Inferring patterns in the multi-week activity sequences of public transport users

Inferring Patterns in the Multi-week Activity Sequences of Public Transport Users

Gabriel Goulet Langlois[a], Haris N. Koutsopoulos[b], Jinhua Zhao[c]

[a]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States*
[b]*Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, United States*
[c]*Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, United States*

Abstract

The public transport networks of dense cities such as London serve passengers with widely different travel patterns. In line with the diverse lives of urban dwellers, activities and journeys are combined within days and across days in diverse sequences. From personalized customer information, to improved travel demand models, understanding this type of heterogeneity among transit users is relevant to a number of applications core to public transport agencies' function. In this study, passenger heterogeneity is investigated based on a longitudinal representation of each user's multi-week activity sequence derived from smart card data. We propose a methodology leveraging this representation to identify clusters of users with similar activity sequence structure. The methodology is applied to a large sample (n=33,026) from London's public transport network, in which each passenger is represented by a continuous 4-week activity sequence. The application reveals 11 clusters, each characterized by a distinct sequence structure. Socio-demographic information available for a small sample of users (n=1,973) is combined to smart card transactions to analyze associations between the identified patterns and demographic attributes including passenger age, occupation, household composition and income, and vehicle ownership. The analysis reveals that significant connections exist between the demographic attributes of users and activity patterns identified exclusively from fare transactions.

Keywords: Travel Behavior, Smart Card Data, Activity Sequence, User Clustering, Public Transportation, Data Mining

1. Introduction

Diverse cities and the varied opportunities they foster are reflected in the heterogeneous travel patterns of the passengers of large urban transit networks. Beyond conventional 9-to-5 commuters, a variety of non-working routines and non-conventional work routines (for example driven by shift work, multi-employment, or self-employment) structure the activity patterns of public transport (PT) users. While these diverse activity patterns are typically considered on a daily basis, considering activity sequences across multiple days and weeks may reveal important differences among users. Segmenting

users based on these differences is useful to gain a better understanding of the PT passenger population. From the provision of passenger information customized to specific user segments, to targeted travel demand management campaigns (Halvorsen, 2015), to service planning informed by the types of passengers traveling along different portions of the network, knowledge of the diversity among transit users provides opportunities to improve passenger experience and service provision.

Exploring heterogeneity in multi-week activity and journey sequences requires longitudinal observations of users. While conventional survey data contain detailed information about most aspects of a user's activity pattern (purpose, location, etc.), their costs typically proscribe large samples of users from being observed over long time periods. In contrast, smart card data provides a continuous stream of information about the PT travel of a large number of users. This information can be used to partially infer, and hence analyze, certain components of each user's general activity pattern (Lee & Hickman, 2014; Kusakabe & Asakura, 2014). Pelletier et al. (2011) present a review of research leveraging smart card data for such analysis.

Specifically related to this research, some studies focus on segmenting the travel patterns of PT users using smart card data. Ortega-Tong (2013) defined 20 different clustering variables related to travel frequency, journey times, origin-destination pairs, activity duration, fare type and public transport mode choice to identify 8 different user segments using the K-medoids algorithm. The resulting 8 groups were aggregated into four categories: non-exclusive commuters, exclusive commuters, non-commuter residents, and leisure travelers. Focusing on travel regularity, Ma et al. (2013) identified journey characteristics, including journey boarding time, bus route sequence, and bus stop sequences, frequently observed for the same user over a 1-week period in Beijing. From the number of days traveled and the number of frequent journey characteristics identified for each user, they define 5 clusters of varying regularity levels using the k-means++ clustering algorithm. Similarly to Ma et al. (2013), Kieu et al. (2014) defined measures of temporal regularity and spatial regularity focused on weekday travel to segment public transport users in South East Queensland, Australia. They subjectively define segment boundaries for the resulting distribution and identify four groups: irregular passengers, regular OD pair passengers, habitual time passengers, and routine OD and time passengers. Finally, the early work of Morency et al. (2007) used k-means to identify typical patterns of bus boarding time in Gatineau, Canada.

While the work of these authors highlights the potential of smart card data to classify travel patterns, the approaches are limited in capturing the sequence within which each journey occurs. The clustering variables used by these studies are all derived from a scalar aggregation of a passenger's journeys which ignores the organization of multiple journeys over time.

Moving away from a fully scalar representation of user's travel patterns, El Mahrsi et al. (2014) use a vector of hour periods to represent the times at which each user is observed traveling. They identify 16 clusters of weekly temporal patterns by comparing the times at which users start journeys on each day of the week. While their approach preserves the order of hours within the week, it relies on aggregating multiple weeks of data to compute an average number of journeys for each hour. As such, it also ignores the sequence in which journeys are completed, and disregards all geographical information about journeys.

Important information about passenger's activity pattern is lost through such aggre-

gation. As described by Hagerstraand (1970), and in line with the precepts of activity based travel theory, certain activity patterns include activities and journeys arranged in 'non-permutable' sequences. Activity patterns are defined not only by the attributes of the activities and journeys they are composed of, but also by the order in which these activities are organized. Abstracting this order may obscure sequence structure specific to certain passenger segments.

The research is organized around two objectives. First, we aim to develop a methodology leveraging smart card data to identify clusters of users sharing similar multi-week activity sequences. This methodology should provide an approach to investigate heterogeneity among passengers which can be applied systematically over time using continuously collected fare transactions. Second, we aim to provide empirical analysis of the heterogeneity among users of an extensive transit network through a large scale application of this methodology in London's transit system. This aim focuses on describing the underlying structure of activity sequences contained in each cluster and on exploring the socio-demographic attributes associated with each pattern.

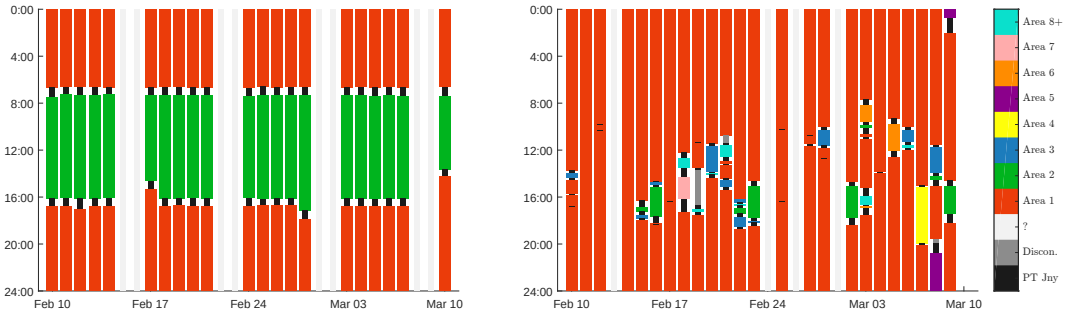
In line with these objectives, the contribution of this work is twofold. From a methodological perspective, we provide a novel representation of travel patterns based on the longitudinal activity sequence of each user, and synthesize pervasive computing and data mining methodologies to identify trends from these sequences. From an empirical perspective, we analyze and expose the nature of heterogeneity among London's public transport users. We also provide evidence of significant associations between the patterns identified from traces of travel alone and socio-demographic attributes, by combining socio-demographic data about individual users to smart card records.

The remainder of this paper is organized as follows. Section 2 provides an overview of the methodology, and section 3 describes the application of this methodology to London's user population. Finally, the conclusions and limitations of the work are discussed in section 4.

2. Methodology

2.1. Representing longitudinal activity sequences

Central to the approach implemented in this research is the representation of each individual's travel pattern. In order to preserve the relationships between journeys and activities organized over multiple days, each user is represented as a time-ordered sequence of activities inferred from smart card data. Figure 1 illustrates two such sequences, each associated with a different individual. Each column along the x-axis shows a day, covering a 4-week analysis period. Time of day is indicated on the y-axis. The different colors indicate different activity locations, revealing two contrasting patterns for both users. The first is characterized by long activities in the green location on weekdays between 8:00 and 16:00, and evenings and mornings spent in the red location. The second is characterized by a large proportion of time spent at the red location, interrupted by shorter activities scattered across the 4-week period. While the travel of all passengers is unique to some degree, similarities with respect to the structure of such sequences exist across individuals. For example, users who use public transport (PT) to commute on conventional 9-to-5 schedules are likely to follow a pattern similar to that of the first sequence.



(a) User 1 (b) User 2

Figure 1: Two Example Activity Sequences

As smart card data provides no explicit information about activity purpose and captures solely PT journeys, it only allows for longitudinal activity sequences to be partially reconstructed based on the location of stops and stations used to access activities. We propose an approach to partially infer activity sequences from smart card data organized in two steps. First, the stops and stations visited by each individual are clustered in user-specific areas aligned with different activity locations. Users' activity sequences are then inferred from the origin and destination areas of consecutive journeys. Given these sequences, the travel of each user is summarized with respect to the underlying structure of sequences, in contrast to the scalar aggregation metrics used by Ortega-Tong (2013), Ma et al. (2013), and Kieu et al. (2014).

