EXPLORING THE RELATIONSHIP BETWEEN TRAVEL PATTERN AND SOCIAL-DEMOGRAPHICS USING SMART CARD DATA AND HOUSEHOLD SURVEY

Yang Zhang [1, *], Tao Cheng [1], Nilufer Sari Aslam [1]

[1] SpaceTimeLab for Big Data Analytics, Dept. of Civil, Environmental & Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT – (yang.zhang.16, tao.cheng, n.aslam.11)@ucl.ac.uk

Commission IV, WG IV/10

KEY WORDS: Smart card data, Travel pattern, Social-demographics, Household survey, Relationships, Clustering

ABSTRACT:

Understanding social-demographics of passengers in public transit systems is significant for transportation operators and city planners in many real applications, such as forecasting travel demand and providing personalised transportation service. This paper develops an entire framework to analyse the relationship between passengers' movement patterns and social-demographics by using smart card (SC) data with a household survey. The study first extracts various novel travel features of passengers from SC data, including spatial, temporal, travel mode and travel frequency features, to identify long-term travel patterns and their seasonality, for the in-depth understanding of 'how' people travel in cities. Leveraging household survey data, we then classify passengers into several groups based on their social-demographic characteristics, such as age, and working status, to identify the homogeneity of travellers for understanding 'who' travels using public transit. Finally, we explore the significant relationships between the travel patterns and demographic clusters. This research reveals explicit semantic explanations of 'why' passengers exhibit these travel patterns.

1. INTRODUCTION

The portable and durable smart card (SC) has been widely used for paying for public transport, such as London's Oyster card (Lathia et al. 2011), Beijing's BMAC card (Yuan et al. 2013), Singapore's SC for MRT service (Sun et al. 2012). SC that stores massive trip transactions of passengers has been drawn a lot of attention in various existing literature (Pelletier et al. 2011). The application domains include mobility pattern analysis (Shi et al. 2014), traffic congestion pattern analysis (Ceapa et al. 2012), home/work location estimation (Sari Aslam et al. 2018), and activity detection (Nassir et al. 2015).

Overwhelming amounts of SC data also provides a promising way to mine mobility patterns for better transport planning and service provision. However, it lacks the social-demographic information of passengers to further explore 'who are the card carriers', 'why they behaved differently' and 'what factors affect their behaviours', which are crucial to better understand the users' travel demand and mobility patterns. Fortunately, leveraging household survey data, it might further explore the relationship between human travel patterns and their social-demographic roles (Zhang et al. 2018; Zhang et al. 2019), which can help operators make better transportation planning and provide passengers with more personalised services.

In this paper, an entire framework is proposed to explore 'how', 'who' and 'why' travels in the PT:
'**How**': We aim to establish an elaborate travel feature extraction process to classify passengers' long-term travel behaviours by using smart card data. Users are then clustered into several groups indicating different travel patterns for the in-depth understanding of '**when**', '**where**' and '**how often**' people travel by '**which travel mode**' in cities.

'**Who**': Leveraging travel survey data, passengers can be also categorised into different demographic groups based on individual or household demographic variables, including age, working status, main occupation, car ownership, household income. This analysis investigates who usually travel via public transit (e.g., bus or underground).

'**Why**': In this step, we link the passengers' travel pattern with the demographic group to find the significant linkages between the two clustering results. This study provides a better understanding and semantic explanations of passengers' movement patterns.

This paper is organised as follows. Section 2 introduces the dataset used in this study. Section 3 illustrates the methodologies to analyse the travel patterns, social-demographics groups and their relationships. Then, Section 4 describes a case study of London, UK. Finally, the conclusions, limitations and future work are discussed in Section 5.

2. DATASET

2.1 London's Oyster Card Data

The SC data used in this study is a sample of Oyster Card transaction records in London, UK, during the full year of 2012. There are two types of SCD, one from the tube system and the other from the bus system. A transaction is recorded automatically when a passenger taps in/out at a tube station or boards at a bus stop. Summarily, the entire dataset contains

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-2-W13-1375-2019

around 2.18 million journeys made by 9708 passengers, made up of 33.7% tube journeys and 66.3% bus journeys. Each SCD record consists of the following fields: (1) Oyster card ID (encrypted), (2) transaction date, (3) start time, (4) end time, (5) boarding station, (6) exiting station, (7) journey mode (bus or tube). Note that in bus trip records, the boarding station indicates the bus line number but not precise locations, and the exit station and end time are unavailable.

2.2 London Travel Demand Survey Data

London Travel Demand Survey (LTDS) is a continuous survey based on the household for collecting individual or household demographic, social-economic and travel-related information. Every year, around 8000 randomly selected households undertake the LTDS annually. All household members aged 5 and over are required to complete the questionnaire. The information provided in LTDS includes: (1) Oyster card ID (2) PAGEI: Age, (3) PMANAGER: If a manager, (4) HCVN: Number of vehicles in total owned, (5) HINCOMEI: household income, (6) PWKSTAT: working status, (7) POCCUPA: occupation type, (8) POFWK: weekly work frequency, (9) PLENN: approximate daily commuting distance, (10) PFRCARD: the frequency of using car as a driver, (11) PFRCARP: the frequency of using car as a passenger. Among them, 'PAGEI' and 'PLENN' are continuous variables, and others are categorical variables.

