Transportmetrica A: Transport Science

ISSN: 2324-9935 (Print) 2324-9943 (Online) Journal homepage: http://www.tandfonline.com/loi/ttra21

Understanding public transit patterns with open geodemographics to facilitate public transport planning

Yunzhe Liu & Tao Cheng

To cite this article: Yunzhe Liu & Tao Cheng (2018): Understanding public transit patterns with open geodemographics to facilitate public transport planning, Transportmetrica A: Transport Science, DOI: 10.1080/23249935.2018.1493549

To link to this article: https://doi.org/10.1080/23249935.2018.1493549

Accepted author version posted online: 28 Jun 2018.
Published online: 12 Jul 2018.

TRANSPORTMETRICA A: TRANSPORT SCIENCE
https://doi.org/10.1080/23249935.2018.1493549

Understanding public transit patterns with open geodemographics to facilitate public transport planning

Yunzhe Liu and Tao Cheng

SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK

ABSTRACT

Plentiful studies have discussed the potential applications of contactless smart card from understanding interchange patterns to transit network analysis and user classifications. However, the incomplete and anonymous nature of the smart card data inherently limit the interpretations and understanding of the findings, which further limit planning implementations. Geodemographics, as 'an analysis of people by where they live', can be utilised as a promising supplement to provide contextual information to transport planning. This paper develops a methodological framework that conjointly integrates personalised smart card data with open geodemographics so as to pursue a better understanding of the traveller's behaviours. It adopts a text mining technology, latent Dirichlet allocation modelling, to extract the transit patterns from the personalised smart card data and then use the open geodemographics derived from census data to enhance the interpretation of the patterns. Moreover, it presents night tube as an example to illustrate its potential usefulness in public transport planning.

ARTICLE HISTORY
Received 1 October 2017
Accepted 22 June 2018

KEYWORDS
Personalised smart card data; transport planning; latent Dirichlet allocation modelling; travel pattern analysis; geodemographics

1. Introduction

Since the publication of the smart card in the late 1960s, the technology itself and its applications have been maturing in various industrial sectors. Particularly, the smart card-based automated fare collection (SCAFC) system has been providing considerable benefits for public transportation in both urbanising and post-urbanised cities. The durability, portability, manageability, and data safety offered by the SCAFC system have made smart cards predominately replacing the manual ticketing and the magnetic cards. Although the original purpose of adopting the SCAFC system is to improve revenue collection, due to its automatic collection of fine-grained travel transaction information, the system also offers extensive opportunities for integrating with the transport planning to provide more intelligent planning solutions.

Plentiful literature and studies focus on the potential applications of the data extracted from the SCAFC, from understanding interchange patterns to transit network analysis and user classifications. In particular, understanding and characterising the passengers' travel

CONTACT Tao Cheng tao.cheng@ucl.ac.uk SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK

behaviour by mining the smart card data is one of the hot topics given its usefulness for transport planning (Pelletier, Trépanier, and Morency 2011). For example, the results of travel behaviour characterisation through the SCAFC data can be utilised as the evidence base to facilitate scientific decision-making and assessment of the current transit network, e.g. transport policy before-and-after assessment (Daraio et al. 2016; Lee, Oh, and Min 2011; Yu and He 2016). The usage of the smart card is valuable to provide a more accurate estimation of the travel demand and accordingly make service adjustment that copes with variations in ridership (Trépanier and Morency 2010). Furthermore, by comprehensively understanding passengers' travel behaviour and the associated attitude (e.g. quantify transit loyalty), travellers that have the inclination to swing between private and public transport modes can be targeted, therefore identify potential changes in the transport sector so as to achieve sustainable development (Webb 2010; Zhao, Webb, and Shah 2014).

Although many studies focus on using smart card data to extract travel patterns, not much work validates their findings, e.g. the explanation and understanding of the patterns is limited. This is because of privacy so that the card data are always anonymised, which leads to the unavailability of personal information about the passengers and their travel purpose. Although transport authorities widely employ travel surveys to make up for the vacancy caused by the anonymity, the sample size captured is extremely smaller than the quantity of public transport users in reality. For instance, the London Area Transport Survey (LATS) and London Travel Demand Survey (LTDS) are carried out by Transport for London (TfL) to monitor the travel demand for London residences (TfL 2015). The LTDS annually samples approximately 8000 households, and the response rate in 2013/2014 was 49.1%. Therefore, how to widely verify the patterns extracted from the smart card is challenging. Furthermore, not much work has demonstrated how the patterns extracted could be used in transport planning in practice. For example, travellers' behaviours have been extracted in Canada (Agard, Morency, and Trépanier 2006) and in Guangzhou (Yu and He 2016) with users' card types to enrich the clusters' interpretability. However, none of them apply their findings to the transit planning.