2.2. Defining user-areas

Let $X_u = \{x_1, x_2, \ldots, x_{n_u}\}$ be the set of all stops or stations visited by a user u as the origin or destination of a journey, where n_u represents the number of distinct locations visited by the user. As a passenger may use different stops or stations to access the same activity location (e.g. depending on time of day, day of the week, or previous location), stops and stations are grouped in k_u geographical user-areas. Denote the set of areas defined for user u by $\mathcal{A}_u = \{A_1, A_2, \ldots, A_i, \ldots, A_{k_u}\}$, such that $A_i \subseteq X_u \, \forall A_i \in \mathcal{A}_u$. A separate set of areas $\mathcal{A}_u$ is defined for each user u using hierarchical clustering with complete distance (Day & Edelsbrunner, 1984).

As described in Algorithm 1, areas are defined by iteratively merging the two closest areas until the smallest distance δ between two areas is greater or equal to a predefined threshold distance D. The distance between two sets of stops or stations A_i and A_j is measured by

$$\delta(A_i, A_j) = \max_{x_l \in A_i, x_m \in A_j} (d'(x_l, x_m)) \tag{1}$$

$$d'(x_l, x_m) = \begin{cases} d(x_l, x_m) & \text{if } t_{l,m}/T_u < \tau \\ D & \text{if } t_{l,m}/T_u \geq \tau \end{cases} \tag{2}$$

where $d(x_l, x_m)$ denotes the euclidean distance between two stops or stations, $t_{l,m}$ is the number of journeys observed between x_l and x_m for user u, T_u is the total number of

Algorithm 1 Agglomerative Hierarchical Clustering

Input: All locations visited by user u in $X_u = \{x_1, x_2, \ldots, x_{n_u}\}$, maximum distance threshold D, δ distance function between two sets of stops or stations

Output: Set of areas $\mathcal{A}_u$ for user u composed of clustered locations

Initialize each x to singleton area such that $\mathcal{A}_u = \{A_1, \ldots, A_{n_u}\}$
while $\min_{\{(A_i, A_j) \in \mathcal{A}_u \times \mathcal{A}_u; i \neq j\}} \delta(A_i, A_j) < D$ **do**
  $(A_i, A_j) \leftarrow \arg\min_{\{(A_i, A_j) \in \mathcal{A}_u \times \mathcal{A}_u; i \neq j\}} \delta(A_i, A_j)$
  Merge A_i and A_j as one cluster
end while

journeys completed by u, and τ is predefined parameter. Stops between which the user frequently travels are likely to be associated with distinct activities. Hence, the parameter τ is used to ensure that pairs of stops or stations between which a high percentage of the user's journeys are observed are not grouped in the same user-area. The maximum distance threshold D ensures that all stops and stations grouped in the same area are separated by no more than a predefined walkable distance.

2.3. Inferring longitudinal activity sequences

In order to reconstruct activity sequences as illustrated in Figure 1, an activity status corresponding to user-areas can be assigned to each interval bounded by the user's journeys. To do so, the journeys of each individual are ordered by time and each journey is considered sequentially. For each journey i, the destination and origin areas of neighboring journeys $i-1$ and $i+1$, respectively, is used to infer the activity status as described below. This approach can be implemented using any smart card data for which both journey origin and destination is explicitly recorded, or indirectly inferred using vehicle location data (Chu & Chapleau, 2010, Munizaga & Palma, 2012, and Gordon et al., 2013). For this research, the algorithm presented by Gordon et al. (2013) is used to reconstruct journeys.

1. If the current journey i started on the same day as journey $i-1$, or on the day directly following the day of $i-1$, establish the users activity status from the end time of $i-1$ to the start time of i by comparing the destination of journey $i-1$ and the origin of journey i. If the destination area of $i-1$ is the same as the origin area of i, infer it to be the activity status. If the areas are different, the user traveled between areas using a non-PT mode during the interval.
2. If the current journey i started on a day later than the day directly following journey $i-1$, or if journey i is the first journey made by the user, infer the location of the user from the start of the day on which the current journey was made to the start of journey i based on the origin of i.
3. If the current journey ended on the same day as journey $i+1$, or on the day directly preceding journey $i+1$, the location from the end of journey i to the start of journey $i+1$ can be inferred as explained in 1 at the next step, when journey $i+1$ is considered as the current journey.
4. If the current journey ended on a day earlier than the day directly preceding journey $i+1$, or if journey i is the last journey made by the user, the location from the

end of journey i to the end of the day can be inferred based on the destination of journey i.

All journeys completed by a user can hence be linked into a sequence of intervals, characterized by an activity status aligned with the user's areas, a start-time, and an end-time. Intervals for which the origin of journey i does not match the destination of journey $i-1$ are assigned status -1, indicating that the user traveled between two areas using a non-PT mode during the interval. Intervals during which the user is on a PT journey are assigned the status -2. For certain intervals, information may be insufficient to make any inference about the user's location. This includes days on which no journeys are observed, or intervals for which data issues result in missing origins and destinations. These intervals are assigned status 0.

Table 1: Activity Status Summary

Activity Status	Meaning
-2	User is in public transit journey
-1	User traveled between two areas using a non-PT mode during the interval
0	User Location cannot be inferred due to insufficient information
1	User is located at Area 1
2	User is located at Area 2
...	User is located at Area ...

The value of each possible activity status is summarized in Table 1. Statuses below 1 indicate intervals for which user-area cannot be inferred. Statuses above 0 indicate intervals for which the user-area was successfully inferred. Once the user's activity status is inferred over the period of analysis, user-areas are ordered with respect to the amount of time spent in each area. Hence, area 1, always aligns with the area in which the user was inferred to spend most time. Beyond this ordering, no specific ordinal scale is implied by the numeric value of statuses.

2.4. Cluster analysis

In order to cluster users based on the organization of their activities over multiple weeks, the inferred activity sequence is discretized into a series of finite time-bins (e.g. 1 hour bins). This is akin to modeling the activity sequence of each individual as an image. Each pixel of this image corresponds to a time-bin and the value assigned to the pixel corresponds to the status of the individual during the hour. By extracting statistical trends in variations of pixel values across the sequences of all users, recurrent elements of sequence structure can be identified and used to summarize each sequence through a small number of dimensions. Sequences can then be clustered based on their low-dimensionality representation. Principal component analysis (PCA) is a commonly used method to identify such statistical trends in high-dimensionality data.

The use of PCA to analyze mobility patterns in this fashion was first introduced by Eagle & Pentland (2009). Using data from MIT's Reality Mining experiment, they identified inherent elements of structure in the daily behavior of 100 individuals. Using

Eagle and Pentland's approach, Jiang et al. (2012) clustered the daily activity patterns of over 23,000 individuals from a Chicago activity diary survey. Building on the work of these authors, we use PCA to represent the multi-week sequence of each user as a combination of recurrent elements of sequence structure (eigen-sequences).

For a period of y days, divided into z bins per day (e.g. 24 bins/day), each user's sequence is represented by a vector of yz activity statuses, assigned one of the possible statuses defined above. As activity status values represent distinct categories on no specific ordinal scale, each vector is transformed from a categorical vector of yz elements to a binary vector of yzs, where s represents the number of possible statuses. All vectors are assembled into a $U \times yzs$ binary matrix, of which each column represents a status-hour and each row one of the U sampled users. This matrix is standardized by subtracting the average of each column from all values in the column. The resulting standardized matrix is denoted by B. In order to compute the principal components of matrix B, the eigenvectors $\mathbf{v}$ and eigenvalues λ of B's covariance matrix C are identified by solving equation 4.

$$C = B^T B \tag{3}$$

$$(C - \lambda I)\mathbf{v} = 0 \tag{4}$$

where I denotes the identity matrix.

The solutions to equation 4 are denoted by the eigenvalue and eigenvector sets $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ respectively, where n represents the rank of C. The result of this process is a set of yzs orthogonal eigenvectors of yzs dimension. These vectors constitute the principal components (PC) of B. Each eigenvector is associated to an eigenvalue which is proportional to the amount of variation observed along the direction of the vector. PCs are ordered according to their associated eigenvalue, such that the first PC explains the most variation in B.