3. METHODOLOGY

3.1 Framework

This article aims to explore 'how' (including 'when' and 'where'), 'who' and 'why' travel in public transit using smart card data and household survey. For such purpose, the proposed framework should be capable of:

- Step 1: Identify long-term travel patterns by using smart card data, telling how passengers travel in the city.
- Step 2: Identify social-demographic groups of travellers, understanding who travels.
- Step 3: Define more significant relationships between travel patterns and social-demographic groups.

The framework is illustrated in Figure 1.

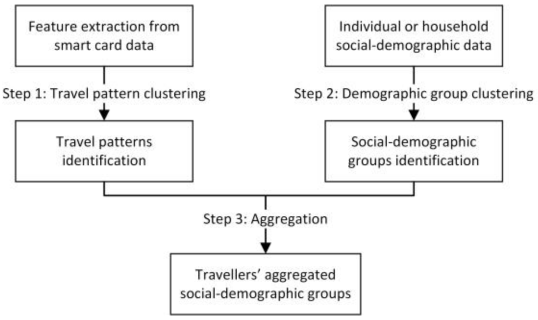

Feature extraction from smart card data
Individual or household social-demographic data
Step 1: Travel pattern clustering
Step 2: Demographic group clustering
Travel patterns identification
Social-demographic groups identification
Step 3: Aggregation
Travellers' aggregated social-demographic groups

Figure 1. Methodology framework

3.2 Travel Pattern Analysis

Traditional travel pattern analysis using smart card data focuses on daily frequent trip pattern recognition, including (Kieu et al. 2014; Tao et al. 2014), which cannot reflect the full and trustworthy portraits of passengers during a long-term range, such as yearly travel pattern. To overcome this issue, the paper proposes to first distinguish travel patterns using travel features extracted from SC data. In addition, two novel statistic measures are employed to identify and quantify the seasonality of different travel patterns.

3.2.1 Travel Feature Extraction

A key issue in passenger segmentation based on their travel behaviours is to extract accurate and comprehensive travel features from SC data. In this study, various travel features are defined as to calibrate passenger profiles in order to differentiate their travel patterns. All features are categorised into four types, related to temporal variability (When), spatial variability (Where), travel mode preference (Which mode) and travel frequency (How often), respectively. Authors have demonstrated and explained the feature extraction process in (Zhang et al. 2017). Here, we just list the features generated from SC data in Table 1. The morning and evening peak for London Underground is between 6:30 and 9:30 and between 16:00 and 19:00 on weekdays, respectively.

3.2.2 Affinity Propagation for Travel Pattern Clustering

In this paper, we propose to use Affinity Propagation (AP) algorithm for travel pattern clustering. AP, first developed by Frey et al. (2007), is a local-message-passing-based clustering approach. It has many advantages in terms of clustering task. Unlike other clustering algorithms, such as centroid-based k-means or k-medoids, AP does not require the predefined number of clusters before running this algorithm. Furthermore, AP takes all data points as candidates of exemplars (the centre of cluster). Since we hardly have any prior knowledge about underlying travel patterns, travel pattern identification can benefit from the above-mentioned advantages. The details of AP can be referred to (Frey et al. 2007).

3.2.3 Identify and Quantify Seasonality of Travel Patterns

Seasonal traffic demand may obviously increase the burden on urban public transit systems. Understanding long-term travel behaviours will help transportation agencies formulate better strategies and make more effective and efficient operating policies. In this paper, we propose two novel statistic measures, skewness, and kurtosis of trip distributions, to identity and quantify the seasonality of travel patterns, revealing more details of passengers' travel habits.

(1) Seasonality identification

This paper proposes to use the skewness of the trip distribution by month as a quantitative measure to detect whether a travel pattern exhibit seasonality. In statistics, skewness is a measure of the asymmetry in a distribution. Suppose the number of trips in each month during a year is $x_1, x_2, \cdots, x_N$, the skewness is:

$$sk = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^3 / N}{s^3} \tag{1}$$

where $\bar{x}$ is the mean value, s is the standard deviation, and N is the sample size. The value of skewness can be positive or negative. Positive skewness indicates data that are right skewed, and vice versa. To interpret the values for skewness, Bulmer (1979) suggests the following rule of thumb:

- If $|sk| > 1$, the distribution is highly skewed.
- If $0.5 < |sk| < 1$, the distribution is moderately skewed.
- If $|sk| < 0.5$, the distribution is approximately symmetric.

