Understanding of representative 24h travel activity sequences of Londoners

Yiqiao Chen*[1] and Elisabete A. Silva†[1]

[1] Lab of Interdisciplinary Spatial Analysis, Department of Land Economy, University of Cambridge

March 20, 2021

Summary

A 24-hour travel activity sequence is a traveller's time-use activity in a day, containing location, function, purpose, and trip mode information in each time interval. Understanding the daily time-use profile of travellers can reveal the travel patterns in a city, uncover the travel behaviour of citizens, and increase the accuracy of activity-based travel forecasting models. The focus of this paper is to cluster the 24h travel activity sequences and learn the daily travel activity patterns of Londoners. We analyse the National Travel Survey data. A three-stage clustering algorithm is applied to group similar travel sequences and find representative travel patterns.

KEYWORDS: 24-h travel sequence, representative activity sequences, pattern clustering, London travel patterns

1. Introduction

A travel activity sequence is a traveller's time-use activity in a time period. The sequences illustrate the information such as location, function, purpose, and mode in each time interval. To understand the travel behaviours of citizens and build more robust activity-based travel models, we can explore the representative travel sequences of travellers in a city (Ahmed et al., 2020). Representative sequences can statistically represent the behavioural patterns and 'optimal' behaviour of individuals in a city (Allahviranloo et al., 2017). Sequence alignment method, clustering algorithms, and machine learning approaches are usually used to finding representative sequential patterns (Hafezi et al., 2018; Allahviranloo et al., 2017; Xu and Kwan, 2020). Existing studies identified sequential travel activity patterns at different spatial scales, including Chicago (Xu and Kwan, 2020), California (Allahviranloo et al., 2017), and Canada (Hafezi et al., 2018) etc. However, the travel sequences in the individual or entire UK cities haven't been investigated and are worth analysing. Previous studies on Londoners' travel activities are either focused on sub-groups or specific behaviours (Transport for London, 2017; Langlois et al., 2016). This study intends to analyse the sequences of the main groups of London travellers on the weekdays and reveal their travel activity patterns. The findings can increase the understanding of travel patterns and guide policymaking.

2. Study Area and Data

This paper chooses Greater London as the study area. The UK national travel survey (NTS) for the year 2012 to 2019 is used for this study (Department for Transport, 2020). NTS provides a representative sample of householders based on a stratified random probability sample (Cornick et al., 2020). Thus, we use the sample of NTS to understand different groups of travellers. It is a reliable source and has been widely used to analyse travel patterns and inform policymaking (Cornick et al., 2020; Luiu and Tight, 2021). We choose the trips within Greater London and the trips on weekdays. This study considers eight types of activities for analysing and generating the 24h activity sequence patterns. They are "Home(H)", "Work(W)", "Education(E)", "Shopping(S)", "Personal Business(P)", "Recreation(R)", "Escort/ Company(C)" and "Other(O)". The main trip chains are listed in **Table 1**.

* yc418@cam.ac.uk

† es424@cam.ac.uk

Table 1 Main trip chains

Trip chains	Frequency	Percentage (%)
HWH	12759	23
HRH	7139	13
HSH	5410	10
HEH	4949	9
HPH	3077	6
HCH	1707	3

3. Method

3.1. Data processing

We first convert the raw data into daily travel sequences. The data of NTS is trip-based so we reorganise an individual's trips in one day by stage numbers. We recode the 23 trip purposes in numeric form into the 8 activity types in string form. Individual travel diaries are reconstructed with 8 activity types at 1-hour time intervals from 0:00am to 24:00pm. A daily sequence is in the form of a sequence of 24 letters, representing activity states in each hour. A total of 52886 daily sequences, with 8953 unique sequences, are generated from 16704 individuals.