Longitudinal activity sequences can be reconstructed by overlaying the correlation patterns described by multiple PCs. A user's sequence is approximated by a weighted sum of PCs. For a given sequence b_u, the weight $w_{i,u}$ of a component $\mathbf{v_i}$ is computed by projecting b_u onto $\mathbf{v_i}$ (equation 5).

$$w_{i,u} = b_u \cdot \mathbf{v}_i^T \tag{5}$$

This projection is a measure of the extent to which the pattern described by $\mathbf{v_i}$ is observed within user u's sequence b_u. A sequence can hence be represented by a smaller set of variables corresponding to its projection onto the most important PCs. The variables are then used to cluster sequences with similar structure through multivariate cluster analysis.

2.5. Socio-demographic associations

While activity sequence clusters can be defined from a large smart card sample of the general passenger population, detailed demographic data from surveys is typically available only for a small sample of users. In order to examine associations between user demographics and each cluster, it is therefore necessary to map users in the smaller sample for whom both smart card and demographic data is available to clusters defined from the

large smart card sample. This is done by assigning each user-sequence from the small sample to the cluster whose centroid is closest to the sequence projection (Manning et al., 2008). Demographic associations are then evaluated using odds ratio analysis (Szumilas, 2010) and a multinomial logit model of cluster membership as a function of demographic attributes. As summarized by Equation 6, the odds ratio estimate measures the ratio of the odds of having a given demographic characteristic a given membership to a given cluster k over the odds of having characteristic a given membership to any other cluster.

$$\widehat{OR}_{a,k} = \frac{N_{a,k}/N_{a',k}}{N_{a,k'}/N_{a',k'}} \tag{6}$$

where $N_{a,k}$ denotes the number of users with characteristic a in cluster k. All clusters other than k are aggregated as k', and similarly for characteristic a. This measure indicates of how much more (or less) likely a user in cluster k is to have characteristic a compared to a user who does not belong to k.

3. Case Study

The approach outlined in section 2 is applied to a case study focused on the Transport for London (TfL) public transport network. With nearly 10 million journeys daily (Transport for London, 2014), 270 underground stations (Transport for London, 2015b), and over 19,000 bus stops (Transport for London, 2015a), TfL's comprehensive PT network serves over 45% of all journey stages in the greater London area (Transport for London, 2014). In line with the ubiquity of the PT network, an important proportion of Londoners' activities can be captured from smart card data. This provides a valuable opportunity to investigate heterogeneity in longitudinal activity sequences in the context of a large metropolitan area.

3.1. Data

3.1.1. Transport for London Smart Card Data

The primary dataset available for this study consists of the smart card records of a sample of TfL passengers observed between February 10th and March 10th 2014. The sample contains transactions associated with over 3 million stages completed during the 4-week period across all public transport modes, including bus and rail. Rail records contain the origin and destination station, as well as the start time and end time of each rail stage. In contrast, bus stages only include boarding time and location. Hence, the data is processed to infer the alighting time and location of bus stages according to the methodology developed by Gordon et al. (2013). This approach uses automatic vehicle location data and information on subsequent stage boarding time and location to infer the alighting location and time of each bus stage. Approximately 80% of bus alighting are successfully inferred. Given the origin and destination of bus and rail stages, intermodal journeys are reconstructed by linking stages based on appropriate rules Gordon et al. (2013).

3.1.2. Sample Selection

As public transport travel represents a different proportion of each individual's general mobility, smart card transactions provide varying levels of information about the activity patterns of different users. Some individuals use public transport frequently for most of their journeys, while others use it occasionally at specific times or for specific purposes. The fare transactions of passengers who use public transport occasionally only reveals a small portion of longitudinal activity patterns. In order to identify users whose activity pattern can be inferred more completely from smart card data, cards are clustered based on their level of public transport usage. Each card is characterized by the number of days it was observed traveling over the 29-day analysis period, and by the spread of days between the first and last day it is observed. Using these two variables and k-means clustering, 3 user clusters are identified: a group of non-recurrent users who are seen traveling few days concentrated over a short period (average of 2 days traveled and 4 days of spread), a group of occasional users who travel on few days spread over the analysis period (average of 8 days traveled and 22 days of spread) and a group of frequent users who travel on many days spanning most of the analysis period (average of 22 days traveled and 28 days of spread). The resulting frequent user group consists of 33,026 cards, accounting for 33% of all cards in the available sample and for over 70% of trips completed during this period. We refer to these 33,026 users as the primary sample.

3.1.3. London Travel Demand Survey

The London Travel Demand Survey (LTDS) is a continuous household survey focused on the travel of Greater London residents. On each survey year, a random sample of approximately 8,000 households, including approximately 19,000 individuals, is interviewed face-to-face. The interview covers questions related to characteristics of the household and household members above 5 years old. Since April 2011, LTDS respondents over 17 years old are asked, on a voluntary basis, to provide the ID number of their Oyster cards, allowing two typically distinct types of data to be connected on an individual level. Of all card IDs provided by respondents interviewed for the 2011-2012 and the 2012-2013 surveys available for this research, 5,713 were observed on TfL's network between February 10th and March 10th 2014. The usage of these 5,713 individuals is classified as described in section 3.1.2, revealing 1,973 frequent users for whom both smart card transactions and detailed socio-demographic information are available. We refer to these 1,973 users as the LTDS sample.

3.2. Application

3.2.1. Inferring Longitudinal Activity Sequences

The 4-week longitudinal activity pattern of each user in the primary sample is reconstructed based on smart card transactions. All stops and stations visited by a given individual are grouped in into user-areas of less than 1000 meter in diameter and such that no origin-destination pair accounting for over 10% of the user's journeys is grouped in the same area. These areas are then used to infer the activity status of users as described in section 2.3. The threshold of 1000 meter was selected to define walkable user-areas, and in line with sensitivity analysis detailed in Goulet-Langlois (2015).

Table 2 summarizes the average proportion of time users were inferred to spend in each activity status. Figure 2 shows the distribution of activity duration for 2 distinct activity

Table 2: Distribution of Time Inferred Across Activity Status

	Status							
	-2	-1	0	1	2	3	4	≥ 5
Average Proportion Time Spent (%)	3.4	9.0	27.7	36.5	14.6	4.1	1.9	2.8

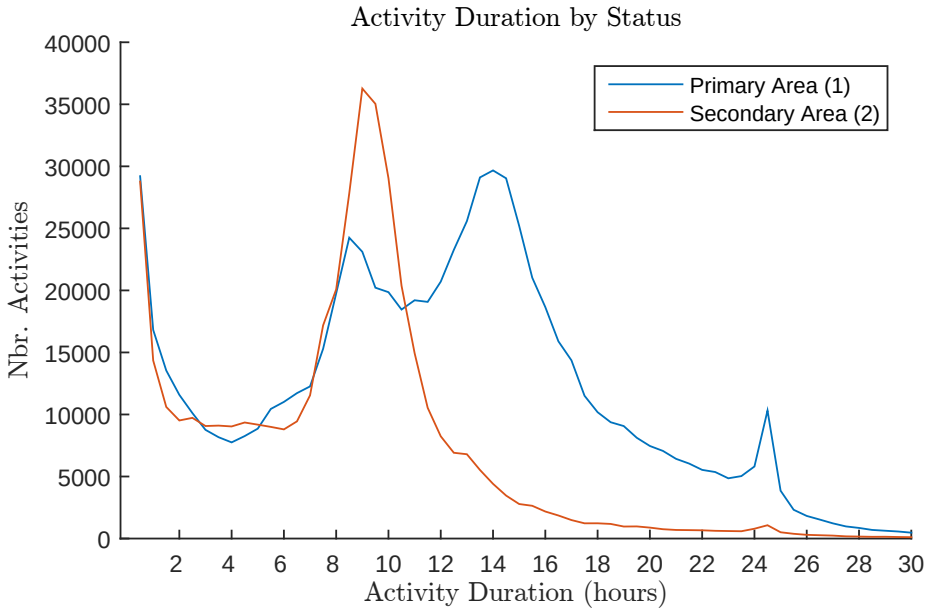


Figure 2: Distribution of Activity Duration

statuses. The distribution of status 1, aligned with users' primary area, is characterized by three peaks. The highest peak, around 14 hours, is associated with spending the night at home. The second peak, around 8 hours, is aligned with the 8 hours working day. The third peak is likely associated with users making a single journey on a given day, followed by another journey at a similar time the next day. Status 2, aligned with users' secondary area, is characterized by a dominant peak also around 8 to 9 hours. This reflects the fact that the secondary area corresponds predominantly to the area in which users work. Intervals shorter than 1 hour make-up an important proportion of activities for both statuses. Overall, these results suggest that each status is associated with a distinct mixture of activities. The curves also appear to support that user's home areas are primarily associated with area 1 and that work areas are primarily associated with area 2.