Taking the aforementioned points into consideration, this paper, therefore, proposes a methodological framework to extract transit patterns from the SCAFC system and conjointly cooperate with socioeconomic data (e.g. open geodemographics) in order to enrich the interpretability of the passenger behaviour in public transportation and subsequently informs the public transit planning. The Oyster card system operating in London is employed to demonstrate the methodology.

The paper is structured as follows. The next section presents a comprehensive literature review of related works using smart card data for analysing travellers. Section 3 develops a methodological framework of the research, which including six steps that are further developed in Sections 4–7, demonstrated by the case of Oyster card data operating in London. The major contribution, limitations, and possible improvements are summarised and discussed in Section 8.

2. Related works: understanding travel behaviours from the SCAFC system

A comprehensive review of the public transit application of the SCAFC data can be found in Pelletier, Trépanier, and Morency (2011). They summarised mainstreams of the existing

researches into three levels: tactical, operational, and strategic. Here, we recap some of their findings by adding the latest development along these three levels.

The tactical-level study focuses on developing algorithms to estimate the transfer interval or create/recompile a detailed origin-destination matrix for the passengers, which can subsequently inform the service schedule adjustment (Pelletier, Trépanier, and Morency 2011). Particularly, studies focus on the smart card used in the bus system, due to the lacking of information about the alighting point (Trépanier, Tranchant, and Chapleaub 2007; Zhang et al. 2015). Additionally, estimating the interchange time between different modes of public transport is also popular (Seaborn 2009).

Studies at the operational level concentrate on assessing the transit network by setting several performance indicators (Pelletier, Trépanier, and Morency 2011). Schedule of public transit and transit fare pricing is one of the popular operational-related topics (Wang, Li, and Chen 2015). Moreover, the ridership and loyalty of smart card are assessed so as to identify and signify eligibility to access certain service (Bagchi, Gleave, and White 2003; Trépanier and Morency 2010). Lathia and Capra (2011) developed an algorithm to estimate travellers' travel behaviour so as to minimise the unnecessary overpaid fare by providing personalised ticket recommendations based on the estimated travel pattern.

The strategic level is one of the most highly active research areas within academia, majorly related to user characterisation and classification (Pelletier, Trépanier, and Morency 2011). Given its relevance to our work proposed, we here examine the literature in this aspect in detail. The early work of Agard, Morency, and Trépanier (2006) analysed the variability of travel behaviour between two types of card (i.e. elderly and regular adult) by clustering the travel profile generated from the smart card data from the bus system. Efforts have been made in integrating smart card data with personal or market information to improve user characterisation and the interpretability. For example, with prior knowing the personal information registered by the smart card user, Utsunomiya, Attanucci, and Wilson (2006) carried out a targeted marketing analysis by using the non-anonymous card data offered by the Chicago Transit Authority in order to analyse the relationship between travellers' behaviour with access distances, frequency of use, and types of residential area. Based upon market supplement information, Kieu, Bhaskar, and Chung (2015) characterised passengers into four classifications, namely 'transit commuters', 'regular OD passengers', 'habitual time passengers', and 'irregular passengers' in order to provide accurate information and services. Moreover, by using data mining techniques, analysts are now attempting to estimate the unobtainable travel purpose from the smart card data that are partially consistent with the travel survey. For example, in order to enhance the understanding of the journey pattern, Kusakabe and Asakura (2014) developed a data fusion methodology combined with the naive Bayes probabilistic model to estimate the behavioural features of trips by utilising data, respectively, derived from the SCAFC system and the personal trip survey. They suggest that this proposed method can supplement behavioural attributes absenting in the smart card dataset.