3.2. Weighted sum distances

After constructing the 24h travel sequences for each individual every day, we select the sequences on weekdays. This study uses a 3-stage clustering algorithm to identify representative travel activity sequences. To align the sequences, the dissimilarities between each pair of sequences need to be measured. SAM is a method from molecular biology for analysing genetic sequences and now one of the most widely used similarity measures in transport research (Shou and Di, 2018). It measures the distance between two sequences by calculating the minimal cost to transform one into the other (Kim, 2014; Wilson, 1998). We choose the most used distance measures in SAM as a starting point to analyse activity sequences in this study. Edit distance calculates the minimum cost of single-character edits (i.e. insertions, deletions or substitutions) to show the dissimilarities (Wilson, 1998; Shou and Di, 2018; Allahviranloo et al., 2017). To better account for the changes of activity types, we use a weighted sum distance combining sequence alignment method (SAM) and agenda dissimilarity (Allahviranloo et al., 2017). Agenda dissimilarity can be measured by Jaccard distance, which is calculated by the relative size of the intersection divided by that of the union of the sample sets (Kosub, 2019). We use Levenshtein distance (the most commonly used edit distance approach in natural language processing) and Jaccard distance to calculate edit distance and agenda dissimilarity respectively. This study uses the weights from Allahviranloo et al. (2017) and the weighted distance is shown in Equation (1).

$$\text{Weighted sum distance} = 0.9 * \text{Edit distance} + 0.1 * \text{Agenda dissimilarity} \tag{1}$$

3.3. Three-stage Clustering

For clustering, we first use the Affinity Propagation algorithm to initialise the cluster numbers and locate the initial centroids as it is a fast clustering technique (Allahviranloo et al., 2017). We apply the Affinity Propagation algorithm to the most frequent (>75%) unique sequences. Then we use K-medoids to cluster sample set of all sequences to identify the representative sequences, setting as reference episodes. Finally, we calculate the pair-wise weighted sum distance between each sequence in the full set and the reference episodes. We compare the distances, and the sequence is clustered into the group where the distance between the sequence and reference episode is the smallest.

For yearly sequences, similar approaches are applied. The Affinity Propagation algorithm is first run to initialise cluster numbers and the K-medoids algorithm is then run to cluster the yearly data.

4. Results

We identify 16 unique clusters from the travel activity sequences from 2012 to 2019 (see **Table 2**).

Based on the temporal pattern and the main activity type, we named each cluster. In general, the cluster model identifies 5 clusters for workers and 11 clusters for non-workers. The 5 workers groups are long-hour workers, morning workers, daytime workers, afternoon workers, and night workers. The 11 non-worker clusters contain 1 student cluster and 1 stay-at-home cluster for individuals who spend major time at home. The rest of 9 non-worker clusters are night recreation non-workers, midday recreation non-workers, afternoon recreation non-workers, all-day recreation non-workers, midnight recreation non-workers, midday shopping non-workers, midday personal business non-workers, morning escort non-workers and afternoon escort non-workers.

Table 2 16 Representative Sequences (2012-2019)

	Category	Cluster name	Exemplar
1	Non-workers	Students	HHHHHHHHEEEEEEEEHHHHHHHH
2	Workers	Long-hours workers	HHHHHHHHWWWWWWWWWWHHHHHH
3	Workers	Morning workers	HHHHHHWWWWWWWWWHHHHHHHHH
4	Workers	Daytime workers	HHHHHHHHHWWWWWWWWHHHHHHH
5	Workers	Afternoon workers	HHHHHHHHHHHHHHWWWWWWWWWH
6	Workers	Night workers	WWWWWWWHHHHHHHHHHHHHHWWW
7	Non-workers	Stay-at-home	HHHHHHHHHHHHHHHHHHHHHHHH
8	Non-workers	Night recreation	HHHHHHHHHHHHHRRRRRRRRRRR
9	Non-workers	Midday recreation	HHHHHHHHHHHRRRRRHHHHHHHH
10	Non-workers	Afternoon recreation	HHHHHHHHHHHHHHHRRRRRHHHH
11	Non-workers	All-day recreation	RRRRRRRRRRRRRRRRRRRRRRRR
12	Non-workers	Midnight recreation	RRRRRRRRRRRRRRHHHHHHHHHH
13	Non-workers	Midday shopping	HHHHHHHHHHHSSHHHHHHHHHHH
14	Non-workers	Midday Personal business	HHHHHHHHHHPPPPPHHHHHHHHH
15	Non-workers	Morning escort	HHHHHHHHCCHHHHHHHHHHHHHH
16	Non-workers	Afternoon escort	HHHHHHHHHHHHHHHCHHHHHHHH