3.2.2. Cluster Analysis

The longitudinal activity sequence of each user is discretized into 696 1-hour time bins (24 hours $\times$ 29 days). All activity statuses above the fourth geographical area are aggregated into a single activity status. Hence, each users' longitudinal activity pattern is

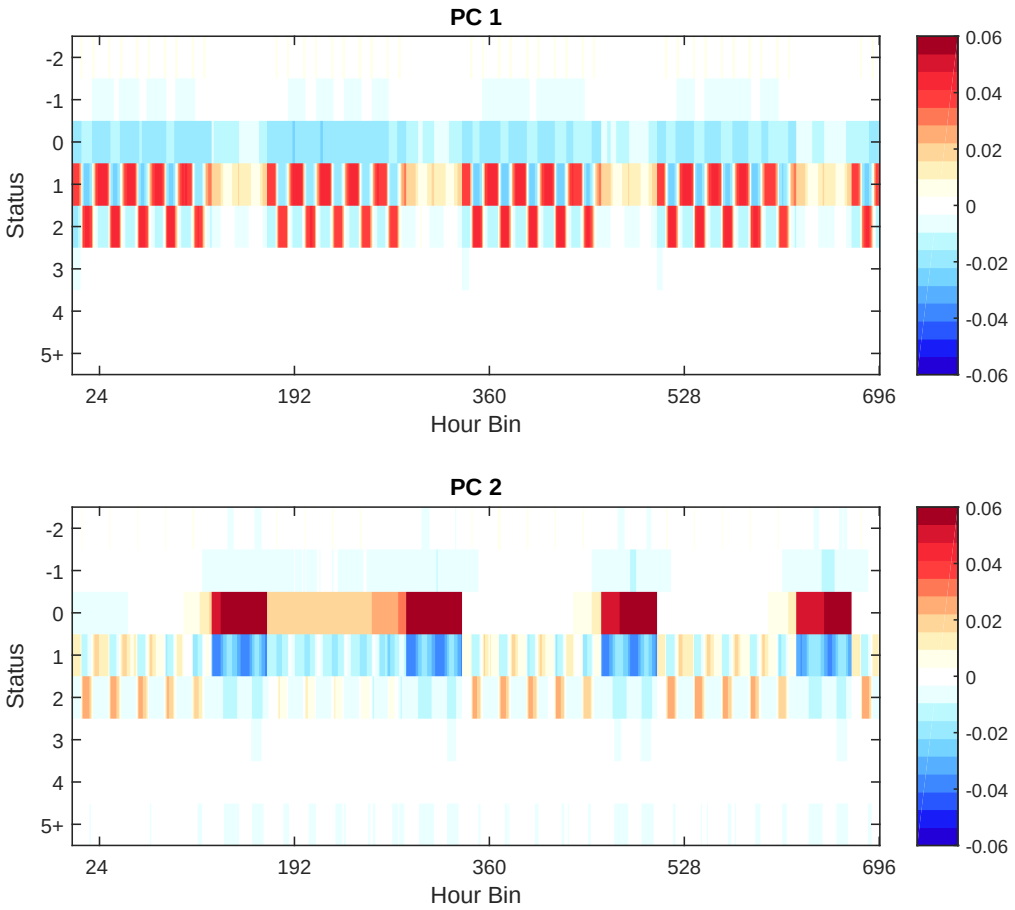


Figure 3: Principal Components 1 and 2

represented by a vector of 696 elements, each of which can take one of 8 possible statuses (PT journey, discontinuity, no inference, areas 1, 2, 3, 4, and areas 5+). Principal component analysis is applied to these 33,026 sequences as described in section 2.4.

Figure 3 illustrates the first two PCs resulting from this analysis. Each principal component represents a different component of longitudinal behavior observed across multiple user-sequences. This common component of longitudinal behavior is summarized by the PC as a pattern of correlation, such that status-hours often observed to co-occur within the same sequence are assigned a high weight in the corresponding PC. In Figure 3, the color of each status-hour represents its weight.

For the first PC, the red status hours are those associated with being observed at area 2 during weekdays and at area 1 during weeknights. This indicates that individuals observed at area 2 on a given weekday were very often also observed at area 2 on other weekdays and at area 1 on weeknights. This is intuitive; a user going to work on one day correlates with this user going to work on other days and returning home at night.

Status-hours with a negative weight are those which correlate negatively with other status-hours. For example, in PC 2, week-end hours at status 0 were all correlated with each other, while they were negatively correlated with week-end hours at area 1. Intuitively, observing no travel for a user on a given week-end correlates with this user not being observed at location 1 on future week-ends. This relationship also holds backwards: observing a user at area 1 on a given week-end correlates with not failing to observe this user on future weekends.

The degree to which each PC is reflected in the sequence of a given user is measured from the projection described by equation 5. For example, the first principal component illustrated in Figure 3 would have a high positive weight for the activity pattern of users who commuted on all week-days. This dimensionality reduction approach is used to extract clustering variables for each sequence. In order to identify the number of PCs to use, 20 bootstrap subsamples of 10,000 users are defined from the primary sample, and PCA is applied to each subsample. To identify the stability of a given PC, the average correlation between pairs of matching components i, $\overline{\rho}_i$, is used (Equation 7).

$$\overline{\rho}_i = \frac{1}{||P||} \sum_{k,l \in P} |\mathbf{v}_{i,k}^T \cdot \mathbf{v}_{i,l}| \tag{7}$$

where $P = \{(k,l) : k,l \in \mathbb{N} \wedge k < l \leq 20\}$ denotes the set of 190 sample pairs defined from the 20 sub-samples, and $\mathbf{v}_{i,k}$ denotes the i^{th} principal component of sample k. Figure 4 shows the stability of the first 13 principal components. The x-axis indicates the principal component number and the y-axis shows $\overline{\rho}_i$. As seen from the figure, the average correlation is above 0.9 for the first 8 principal components, indicating high stability, and drops below 0.8 for the following PCs. Hence, the projections of user-sequence onto the first 8 PCs are used as input variables to the k-means clustering process.

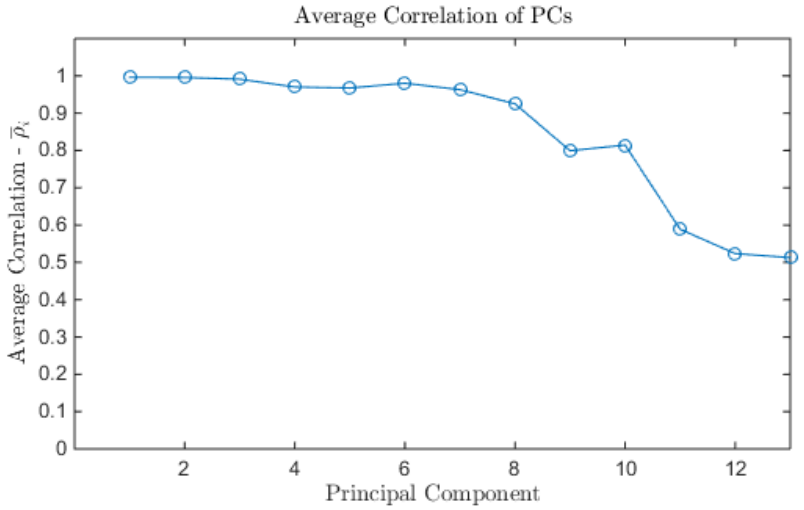


Figure 4: PC Stability from 20 Bootstrapped Samples

Figure 5: DB Index for 8 Principal Components

The k-means++ initialization approach is used and 150 replications are ran for each k evaluated to avoid local-optimum solutions. Figure 5 shows the DB-index, a measure of internal cluster fit based on the ratio of the within cluster distances to across cluster distances (Davies & Bouldin, 1979), for values of k between 2 and 20. The distribution shows that the clustering solutions are most compact for $k = 4$ and $k = 11$. The $k = 11$ solution is chosen as it provides a more detailed segmentation of users.

3.3. Results

As the 33,026 users are clustered based on the underlying structure of their 29-day activity sequence, a distinct sequence structure is associated with each one of the 11 clusters identified. The pattern associated with each group is visually reflected in the sequences it contains. Figure 6 illustrates the sequence pattern of the first cluster. The figure shows the activity sequence of 500 users randomly selected from all users assigned to cluster 1. Each row in the figure corresponds to a single user-sequence, and all sequences are aligned with respect to time, shown on the x-axis. The activity status of each hour of the 29-day period is symbolized by its color. The figure reveals that the sequence structure of cluster 1 is characterized by two dominant attributes: clear working days reflected by the vertical green bands delineating weekdays, and reduced transit travel during weekends reflected by the vertical white bands delineating Saturdays and Sundays. A similar representation of each cluster is presented in Figure 7. Clusters are organized in 4 sets according to their structure similarity: working day, homebound, complex activity pattern, and interrupted pattern. The characteristics associated with each cluster are summarized in the following 4 sections.