Hence, if the skewness of a travel pattern's trip distribution within (-0.5, 0.5), it is regarded as an unseasonal travel pattern. Otherwise, the travel pattern should be a seasonal one.

No.	feature	Description
Temporal feature	AFTI_WD	The average start time of the first trip on weekdays
	LFTI_WD	The average start time of the last trip on weekdays
	AFTI_WE	The average start time of the first trip on weekends
	LFTI_WE	The average start time of the last trip on weekends
	MPT_TUBE_NUM	the number of trips by tube during morning peak
	EPT_TUBE_NUM	the number of trips by tube during evening peak
	MPT_BUS_NUM	the number of trips by bus during morning peak
	EPT_BUS_NUM	the number of trips by bus during evening peak
	MPTR_TUBE	Morning peak travel rate by tube
	EPTR_TUBE	Evening peak travel rate by tube
	MPTR_BUS	Morning peak travel rate by bus
	EPTR_BUS	Evening peak travel rate by bus
	SEASON_1/2/3/4	The number of trips during the 1/2/3/4-th season
	SEA_PER_1/2/3/4	The percentage of trips during the 1/2/3/4-th season
Spatial Features	AVG_T_WD	The average of tube trip time on weekdays
	AVG_T_WE	The average of tube trip time on weekends
	VAR_T_WD	The variance of tube trip time on weekdays
	VAR_T_WE	The variance of tube trip time on weekends
	AVG_MAX_TD	The average radius travelled by tube per day
	VAR_MAX_TD	The average radius travelled by tube per day
	TOTAL_TD	The total travel distance by tube in the whole year
	AVG_TS	The daily average of the number of visited tube stations
	VAR_TS	The daily variance of the number of visited tube stations
	AVG_BL	The daily average of the number of visited bus lines
	VAR_BL	The daily variance of the number of visited bus lines
	ZONE_T_R	How often a passenger transfers the travel zone per day
	AVG_INNER	The mean value of the inner zone number
	AVG_OUTER	The mean value of the outer zone number
	VAR_ZONE_IO	The variance differences of inner-zone and outer-zone
Travel Mode Features	TUBE_NUM	The total number of the tube journeys
	BUS_NUM	The total number of the bus journey
	TUBE_PER	The percentage of tube journeys
Travel Frequency Features	TRA_DAY	How many days a passenger travels in the whole year
	TRA _DUR	Travel duration in the whole year
	TRA_WD	How many weekdays a passenger travels in the whole year
	TRA_WE	How many weekends a passenger travels in the whole year
	TRA_R_WD	Weekday travel rate (TRA_WD/ TRA _DUR)
	TRA_R_WE	Weekend travel rate (TRA_WE/ TRA _DUR)
	WD_TRIP	The total number of weekday trips
	WE_TRIP	The total number of weekend trips
	AVG_WD_TRIP	The average number of weekday trips per day
	AVG_WE_TRIP	The average number of weekend trips per day

Table 1. Feature extracted from smart card data

(2) Seasonality quantitation

To quantitative analysis each travel pattern's preferred seasons or months for travelling via public transit, we employ a statistic measure 'excess kurtosis' to evaluate the heaviness of the tails of a distribution relative to a normal distribution. Given a set of data $x_1, x_2, \cdots, x_N$, the formula of kurtosis is:

$$\text{ek} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4 / N}{s^4} - 3 \qquad (2)$$

where $\bar{x}$ is the mean, s is the standard deviation, and N is the sample size. Positive excess kurtosis indicates a 'heavy-tailed' distribution while negative indicates a 'light-tailed' distribution.

3.3 Social-demographic Groups Analysis

This step intends to identify the passengers' social-demographic groups by clustering LTDS data. Comparing to classical clustering tasks, LTDS data contain both continuous (e.g. age and income) and categorical variables (e.g. main occupation). TwoStep Cluster (Bacher et al. 2004) is a suitable algorithm to deal with this clustering task. In addition, TwoStep algorithm is a scalable cluster method, allowing to analyse large dataset and it can automatically determine the optimal number of clusters.

TwoStep Cluster algorithm involves three main steps: pre-clustering, outlier handling (optional) and clustering. The pre-cluster step is implemented by building a modified cluster feature tree. The clustering procedure is to group the sub-clusters resulting from the pre-cluster step into an optimal or a desired number of clusters. This process is implemented by using the hierarchical clustering algorithm, which can produce a sequence of partitions in one run. To determine the optimal cluster solutions, each potential number of clusters is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

3.4 Association analysis

The discussion of the relationship between travel patterns and social demographics are still somewhat ambiguous in existing literature. Previous work usually summarised the demographic attributes based on the results of travel pattern segmentations (Ortega-Tong 2013), which can be regarded as a one-to-one relationship mode. A more reasonable assumption of the relationship between travel patterns and social-demographics should be a many-to-many mode considering the following reasons.