The number of clusters varied in the range of 10-13 for the years 2012-2019, as shown in **Figure 1**. A comparison of the representative sequences in different years can show the similarities in the past 8 years and some changes. In the previous years, most Londoners spent their time at home, work and recreation. Activities usually took place in the daytime. The main activity that happened at nighttime is working.

The most stable clusters in the past years are students and stay-at-home non-workers. They appeared in all 8 years as one of the representative sequences. Generally, 3-4 sequences are for the workers' groups. Daytime workers, long-hour workers, and afternoon workers are the most common clusters. The recreation non-workers groups have 2-5 sequences in the representative episodes, varied each year. Daytime and afternoon-night recreations can be found in all years among the non-workers. In all the years, we can find a shopping non-workers episode and a personal business non-workers episode. The shopping activities usually happened in the afternoon and the personal businesses often happened in the daytime. The escort non-workers activities can be found in some year while the sequences contain escort activities were less seen in other years.

Further investigation is required into how determinants such as socio-demographic variables are linked with representative sequences and why representative sequences change over time. Travel demands of different population groups can be better understand and simulated, which can further support the healthy, active and inclusive transport objectives in London.

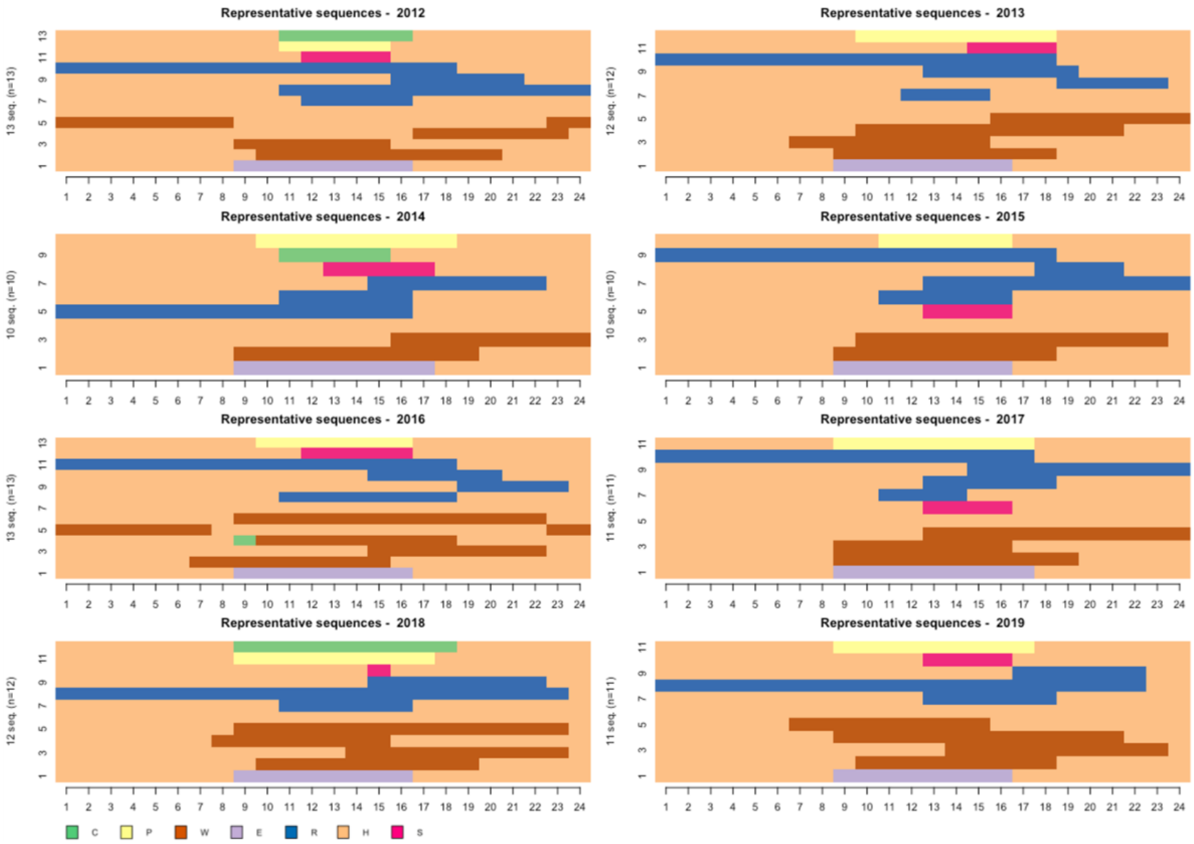

Figure 1 Yearly representative sequences

5. Acknowledgements

The authors express thanks to the UK Data Service for the data used in this work.

References

Ahmed U, Moreno AT and Moeckel R (2020) Microscopic activity sequence generation: a multiple correspondence analysis to explain travel behavior based on socio-demographic person attributes. *Transportation*. DOI: 10.1007/s11116-020-10103-1.

Allahviranloo M, Regue R and Recker W (2017) Modeling the activity profiles of a population. *Transportmetrica B-Transport Dynamics* 5(4): 426-449.