In line with the close connection between short-term activity patterns and longer-term factors such as occupation status, residential choice and vehicle ownership, similarities in sequence structure are likely to be associated with similarities in user socio-demographic.

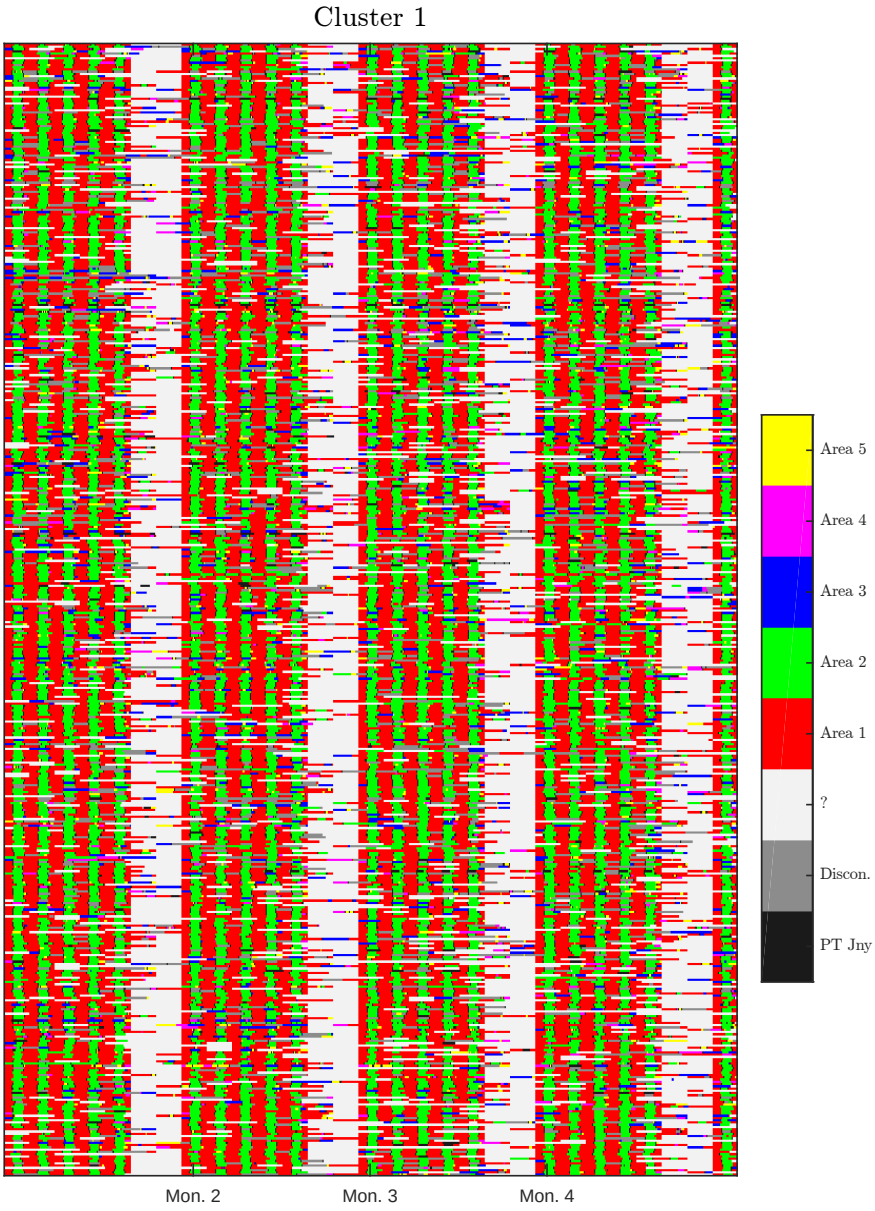
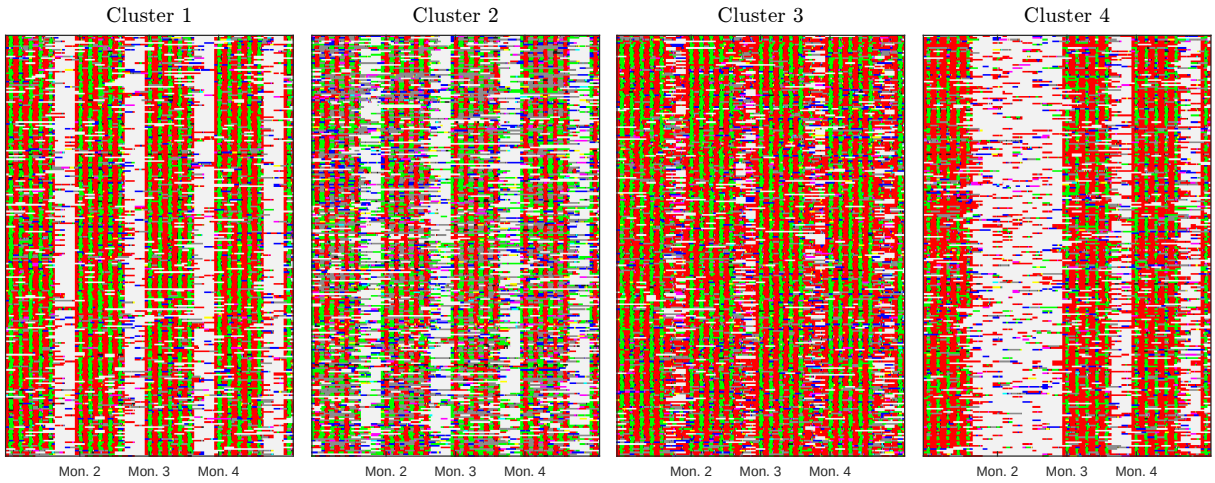
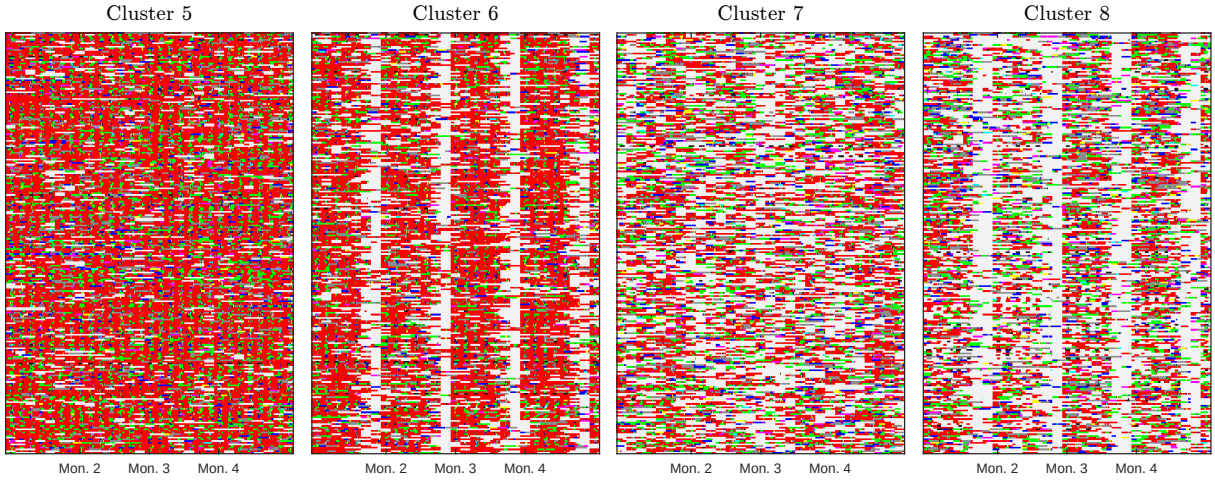