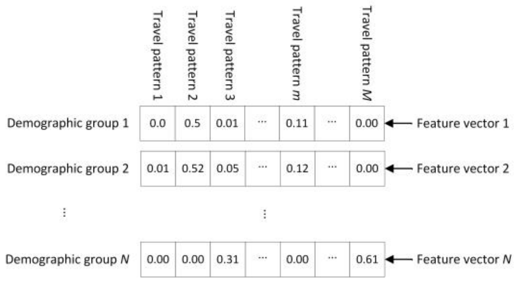
First, the previous passenger segmentation totally depends on the individual or household social-demographics. The segmentation results cannot reflect whether the selected social-demographic characteristics are indeed significant determinants of travel patterns at the individual level. Secondly, according to previous researches, some social-demographic characteristics, such as age, income, and car ownership, can largely affect personal travel patterns. However, the complex travel behaviours are not determined by a single demographic feature, but the combination of diverse social-demographic attributes, as well as some other unknown latent factors.

To achieve a better explanation of the individuals' complex travel patterns, we need to find more significant relationships between travel patterns and the social-demographic characteristics while keeping the diversity of travel patterns to the largest extent. Thus, we aggregate the initial social-demographic categories by applying hierarchical clustering (HC) (Kraskov et al. 2005).

The third step is based on the results of the first two steps. After we obtained the travel patterns in the first step and the demographic groups in the second step, it is found that people in the same demographic group may exhibit different travel patterns. Thus, we use the distribution of passengers over different travel patterns as the feature vector of each demographic group, as illustrated in Figure 2. For example, the demographic group 1 has 50% of passengers exhibit the second travel pattern and 11% of passengers exhibit the *M*-th travel pattern, as shown in Figure 2. Using these feature vectors, HC clustering is then applied to aggregate demographic groups to identify significant relationships between demographic groups and travel patterns. HC starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are the closest together, and (2) merge the two most similar clusters. This continues until all the number of clusters are equal to the predetermined value. This is illustrated in the diagrams below. To determine the optimal number of clusters, we use the Dunn Index to measure the clustering performance.

4. CASE STUDY

In this section, we use London's Oyster Card and LTDS data from 9708 passengers to demonstrate the proposed framework of exploring the relationship between travel patterns and social-demographics. Details are given bellow.



	Travel pattern 1	Travel pattern 2	Travel pattern 3	···	Travel pattern *m*	···	Travel pattern *M*	
Demographic group 1	0.0	0.5	0.01	···	0.11	···	0.00	⟵ Feature vector 1
Demographic group 2	0.01	0.52	0.05	···	0.12	···	0.00	⟵ Feature vector 2
⋮				⋮				
Demographic group *N*	0.00	0.00	0.31	···	0.00	···	0.61	⟵ Feature vector *N*

Figure 2. The feature vector of each demographic group is the passenger distribution across different travel patterns

4.1 Travel Patterns of Residents in London

4.1.1 Data pre-processing

Travel features of 9708 passengers are extracted as described in section 3.2.1. Before clustering, features should be first rescaled to remove the influence of the different data range. Second, the extracted features include spatial, temporal, mode preference and travel frequency characteristics. Since the dimension of the travel measures is large and some of them are intercorrelated, PCA is applied to reduce the dimensionality. The number of principal components to be retained is automatically estimated by using the method proposed by Minka (2000). Finally, the first 20 components are kept, explaining around 96.8% of the total variance.

4.1.2 Travel Pattern Clustering Results

AP is used to detect travel pattern clusters. We calculate the Dunn index by running AP with the different number of predefined clusters ranging from 2 to 20. According to Figure 3, the Dunn index reaches the local maximum value at 15 clusters, indicating the optimal segmentation.

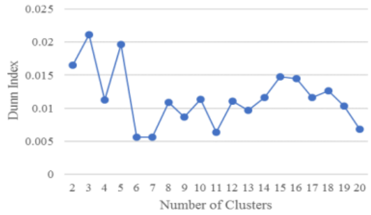


Figure 3. The Dunn index changes with the number of clusters obtained by Affinity Propagation algorithm

The 9708 passengers are classified into 15 clusters, as shown in Figure 4. The largest group contains around 14% passengers while the smallest cluster (cluster 15) only consists of 94 passengers (less than 1%). Observing the travel features of cluster 15, we find over 95% of individuals in this cluster only used their Oyster cards once or twice during the whole year. Thus, we think these Oyster cards are just for disposable use and we do not further discuss it.