Cornick P, Cant J, Yarde J, et al. (2020) National Travel Survey 2019 Technical Report. Reportno. Report Number|, Date. Place Published|: Institution|.

Department for Transport (2020) National Travel Survey, 2002-2019: Special Licence Access, [data collection], UK Data Service, 9th Edition. In: Transport Df (ed).

Hafezi MH, Liu L and Millward H (2018) Learning Daily Activity Sequences of Population Groups using Random Forest Theory. *Transportation Research Record* 2672(47): 194-207.

Kim K (2014) Discrepancy Analysis of Activity Sequences What Explains the Complexity of People's Daily Activity-Travel Patterns? *Transportation Research Record*. DOI: 10.3141/2413-03.(2413): 24-33.

Kosub S (2019) A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters* 120: 36-38.

Langlois GG, Koutsopoulos HN and Zhao JH (2016) Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C-Emerging Technologies* 64: 1-16.

Luiu C and Tight M (2021) Travel difficulties and barriers during later life: Evidence from the National Travel Survey in England. *Journal of Transport Geography* 91: 102973.

Shou ZY and Di X (2018) Similarity analysis of frequent sequential activity pattern mining. *Transportation Research Part C-Emerging Technologies* 96: 122-143.
Transport for London (2017) Transport Classification of Londoners. Reportno. Report Number|, Date. Place Published|: Institution|.
Wilson C (1998) Analysis of travel behavior using sequence alignment methods. *Transportation Research Record* 1645(1): 52-59.
Xu L and Kwan MP (2020) Mining sequential activity-travel patterns for individual-level human activity prediction using Bayesian networks. *Transactions in Gis* 24(5): 1341-1358.

Biographies

Miss Yiqiao Chen is a PhD candidate in the Department of Land Economy, University of Cambridge. Her research interests cover urban data analytics, machine learning, smart city and smart transport governance.

Prof. Elisabete A. Silva is a University Professor in Spatial Planning in the Department of Land Economy, University of Cambridge. Her research interests are centred on the application of new technologies to spatial planning in particular city and metropolitan dynamic modelling through time.