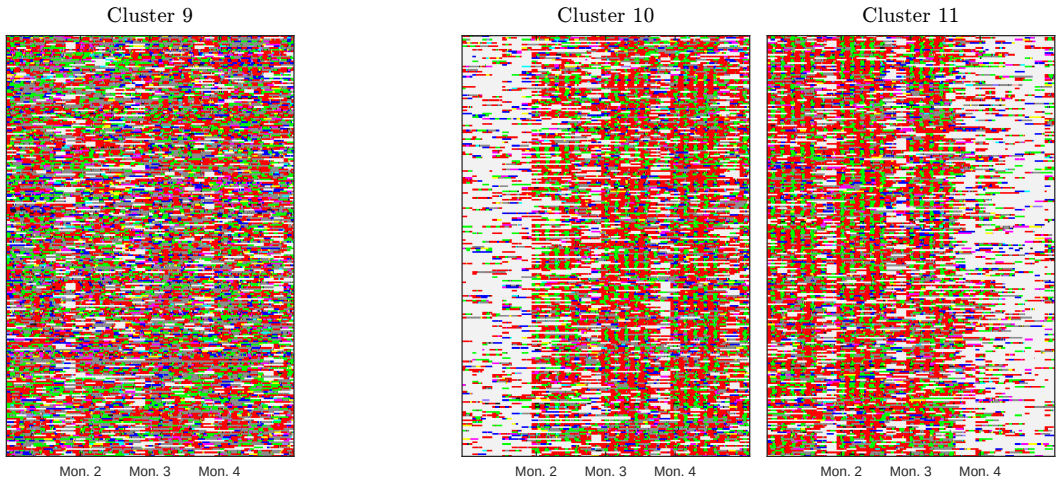
Figure 6: Cluster 1 Sequence Structure



(a) Working Day Clusters



(b) Homebound Clusters



(c) Complex Activity Pattern Cluster

(d) Interrupted Pattern Clusters

Figure 7: Sequence Structure Associated with each Cluster

In order to evaluate these associations, the 1,973 LTDS respondents (for whom both demographic data and smart card data is available) are assigned to one of the 11 clusters based on the k-means centroid they lie nearest to. The set of cluster centroids resulting from the analysis define a voronoi partition according to which the sequence of LTDS users is classified.

3.3.1. Working Day Clusters

The sequence characteristic shared by all clusters in this set is the distinctive working days reflected by the vertical lines delineating weekdays as seen in Figure 7a.

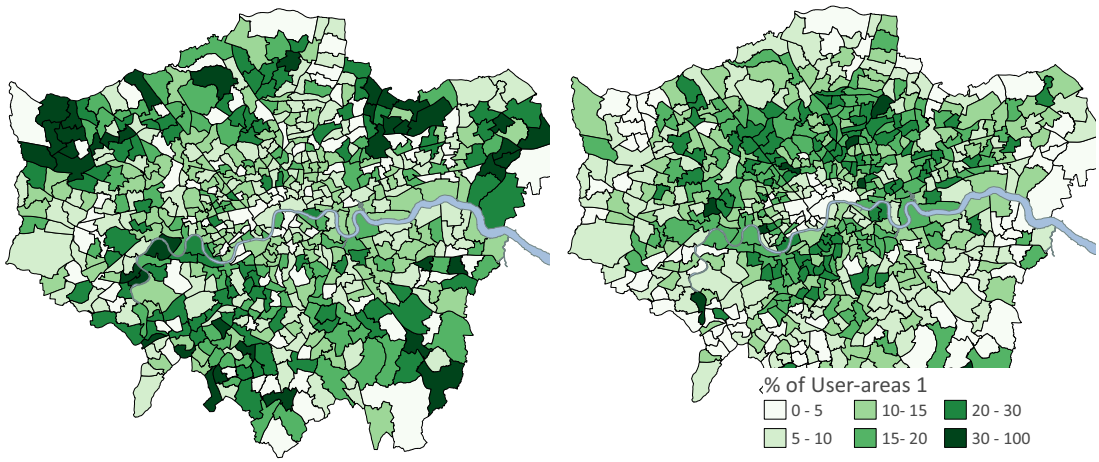
Cluster 1: User-sequences in cluster 1 are characterized by clear working days on weekdays and reduced travel on week-ends. As summarized in Table 3, this cluster accounts for 15% of frequent users in the primary sample. On average, users in this cluster have the early median first journey departure time on weekdays, and visited the smallest number of distinct locations. Odds-ratio analysis reveals that LTDS respondents assigned to this cluster are 5.3 times more likely to be full-time employed, have the highest median household income, and are 3.5 times more likely to live in a household with access to a vehicle than users assigned to other clusters. These characteristics suggests that users in cluster 1 use public transportation for work purposes exclusively. In line with this observation and with Ortega-Tong (2013) users in this group could be referred to as exclusive commuters.

Cluster 2: User-sequences in cluster 2 share sequence attributes similar to those of cluster 1: distinct working days and limited travel on weekends. However, unlike cluster 1, users in cluster 2 spend working hours in their primary area (red vertical band delineating workdays) and mornings and evening in their secondary area. This inversion may be caused by two different patterns. First, users who frequently spend the night outside their home, for example at a partner's home, are inferred to spend more time in the work area than in any single home area. Hence, their primary area, the area in which they were inferred spending most time, is associated with work. Alternatively, journey sequence discontinuities between the last trip of the day and the first trip of the following day result in failure to infer location over night. For instance, the travel sequence of individuals who occasionally use non-public modes (e.g. car-pooling or taxi) for the first stage of their morning commute would be characterized by such discontinuities. The demographic attributes of users assigned to this cluster are largely similar to those of exclusive commuters, with strong positive associations with full-time employment, higher household income, and car access. The differences in travel patterns of cluster 1 and 2, may relate, in part, to age differences between the two clusters. While the first cluster is most strongly associated with individuals in their late thirties and early forties, the second is most strongly associated with younger users in their late twenties and early thirties.

Cluster 3: Users in this cluster are associated with distinct working days and frequent travel to non-secondary areas on weekends. Like the first two clusters, cluster 3 is associated with early departure times on week-days, but on average users in this group visit a higher number of distinct locations and complete more journeys. This suggests the cluster includes passengers who use PT for work but also for other purposes. This is supported by the demographics of users assigned to the cluster, who are primarily employed full-time, over twice as likely to be in their twenties than users in other clusters, and associated with median household income of £25,000 to 35,000 pounds. Users in this group are also 1.4 times more likely to live in households without access to a vehicle.

Cluster 4: The dominant feature characterizing cluster 4 is a marked reduction in travel over the second week of the analysis period, combined with clear working-days, early first journey start time on weekdays and decreased travel on weekends for the remaining three weeks. The second week of the analysis period corresponds to the school

half-term in London. The analysis of Oyster card types reveals that under-18 student cards were 20 times more likely to be assigned to this group. Additionally, the LTDS sample which includes no user under 18 years old reveals another interesting association. LTDS respondents assigned to cluster 4 were over 2 times more likely to belong to a household with children and to be between 40 and 45 years old. The cluster is also positively correlated with part-time employments. Overall, these characteristics suggest that users in cluster 4 are not only pupils but also parents of pupils on holiday during the half-term.



(a) Cluster 1 (b) Cluster 3

Figure 8: Geographical Distribution of User-Areas 1

Figure 8 contrasts the distribution of user-areas 1 for clusters 1 and 3. Each user's primary area (area 1) is assigned to the zone which contains the stop or station most used by the user. For each zone, the percentage of all stops or stations associated with users in a given cluster is indicated on the map. For example, dark green zones on the left map indicate that a high percentage of primary user-areas in those zones were associated with users in cluster 1. As area 1 is most likely associated with the home location of users in clusters 1 and 3, the figure suggests that users in cluster 1 are more likely to live in outer London, while users in cluster 3 are more likely to live in periphery of the urban core. This trend is consistent with the patterns identified solely from users' activity sequences: passengers in cluster 1 use PT more occasionally on weekends, while passengers in cluster 3 use it more frequently likely for non-work related activities.

Table 3: Descriptive Attributes of the Clusters

	Working Day				Homebound				Complex	Interrupted	
Travel Pattern Attributes											
Attribute	1	2	3	4	5	6	7	8	9	10	11
% of 33,026 Frequent Users	15	6	14	10	10	7	9	9	9	5	5
Med. Nbr. Journeys (weekdays)	43	43	52	37	59	44	27	31	53	36	37
Med. Nbr. Journeys (weekends)	2	5	15	4	20	8	11	3	16	11	9
Area 1 (avg. % of time inferred [a])	35%	24%	45%	36%	56%	49%	29%	23%	29%	32%	32%
Area 2 (avg. % of time inferred[a])	20%	17%	22%	13%	13%	8%	10%	9%	15%	12%	12%
Avg. Nbr. of Distinct Locations	6.4	8.0	10.5	6.8	12.1	9.0	9.7	8.1	13.1	9.9	9.7
Avg. Med. Departure Time (week)	8:04	8:26	7:58	8:27	10:08	10:30	11:46	10:43	9:40	9:34	9:24
Avg. Nbr. Weekend Days Traveled	2.2	3.5	6.0	2.8	7.1	3.5	5.4	2.3	6.4	4.7	4.1
Avg. Nbr. Weekdays Traveled	19.2	18.9	20.1	16.1	19.2	17.7	12.5	15.0	19.2	14.7	15.1
Socio-Demographic Attributes											
Age	37	33	35	42	56	64	52	48	35	40	36
Emp. Full-Rime (%)	78	83	69	50	20	14	26	34	46	43	54
Annual Household Income (£1,000)	50-75	50-75	25-35	35-45	10-15	10-15	15-20	25-35	20-25	20-25	25-35
Household with Children (%)	27	25	20	41	18	13	18	27	19	24	18
Vehicle (%)	70	72	38	56	26	28	34	55	24	42	41

[a] Percentage of inferred time which the user spent in given status over 29 days. Days on which no journeys are observed are excluded.