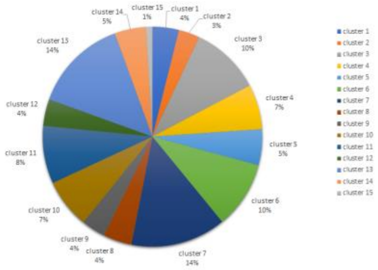


Figure 4. The size of each travel pattern

4.1.3 Seasonality identification

For the rest of 14 clusters, we would like to further identify the seasonal travel patterns. The values of skewness of the 14 clusters are listed in Table 2. Referring to the above-mentioned rules, we summarise the 14 clusters into two main categories, unseasonal travel patterns (cluster 1 to 7) and seasonal travel patterns (clusters 8 to 14). Figure 5 illustrates three distinct distribution of trips over the overall year. Overall, the number of unseasonal passengers in the first seven clusters is 5515, accounting for around 53.1% of the total population, and seasonal passengers are 4533 (near 46.9%), a little fewer than the unseasonal. Since the second step for seasonality quantitation is only applied to seasonal travel patterns, we move this part to the semantic analysis in the next subsection 4.1.4.

Unseasonal travel patterns							
Cluster No.	1	2	3	4	5	6	7
skewness	-0.06	-0.03	-0.19	-0.20	-0.12	-0.29	-0.16
Seasonal travel patterns							
Cluster No.	8	9	10	11	12	13	14
skewness	-1.05	-1.17	-1.51	-1.08	-0.92	-0.61	0.53

Table 2. The values of skewness of the 14 clusters

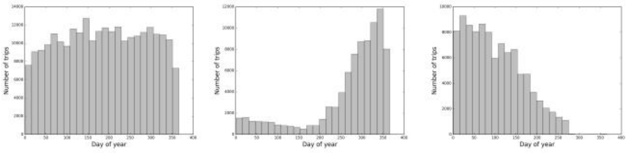


Figure 5. The distribution of trips of Cluster 2, 10 and 14 during the year 2012

4.1.4 Semantic analysis of travel patterns

(1) Unseasonal Passengers

Clusters 1 to 7 denote unseasonal daily routine travel behaviours. According to Table 3, a first similarity can be made among these clusters: clusters exhibiting relative long travel duration. Then, the distinct travel frequencies and preferred travel modes identify varying types of typical unseasonal travel patterns.

- **Cluster 1/2: Unseasonal heavy bus/tube users**

There are 378 passengers in **Cluster 1** identified as the most frequent usage of bus services and the longest travel duration (number of days before the first and the last day on which using public transit). According to Table 3 their first and last daily trip times are relatively early among all clusters. And these passengers took public transit over 4 times per day. In addition, **Cluster 2** shows quite similar travel frequency, travel days and durations. The difference between Cluster 1 and 2 is the opposite travel mode preference. In addition, on average, users in this cluster have the earliest mean time of the first journey and the latest mean time of the last trips on weekdays. These attributes strongly imply the purpose of commuting.

- **Cluster 3/4/5: Unseasonal moderate bus/tube/mixed-mode users**

The second subgroup contains three clusters consisting of unseasonal, moderate public transit users with different travel mode preferences. Specifically, **Cluster 3** represents passengers who usually travel by bus. However, most trips occurred during off-peak time. Passengers in **Cluster 4** exhibit very similar temporal behaviours with Cluster 2, but the evening use of Cluster 4 is around one-hour earlier than that of Cluster 2. In addition, passengers in Cluster 4 travel more on weekdays than any other unseasonal type. Comparing to Cluster 3 and 4, **Cluster 5** has the latest first and last departure time. The travel mode is somewhat irregular.

- **Cluster 6/7: Unseasonal occasional bus/mixed-mode users**

Regarding the remaining two **Clusters 6** and **7**, the most common features are the long travel duration but few and diffuse travel days. Additionally, the proportions of trips during rush hours of the two clusters are both lower than 15%, but their travel modes are different. The former prefers to use bus while the latter has no obvious preference. On weekends, residents in Cluster 7 made more evening trips than Cluster 6. Another significant distinction exists in the spatial features. The range of motion of passengers in Cluster 6 aims at the inner city, and Cluster 7 is the opposite.

(2) Seasonal Passengers

Clusters 8 up to 14 are identified as seasonal travel patterns, thus we compute the excess kurtosis to identify their season preference. Results are listed in Table 4. We can see that the first 4 clusters are heavy-tailed, and the rest are light-tailed, and we also point out the favourite travel season of each cluster (the fifth row in Table 4).

In addition, passengers in the first seven clusters have similar seasonal behaviours, trending of which is like the subplot (b) in Figure 5. Only the last cluster No.14 shows the opposite trend like Figure 5 (c). The similarity is that all the most-frequent travel periods are in the winter, indicating the influence of seasonality on residents' travel behaviour. More details of the semantic analysis of each travel patterns are given as follows.

- **Cluster 8: Seasonal heavy bus users**

These passengers heavily rely on bus for daily trip, but the trip distribution is seasonal. Passengers averagely use the public transit for about 92 days out of the 113-day travel duration, which results in a very dense bus travel demand. With regards to daily temporal behaviour, people in this cluster travel earlier in the morning and later in the evening than other seasonal patterns on weekdays. In spatial respect, according to these passengers' sparse tube journeys, the average tube travel zones reveal that they only use tube very far away from central London.

