* **38**, 122–135.

Joh, C.-H., Arentze, T. A. & Timmermans, H. J. (2001*a*), 'Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms', *Geographical Analysis* **33**(3), 247–270.

Joh, C.-H., Arentze, T. A. & Timmermans, H. J. (2001*b*), 'A position-sensitive sequence-alignment method illustrated for space–time activity-diary data', *Environment and Planning A* **33**(2), 313–338.

Joh, C.-H., Arentze, T., Hofman, F. & Timmermans, H. (2002), 'Activity pattern similarity: a multidimensional sequence alignment method', *Transportation Research Part B: Methodological* **36**(5), 385–403.

Joh, C.-H., Arentze, T. & Timmermans, H. (2001*c*), 'Pattern recognition in complex activity travel patterns: comparison of euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods', *Transportation Research Record: Journal of the Transportation Research Board* (1752), 16–22.

Jones, P. & Clarke, M. (1988), 'The significance and measurement of variability in travel behaviour', *Transportation* **15**(1-2), 65–87.

Kang, H. & Scott, D. M. (2010), 'Exploring day-to-day variability in time use for household members', *Transportation Research Part A: Policy and Practice* **44**(8), 609 – 619. Special Section on Climate Change and Transportation Policy: Gaps and Facts.

Kitamura, R., Yamamoto, T., Susilo, Y. O. & Axhausen, K. W. (2006), 'How routine is a routine? an analysis of the day-to-day variability in prism vertex location', *Transportation Research Part A: Policy and Practice* **40**(3), 259–279.

Kwan, M.-P., Xiao, N. & Ding, G. (2014), 'Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm', *Geographical Analysis* **46**(3), 297–320.

Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals, *in* 'Soviet physics doklady', Vol. 10, pp. 707–710.

Lin, C.-Y. & Och, F. J. (2004), Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, *in* 'Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, p. 605.

Manley, E., Zhong, C. & Batty, M. (2016), 'Spatiotemporal variation in travel regularity through transit user profiling', *Transportation* pp. 1–30.

Minnen, J., Glorieux, I. & van Tienoven, T. P. (2015), 'Transportation habits: evidence from time diary data', *Transportation Research Part A: Policy and Practice* **76**, 25–37.

Moiseeva, A., Timmermans, H., Choi, J. & Joh, C.-H. (2014), 'Sequence alignment analysis of variability in activity travel patterns through 8 weeks of diary data', *Transportation Research Record: Journal of the Transportation Research Board* (2412), 49–56.

Needleman, S. B. & Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of molecular biology* **48**(3), 443–453.

Neutens, T., Delafontaine, M., Scott, D. M. & De Maeyer, P. (2012), 'An analysis of day-to-day variations in individual space–time accessibility', *Journal of Transport Geography* **23**, 81–91.

Pas, E. I. (1983), 'A flexible and integrated methodology for analytical classification of daily travel-activity behavior', *Transportation science* **17**(4), 405–429.

Pas, E. I. (1987), 'Intrapersonal variability and model goodness-of-fit', *Transportation Research Part A: General* **21**(6), 431–438.

Pas, E. I. & Koppelman, F. S. (1987), 'An examination of the determinants of day-to-day variability in individuals' urban travel behavior', *Transportation* **14**(1), 3–20.

Prelipcean, A. C., Gidofalvi, G. & Susilo, Y. O. (2015), Comparative framework for activity-travel diary collection systems, *in* '2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)', IEEE, pp. 251–258.

Prelipcean, A. C., Gidofalvi, G. & Susilo, Y. O. (2016), 'Measures of transport mode segmentation of trajectories', *International Journal of Geographical Information Science* **30**(9), 1763–1784.

Prelipcean, A. C., Gidofalvi, G. & Susilo, Y. O. (2018), 'MEILI: A travel diary collection, annotation and automation system', *Computers, Environment and Urban Systems* **forthcoming**, 1–11.

Recker, W. W., McNally, M. G. & Root, G. S. (1985), 'Travel/activity analysis: pattern recognition, classification and interpretation', *Transportation Research Part A: General* **19**(4), 279–296.

Recker, W. W., McNally, M. G. & Root, G. S. (1987), 'An empirical analysis of urban activity patterns', *Geographical Analysis* **19**(2), 166–181.

Schlich, R. & Axhausen, K. W. (2003), 'Habitual travel behaviour: evidence from a six-week travel diary', *Transportation* **30**(1), 13–36.

Schönfelder, S. & Axhausen, K. W. (2000), 'Analysing the rhythms of travel using survival analysis'.

Schönfelder, S. & Axhausen, K. W. (2003), 'Activity spaces: measures of social exclusion?', *Transport policy* **10**(4), 273–286.

Srivastava, G. & Schönfelder, S. (2003), *On the temporal variation of human activity spaces*, ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme Zurich.

Susilo, Y. & Kitamura, R. (2005), 'Analysis of day-to-day variability in an individual's action space: exploration of 6-week mobidrive travel diary data', *Transportation Research Record: Journal of the Transportation Research Board* (1902), 124–133.

Susilo, Y. O. & Axhausen, K. W. (2014), Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman index, Vol. 41, Springer, pp. 995–1011.

Thøgersen, J. (2006), 'Understanding repetitive travel mode choices in a stable context: A panel study approach', *Transportation Research Part A: Policy and Practice* **40**(8), 621 – 638.

Wilson, W. C. (1998), 'Activity pattern analysis by means of sequence-alignment methods', *Environment and Planning A* **30**(6), 1017–1038.

Zhong, C., Manley, E., Arisona, S. M., Batty, M. & Schmitt, G. (2015), 'Measuring variability of mobility patterns from multiday smart-card data', *Journal of Computational Science* **9**, 125–130.