3.3.2. Homebound Clusters

The users assigned to clusters in this set are characterized by a high proportion of time spent in the home area. This is reflected either by the high percentage of time periods spent in the primary area for users in clusters 5 and 6, or by the relatively low number of days traveled for users in clusters 7 and 8.

Clusters 5: This cluster is characterized by a high proportion of time spent in the primary area, and by a high number of short activities scattered equally throughout both weekdays and weekends. The average first journey departure time for users in this group is later than average, and users in this group tend to visit a high number of distinct locations. Users in this group are 4.5 time more likely to be unable to work due to illness or disability than users in other clusters, and 2.7 times more likely to be retired. This is reflected by the high average age, low median income bracket, and low rate of full-time employment associated with this cluster.

Cluster 6: Individuals in this cluster were inferred to spend the highest percentage of time in their primary area on average. Like cluster 5, their activity sequence is characterized by short activities. They completed fewer journeys than users in cluster 5, with a marked reduction of travel on Sundays (illustrated by the white vertical bands in Figure 7. Cluster 6 is most strongly associated with retired users, (with users in this group being 5.6 times more likely to be retired), but also associated with individuals unable to work due to disabilities. This is reflected by an average age of 64 for users in this group, the highest among all clusters.

Cluster 7: As for the previous two clusters, sequences in this cluster are characterized by shorter activities and late departure times. However, users in this group traveled on fewer days than those in cluster 5 and 6, on both weekdays and weekends. This results in a higher number of periods during which no activity status inference can be made. The socio-demographic association for this group are not as marked as for clusters 5 and 6, with users in this group being around twice as likely to be either retired or disabled. All three clusters are also associated with low rates of vehicle access.

Cluster 8: This group is characterized by no distinct working days, a low number of journeys and days traveled, and late first journey departure time. Most journeys completed by users in this group are concentrated on weekdays, but are not commuting journeys as indicated by the late average first journey departure time. These characteristics may suggest the cluster includes homebound passengers who use public transport for non-commute journeys conducted on weekdays (e.g business related journeys). The cluster is significantly associated with self-employed, retired and stay-home LTDS respondents. Unlike other homebound clusters, over 55% of users in this group live in households with access to at least 1 vehicle. This might explain why this group has the lowest average number journeys and days traveled among homebound clusters.

3.3.3. Complex Activity Pattern Cluster

Cluster 9: This cluster includes users who traveled on almost every day of the period of analysis, weekdays and weekends. Their journey sequence are characterized by multiple discontinuities, indicating many intervals during which a non-PT journey was completed. In addition, these users completed a high number of activities per day and visited a higher number of locations. Consistent with this pattern, this cluster is most strongly associated with individuals in their twenties. Users in this group are twice as likely to be between 20 and 29. They are also over 50% more likely to live in single adult

households or households classified as 'other', which includes multiple adults sharing housing. The median household income bracket of users in this group is also lower than average. These characteristics indicate the cluster is composed of users who entwine public transport journeys within a complex activity schedule not primarily driven by full-time employment, likely in combination to other modes such as walking and cycling.

3.3.4. Interrupted Pattern Clusters

Clusters 10 and 11: Clusters 10 and 11 are marked by reduced travel on the first and last week, respectively, of the period of analysis. Unlike cluster 4, the activity sequence structure observed on the remaining weeks varies significantly across individuals. As a result no clear socio-demographic trends are observed for these two clusters. This is likely due to the fact that the only characteristic shared by members of each cluster is the change in activity pattern for the first or last week. Unlike cluster 4, this change does not correlate to meaningful demographic attributes. The change in travel pattern observed on the first and last week may also reflect card churn of the finite analysis period considered. For example, cards may have come in use during the first week or may have stopped being used after the third week.

3.3.5. Multivariate Demographic Analysis

The odds-ratios described above reveal demographic trends for each cluster consistent across multiple demographic variables. In order to verify that these trends are not the result of correlation between demographic variables, it is necessary to evaluate the associations between clusters and multiple demographic variables simultaneously. For this purpose, we develop a multinomial logit model to explain cluster membership as a function of individual demographic attributes. Given the limited LTDS sample size, clusters with similar attributes and interpretation are grouped together to limit the number of parameters to estimate. The model coefficients are estimated via maximum likelihood estimation using BIOGEME (Bierlaire, 2003).

The model results, summarized in Table 4, are overall consistent with the associations observed from the odds ratios analysis. The coefficient of each variable indicates the direction and the magnitude of its association to each cluster, controlling for the effect of other variables included in the model. The explanatory power of socio-demographics on cluster membership is limited, as indicated by the adjusted rho squared of 0.206, but the trends revealed by the model coefficients are aligned with the travel characteristics of each cluster.

In addition to confirming the trends observed through the odds ratio analysis, the model reveals the following patterns. Income remains positively associated with clusters 1 and 2 even after controlling for employment status, suggesting that the observed association with income is not only a result of correlation with employment status. In contrast, controlling for full-time employment, cluster 3 is negatively associated with income. This suggest that full-time employees in cluster 3 tend to earn less than full-time employees in other clusters. Along with the positive, but insignificant, coefficient for being younger than 35, this further supports the hypothesis that cluster 3 is composed of younger, mid-range income users.

Table 4: Cluster Membership Model - Coefficient Estimates

	Working Day			**Homebound**		**Complex**	**Interrupted**
Clusters	1, 2	3	4	5,6,7	8	9	10,11
ASC	**-0.50****	0.08	**-0.75*****	**1.61*****	-0.08	-0.14	
Age < 35	0.05	0.13	**-0.53***	**-0.43****	**-0.69*****	**0.47***	
HH Car	**0.97****	-0.17	0.39	-0.18	**0.56*****	**-0.60****	
Income (£1000)	**0.0058***	**-0.0066***	0.0014	**-0.010*****	0.0018	**-0.0078***	
Retired or Disabled	**-1.3*****	**-1.7*****	**-1.0****	**0.95*****	**0.68****	0.21	
Full-Time Employed	**1.0*****	**0.82*****		**-0.53*****			
Kids			**0.64*****				
Self-Employed or Stay-Home					**0.52****		
Unemployed						0.35	

* indicates significance at the 90% confidence level (t-statistic ≥ 1.64)

** indicates significance at the 95% confidence level (t-statistic ≥ 1.96)

*** indicates significance at the 99% confidence level (t-statistic ≥ 2.58)

$n = 1968$, $k = 36$

$LL_0 = -3829.55$, $LL_{cst} = -3459.51$, $LL = -3005.98$

$\overline{\rho}^2 = 0.206$

The model confirms that cluster 4 is positively associated with users living in households with children. The strong associations between the low income of users in clusters 5, 6, and 7 remains significant even after controlling for retired or disabled occupation statuses. The model also confirms that self-employed, stay-home and retired users with access to cars are more likely to be in cluster 8. Controlling for the low income of users in cluster 9 reveals that the association between unemployment and this cluster is positive, but insignificant.

3.4. Stability of Sequence Patterns

In order to evaluate the validity of the approach, it is useful to examine how the identified clusters generalize across different time periods. When comparing user clusters defined from data extracted on the same week of two different years, Ortega-Tong (2013) identified important temporal stability issues. She found that different clusters emerged from the two time periods, despite no significant changes in the passenger population. To examine this issue, a random sample of 5,724 frequent users is extracted for the period between October 20th and November 19th 2014. Similarly to the February-March period, this period spans 29 days and is aligned with the school half-term on the second week.

First, the analysis applied to the February sample is independently applied to the October sample. This results in 11 clusters defined exclusively from the user-sequences of the October period. Second, October sequences are classified with respect to the February Clusters. This is done by projecting October user-sequences onto the February principal components (using equation 5) and then mapping the resulting projection to the nearest February cluster centroids. This indicates how a user-sequence observed in October would have been classified with respect to the February clusters had it been observed during the Feburary period.