- **Cluster 9/10/11: Seasonal moderate bus/tube/mixed-mode users**

Passengers in Clusters 9 to 11 show a moderate travel frequency with distinct travel modes (bus, tube and mixed, respectively) during winter. Among them, Cluster 9 and 11 exhibit a similar temporal behaviour during weekdays. The significant features of Cluster 9 are the extremely short average travel distance and narrow travel zone by tube. In addition, Cluster 10 exhibits a remarkable temporal similarity with Cluster 8 on weekdays. Comparing with other seasonal travel patterns, another considerable feature of Cluster 10 is the high proportion of weekday trips, which proportioned for over 75% of the total number of trips. In addition, approximate a quarter of trips occurred during morning and evening peak hours. These features strongly indicate Cluster 10 is a typical seasonal travel pattern with a main purpose of commuting.

- **Cluster 12/13/14: Seasonal occasional bus/tube/mixed-mode users**

Unseasonal passengers							
	Heavy		Moderate			Occasional	
Cluster No.	1	2	3	4	5	6	7
First/Last tap-in time on weekdays	10:20 17:30	9:10 18:40	11:30 15:50	9:10 17:20	11:50 17:20	12:10 14:15	12:30 15:25
First /Last tap-in time on weekends	12:00 16:40	12:50 17:40	12:40 16:00	12:50 16:30	13:40 17:40	12:00 13:50	13:30 16:10
Tube trip rate in morning/evening peak (%)	2.0/2.6	24.9/22.1	2.7/3.5	29.0/26.8	3.1/8.2	0.5/0.8	4.2/6.3
Bus trip rate in morning/evening peak (%)	14.0/15.1	6.5/1.2	16.2/13.2	6.4/5.2	11.2/11.6	13.1/10.2	9.6/8.9
Tube travel zones	1.68-2.98	1.41-2.77	1.73-3.18	1.41-3.32	1.69-3.27	1.36-1.39	1.69-3.69
Mean distance of tube trip (km)	4.95	5.04	4.97	5.96	5.51	0.42	6.54
Mode preference	Bus	Tube	Bus	Tube	Mix	Bus	Mix
Total tube/bus trips	115/998	567/169	62/361	303/102	98/126	2/87	43/68
Weekdays percentage	59.42%	63.53%	63.96%	82.34%	64.12%	63.96%	62.64%
Travel days	266	256	147	149	98	38	36
Travel duration	344	339	307	292	198	206	273

Table 3. Selected travel features of unseasonal travel patterns

Seasonal passengers							
	Heavy	Moderate			Occasional		
Cluster No.	8	9	10	11	12	13	14
Excess Kurtosis value	0.458503	0.626613	1.382376	0.444484	-0.23215	-0.96314	-0.59545
Favourite season	4	4	4	4	4	4	1
First/Last tap-in time on weekdays	10:40 17:40	11:10 15:40	10:40 17:40	12:25 16:10	12.25 14:10	13:15 16:50	12:10 15:30
First /Last tap-in time on weekends	12:10 17:30	12:20 15:30	10:40 13:50	13:30 16:40	- -	13:50 17:15	- -
Tube trip rate in morning/evening peak (%)	4.0/4.7	0.5/0.7	22.4/22.6	7.7/4.6	0.4/0.8	12.2/22.8	11.4/14.8
Bus trip rate in morning/evening peak (%)	13.9/12.8	18.1/12.5	6.7/5.5	10.1/8.5	14.0/11.8	1.9/1.9	6.8/10.0
Tube travel zones	1.79-3.26	1.45-1.49	1.42-2.93	1.74-3.39	1.15-1.17	1.49-3.28	1.48-3.66
Mean distance of tube trip (km)	5.13	0.53	5.25	5.51	0.22	5.70	6.60
Mode preference	Bus	Bus	Tube	Mix	Bus	Tube	Mix
Total tube/bus trips	74/314	2/172	83/28	43/61	0.6/16	51/12	16/12
Weekdays percentage	55.29%	58.52%	75.45%	59.34%	99.76%	51.46%	99.76%
Travel days	92	52	43	38	8	28	12
Travel duration	113	78	66	104	44	175	135

Table 4. Some selected travel features of seasonal passengers

These passengers rarely travel by public transit, and their infrequent travels always occurred during a short period. Clusters 12 and 14 have quite similar temporal behaviours (except seasonality) that they both exhibit a late morning usage at about 12:00 and early evening ending before 16:00 on weekdays, and they hardly use public transit during weekends, therefore the trips on weekdays take account for almost 100% of the total trips. In terms of Cluster 13, it shows the longest travel duration among all seasonal travel patterns, resulting in a more diffuse usage.

4.2 Social-demographic groups

In this case study, ten socio-demographic features collected in LTDS are considered for clustering. In LTDS, age (PAGEI) and the distance between home and work/education (PLENN) can be treated as continuous variables and others are categorical variables.

In the TwoStep clustering, the BIC is calculated as the clustering performance metric to determine the optimal cluster number. As smaller values of BIC indicating better models, 32 clusters are chosen as the most efficient and practical number, preserving a significant diversification of the residents in London. For privacy reason, we cannot provide the details of the social-demographic characteristics for each group. We only present the 32 clusters' average social-demographic features ordered by the average age in Figure 6. To achieve a better visualisation, each demographic feature has been scaled by the maximum value.

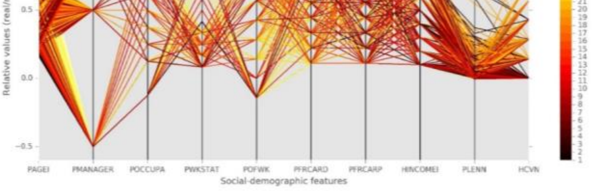


Figure 6. Sankey plot of 32 social-demographic groups

In summary, the first three demographic groups can be regarded as teenagers under education. The dominant distinctions are their household characteristics, including income, the distance between home education place, and car ownership, as well as two personal characteristics (car driver/passenger frequency). Then, we treat groups 4 up to 23 as middle-aged adults, which exhibit the most diverse demographic features at both household and individual level. Among them, group 6, 9, and 14 are unemployed. Finally, the rest 9 groups (from 24 up to 32) mainly consist of retired old-age people grouped by using the household characteristics.

4.3 Significant relationship analysis

Passengers' travel behaviours strongly depend on their demographics. However, because of some unknown factors, such as subjective travel preference, and the accessibility of PT, individuals in the same demographic groups may exhibit different travel patterns. However, comparing the passenger distribution across the travel patterns, we find that some demographic groups presented a quite similar distribution. Thus, Hierarchical Clustering is applied to this distribution to aggregate original social-demographic groups. The aggregation process and the relationship between the aggregated demographic groups and the travel patterns are presented using a flowchart in Figure 7.

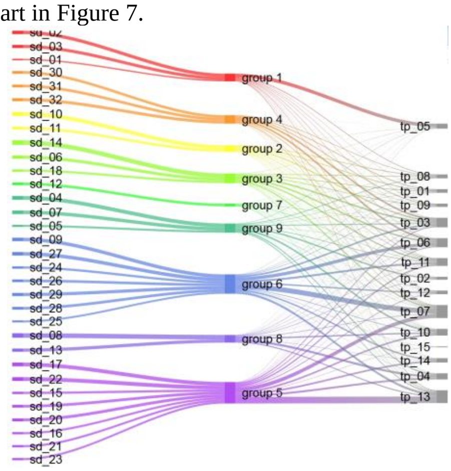


Figure 7. The aggregation process of 32 demographic groups and the relationship between demographic groups and travel patterns. The left denotes the original 32 social-demographic groups, the middle is the 9 groups aggregated by the passenger distribution across travel patterns, and the right are the 15 travel patterns.

To further explain the semantic meaning of the aggregation results, we selected two typical examples to give more details of the semantic analysis of the relationship.

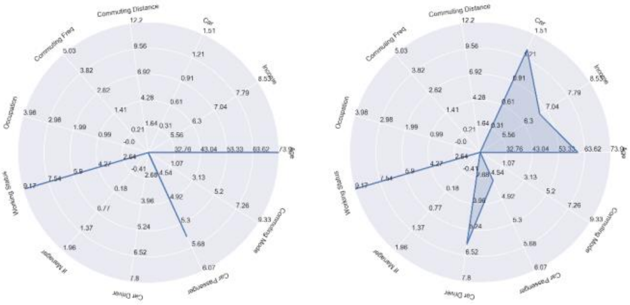
(1) Young passengers

Young passengers in the first three original social-demographic groups are merged together as group 1 in Figure 7 in this aggregation process. Observing the passenger distribution across travel patterns, almost half of them (total 807 persons) belong to travel pattern 5, which is described as unseasonal moderate mixed-mode travellers. It means that young passengers, most of whom are students, have no obvious preference for a certain travel mode. What's more, because the working time is not as fixed as office workers, they did not always travel during the morning peak.

(2) Old passengers

The old passengers are merged into two groups, group 4 and group 6 in Figure 7, respectively. The former is combined of demographic groups 30 to 32 (the oldest three) and the latter includes demographic groups 24 to 29. The average social-demographic characteristics of the two aggregated groups are presented using a radar plot in Figure 8. It can be seen in Figure 8, passengers in group 4 (average 74-year-old, 879 people) are slightly older than those in group 6 (average 61-year-old, 2037 people). In addition, although the working status of the two groups indicate that most of the people are retired the average levels of car ownership, household income and frequency of car driver of group 4 are considerably lower than that of group 6.