Hence, each user-sequence in the October sample is associated to two clusters: one from the October clustering, and another from the February clustering solution. Stability of the clusters can be evaluated from the percentage of user-sequences which are assigned to matching clusters across both partitions. Table 5 summarizes the percentage of users in October clusters assigned to each February cluster. For example, considering the intersection of the second row and the first column, 1.1% of all user-sequences assigned to 1_{Oct} would also have been assigned to cluster 2_{Feb} had they been observed in February. Diagonal values indicate the degree of cluster stability. For instance, 99.4% of users in cluster 6_{Oct} were also classified in the equivalent February cluster, indicating that cluster 6 is 99.4% stable. Overall, 91% of all frequent users in the October sample were allocated to the same cluster across both periods. This high overlap indicates that the clusters are stable over different periods of analysis.

Table 5: February-October Clustering Overlap

		Oct. Users Assigned to Oct. Clusters (%)										
	Cluster	1_{Oct}	2_{Oct}	3_{Oct}	4_{Oct}	5_{Oct}	6_{Oct}	7_{Oct}	8_{Oct}	9_{Oct}	10_{Oct}	11_{Oct}
Oct. Users Assigned to Feb. Clusters	1_{Feb}	95.7	0.1	0.0	0.0	1.4	0.4	0.0	0.0	1.3	0.0	0.2
	2_{Feb}	1.1	95.8	0.2	0.0	0.0	0.0	0.0	0.3	0.9	0.3	0.0
	3_{Feb}	0.0	2.0	96.0	1.1	0.9	0.0	0.0	1.4	0.2	5.3	0.0
	4_{Feb}	0.6	0.1	0.0	87.9	1.2	0.0	0.3	0.0	0.4	0.3	2.9
	5_{Feb}	1.1	1.6	3.3	0.3	81.0	0.0	0.3	1.7	5.0	1.2	0.0
	6_{Feb}	0.0	0.1	0.0	0.6	2.3	99.4	0.0	0.0	0.0	0.3	3.4
	7_{Feb}	0.0	0.0	0.0	0.0	1.4	0.0	98.7	0.0	0.4	0.3	0.4
	8_{Feb}	0.9	0.0	0.0	0.3	1.4	0.0	0.0	91.0	6.1	0.5	0.9
	9_{Feb}	0.0	0.1	0.3	9.9	0.5	0.2	0.0	4.5	78.1	3.3	7.9
	10_{Feb}	0.2	0.1	0.2	0.0	0.5	0.0	0.3	1.0	0.7	88.3	1.8
	11_{Feb}	0.4	0.0	0.0	0.0	9.5	0.0	0.3	0.0	6.8	0.0	82.6

4. Conclusion

The contributions of this research can be summarized in two parts. Our methodological contributions reside in the proposed representation of longitudinal activity sequences and in the synthesis of statistical approaches allowing for the analysis of these sequences. By representing each individual as an ordered sequence of activities spanning multiple weeks, we capture important information relating to the temporal organization of journeys and activities typically lost through the scalar aggregation of the passenger's journeys. This information is leveraged by clustering user sequences with respect to the structure of longitudinal activity sequences. Principal component analysis is used to extract common elements of structure across all frequent-user sequences, and the linear projection of each sequence vector onto the most important principal components is used as input to the cluster analysis.

Our empirical contributions emerge from the large scale application of this methodology to London's PT network. The application reveals 11 clusters of users, each associated with a distinct sequence structure. The structure of these clusters suggest that while conventional working days are an important element of structure for many users (clusters 1 to 4), they do not structure the activity sequence of over 40% of frequent users (clusters 5 to 9). The implications of this finding may be especially important in aspects of transit operations typically focused on planning for 'typical commuters'. Additionally, the results demonstrate the benefit of considering the activity sequence of users over multiple weeks. As illustrated by cluster 4, changes in behavior from week to week can provide insight into the constraints driving individuals' activity pattern, and

hence into the demographic attributes of these individuals. This result also suggests that it may be informative to revisit studies on multiweek travel behavior typically focused on variability within days (e.g. Hanson & Huff, 1988, or Axhausen et al., 2002) to consider longer analysis units than the day. Overall, the activity sequence structure of all but two clusters, derived from traces of travel alone, are associated with distinct and significant socio-demographic characteristics related to occupation status, age, household income and composition, and household vehicle access. These results may inform the practice of transit agencies in a number of ways. With respect to real-time service, the information disseminated to users could be tailored by cluster. For example, knowing that users in clusters 1 to 4 are likely to travel between their primary and secondary area on weekday mornings and evenings, preventive service alerts could be sent to passenger's whose identified commute journeys are expected to be disrupted. With respect to medium-term planning, it would be possible to customize the services provided in different stations based on the station population. Clusters could be used to optimize commercial opportunities associated with in-station and in-vehicle advertising and retail space. For example, retailers targeting elderly customers may be willing to pay a premium for commercial spaces located in stations primarily used by users in homebound clusters, especially cluster 6. With respect to long-term planning and network development, cost-benefit analysis could be designed to accommodate differences between the identified clusters. For example, full-time employees in clusters 1 and 2 likely have different value of time than homebound users in clusters 5 and 6. Such differences may inform how potential projects could be prioritized.

A number of limitations to the research presented in this paper provide basis for further investigation. First, while the case study was implemented using data covering a 4-week period, the effect of analysis period length on the observed clusters should be investigated. Second, the analysis presented focused on frequent public transport users. While these users account for over 70% of journeys completed in the London network, valuable insight may be obtained from analysis of the heterogeneity among non-recurrent and occasional users. The travel survey sample available for the socio-demographic analysis contained a limited number of users and demographic variables. Additional studies could further investigate the relationship between multi-week activity sequence and demographic attributes beyond the attributes considered in this study. For example, an online survey distributed by email to registered smart card users could provide demographic data for a larger number of users.

Finally, the results of the study also hint at interesting future research questions. In line with the strength of the demographic associations observed for certain clusters, future research could evaluate the value of smart card data to predict certain demographic characteristics of users. Additionally, having demonstrated the stability of the clusters in section 3.4, tracking the evolution of individuals across clusters over multiple years may provide interesting opportunities to examine life-stage changes.

Acknowledgment

The authors would like to thank Transport for London for providing the data used in this study and many useful discussions and for the generous support provided for this research.

References

Axhausen, K. W., Zimmermann, A., Schnfelder, S., Rindsfser, G., & Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, *29*, 95–124.

Bierlaire, M. (2003). BIOGEME : A free package for the estimation of discrete choice models,. In *Proceedings of the 3rd Swiss Transportation Research Conference*. Ascona, Switzerland.

Chu, K., & Chapleau, R. (2010). Augmenting Transit Trip Characterization and Travel Behavior Comprehension. *Transportation Research Record: Journal of the Transportation Research Board*, *2183*, 29–40.

Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*, 224–227.

Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*, 7–24.

Eagle, N., & Pentland, A. (2009). Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*, 1057–1066.

El Mahrsi, M., Come, E., Baro, J., & Oukhellou, L. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. In *Proceedings of the 3rd International Workshop on Urban Computing*. New York City, USA.

Gordon, J., Koutsopoulos, H., Wilson, N., & Attanucci, J. (2013). Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board*, (pp. 17–24).

Goulet-Langlois, G. (2015). *Exploring Regularity and Structure in Travel Behavior Using Smart Card Data*. Thesis MIT.

Hagerstraand, T. (1970). What About People in Regional Science? *Papers in Regional Science*, *24*, 7–24.

Halvorsen, A. (2015). *Improving Transit Demand Management with Smart Card Data: General Framework and Applications*. Thesis MIT.

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

Jiang, S., Ferreira, J., & Gonzlez, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, *25*, 478–510.

Kieu, L., Bhaskar, A., & Chung, E. (2014). Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, *PP*, 1–12.

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: a data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179–191.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*, 1–20.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders travel patterns. *Transportation Research Part C: Emerging Technologies*, *36*, 1–12.

Manning, C. D., Raghavan, P., Schütze, H. et al. (2008). *Introduction to information retrieval* volume 1. Cambridge university press Cambridge.

Morency, C., Trpanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, *14*, 193–203.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9–18.

Ortega-Tong, M. A. (2013). *Classification of London's public transport users using smart card data*. Thesis MIT. URL: `http://dspace.mit.edu/handle/1721.1/82844`.

Pelletier, M.-P., Trepanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*, 557 – 568. URL: `http://www.sciencedirect.com/science/article/pii/S0968090X1000166X`.

Szumilas, M. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*, 227–229.

Transport for London (2014). *Travel in London - Report 7*. Technical Report London. URL: `https://tfl.gov.uk/cdn/static/cms/documents/travel-in-london-report-7.pdf`.

Transport for London (2015a). Buses. tfl.gov.uk/corporate/about-tfl/what-we-do/buses. Online; accessed 2015-07-13.

Transport for London (2015b). Facts and figures. tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/facts-and-figures. Online; accessed 2015-07-13.