The passenger distribution of the two groups across 15 distinct travel patterns can be seen in Figure 9. For group 4, a significant feature is that most of the oldest prefer to use bus (travel pattern 1, 3, 6, 8, 9 and 12) and over 60 % passengers in group 4 exhibit unseasonal patterns, which implies that their daily mobility highly depends on the public transport, especially bus system. The potential reasons include the cheap ticket prices and no demand for commuting. The travel mode preference of group 6 is similar to group 4. However, the tube usage of group 6 is more frequent than group 4.



group 4 group 6

Figure 8. The average demographics of group 4 and 6.

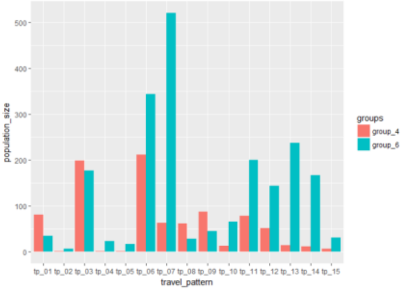


Figure 9. The passenger distribution of group 4 and 6 across 15 distinct travel patterns

5. CONCLUSIONS AND FUTURE WORK

Smart card data provide a promising opportunity to investigate the complex travel behaviours in public transport system. This paper proposes a novel and entire framework to analyse the significant relationships between travel patterns and social-demographics of passengers using smart card data and household survey. This effort provides some new insights into the spatio-temporal travel patterns and their linkage between demographic roles of passengers.

Future work can be conducted based on the research presented in this paper. First, the extracted features from SC data can reveal travel behaviours from the spatial, temporal, travel mode and frequency perspectives, but each feature is just the mean value during the research period, which may miss some useful behaviour features for travel patterns analysis. Thus, other more effective methods should be explored to represent the SC data to describe the travel behaviour of passengers. Second, exploring the possibility of predicting social-demographic roles using SC data is an interesting feature research direction.

ACKNOWLEDGEMENTS

Transport for London is acknowledged for sharing London's Oyster card data and LTDS data, which are used in this study. This work was supported in part by the Consumer Data Research Centre, Economic and Social Research Council, U.K., under Grant ES/L011840/1. The first author's research is jointly supported by the China Scholarship Council under Grant No. 201603170309 and the Dean's Prize from the University College London.

REFERENCES

Bacher, J., Wenzig, K. and Vogler, M., 2004. SPSS TwoStep Cluster-a first evaluation.

Bulmer, M., 1979. Principles of Statistics Dover Publications. *New York*.

Ceapa, I., Smith, C. and Capra, L., 2012. Avoiding the crowds: understanding tube station congestion patterns from trip data. *Proceedings of the ACM SIGKDD international workshop on urban computing.* ACM.

Frey, B. J. and Dueck, D., 2007. Clustering by passing messages between data points. *science* 315 (5814), 972-976.

Kieu, L. M., Bhaskar, A. and Chung, E., 2014. Transit passenger segmentation using travel regularity mined from Smart Card transactions data.

Kraskov, A., Stögbauer, H., Andrzejak, R. G. and Grassberger, P., 2005. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)* 70 (2), 278.

Lathia, N. and Capra, L., 2011. How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. *Proceedings of the 13th international conference on Ubiquitous computing.* ACM.

Minka, T. P., 2000. Automatic choice of dimensionality for PCA. *Nips.* Vol. 13.

Nassir, N., Hickman, M. and Ma, Z.-L., 2015. Activity detection and transfer identification for public transit fare card data. *Transportation* 42 (4), 683-705.

Ortega-Tong, M. A., 2013. *Classification of London's public transport users using smart card data*. Massachusetts Institute of Technology.

Pelletier, M.-P., Trépanier, M. and Morency, C., 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies* 19 (4), 557-568.

Sari Aslam, N., Cheng, T. and Cheshire, J., 2018. A high-precision heuristic model to detect home and work locations from smart card data. *Geo-spatial Information Science*, 1-11.

Shi, X. and Hangfei, L. (2014) *The analysis of bus commuters' travel characteristics using smart card data: the case of Shenzhen, China.*

Sun, L., Lee, D.-H., Erath, A. and Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing.* Beijing, China. 2346519: ACM. 142-148.

Tao, S., Rohde, D. and Corcoran, J., 2014. Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography* 41, 21-36.

Yuan, N. J., Wang, Y., Zhang, F., Xie, X. and Sun, G., 2013. Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach. *2013 IEEE 13th International Conference on Data Mining.* 7-10 Dec. 2013.

Zhang, Y. and Cheng, T., 2017. Feature Extraction for Long-term Travel Pattern Analysis. *Proceedings of the 25th GISRUK conference.* Manchester, UK.

Zhang, Y. and Cheng, T., 2018. Inferring Social-Demographics of Travellers based on Smart Card Data. *2nd International Conference on Advanced Research Methods and Analytics.* Valencia, Spain. Editorial Universitat Politècnica de València.

Zhang, Y. and Cheng, T., 2019. A Deep Learning Approach to Infer Employment Status of Passengers by Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*.