To this end, however, most of the existing studies have been conducted at the macroscopic level (Ali, Kim, and Lee 2016). Only a few studies use truly personalised smart card data, which mean to recognise each traveller with trips pertaining to him/her as a single observation (El Mahrsi et al. 2014). For instance, through reformatting public transit journeys into a weekly profile, Lathia et al. (2013) emphasise that the usage of public transit does vary dramatically between individuals. The key finding reinforces the advocate of

personalising transport information service based on SCAFC data. In addition, Lathia et al. (2013) also point out that since personalisation has become a mainstream of research in the context of Internet services, the existing algorithms that have been successfully employed on web-based preference data that can possibly be implemented in the urban transport network.

Additionally, although extensive studies reviewed above consider organising all travellers into categories, few studies focus on identifying the residents, the key groups of people configuring the demand-side, and more importantly to understand the social demographic impact on travel patterns which are mainly attached to residents (e.g. national census). As mentioned above, Trépanier and Morency (2010) assessed the ridership and loyalty of smart card user. However, their work is more partial to monitor the card usage rather than to analyse the characteristics of frequent passengers. One of the similar works can be found in the sociology report done by Lathia, Quercia, and Crowcroft (2012), which aims at inferring the community well-being from smart card data and socioeconomic data. They calculated the correlation between station-by-station flow and station-to-station IMD (Index of Multiple Deprivations) and state that 'deprived areas tend to preferentially attract people living in other deprived areas' (91). Although their work mainly studies how communities relate to each other rather than revealing the effects of the socioeconomic attributes on passengers' transit behaviours, it also derivatively provides an attempting of setting criteria to identify residents.

More recently, El Mahrsi et al. (2014) published a work that uses personalised smart card information to determine local residences and examines how socioeconomic attributes can affect travel patterns. They identified frequent passengers and their 'Residential Station' by setting several thresholds and conditions. Then, they clustered passengers' temporal behaviours by using a generative model-based clustering technique, which are supplemented with socioeconomic clusters that are derived from demographic data (aggregated in 200 m per 200 m raster cells) to identify how the temporal clusters are influenced by the socioeconomic cluster of the city. However, there are three major drawbacks in their study. Firstly, the residence identification is novel but not systematically rigorous, which contains non-residence passengers (e.g. long-distance commuters). Secondly, the selection of variables used to construct the socioeconomic cluster is questionable and mainly contains population density and income. As discussed previously, however, several studies, such as Lathia, Quercia, and Crowcroft (2012), have manifested that the variation of passengers' travel behaviour is affected by many factors including both demographic, socioeconomic, and also physical environmental domains. Therefore, merely choosing a small subset of variables from the multidimensional dataset to build the classification is neither comprehensive nor convictive. Finally, as also admitted by El Mahrsi et al. (2014), the results generated from their work are required to be examined and evaluated through practical applications, which again can recall back to the research gap we discussed above.

In summary, there are rapid progresses in all the three mainstreams of using data from the SCAFC system. The strategic-level category, particularly, consists some limitations and research gaps that we are aiming to bridge. Similar to El Mahrsi et al. (2014), this paper also utilises a model-based generative model to characterise passenger's travel behaviour from a personalised smart card dataset. However, our work is substantially different in the following aspects:

(1) The residence identification is progressively refined before the travel pattern analysis, particularly the long-distance commuters are filtered out from this study.
(2) Latent Dirichlet allocation (LDA), a generative model commonly utilised in the text mining, is introduced to conduct the temporal clustering analysis aiming to characterise passenger's travel pattern.
(3) A fine and comprehensive geodemographic classification is attached to the clustered patterns to improve the interpretability.
(4) More importantly, this project attempts to utilise the empirical results to guide the London Night Tube planning, which accordingly add practise significance of this project, and is usually missing in current research.

3. Methodology

Figure 1 presents the five steps of the methodology framework of this paper, namely (1) data pre-processing, (2) residence and home station identification, (3) temporal pattern extraction, (4) geodemographics analysis, and (5) policy demonstration. Moreover, each step can be partitioned into several sub-steps that will be further illustrated. We will explain the first step of data pre-processing in this section, and the other four steps will be explained in the following sections, Sections 4–7.

3.1. Data pre-processing

Figure 2 illustrates the procedures for constructing the data pre-processing phase. This stage involves two main sub-steps, namely data extraction and data cleansing. The outputs generated by following the series of procedures can act as the 'raw dataset' that will be subsequently inputted to the next stage which will be discussed in Section 4.

The Oyster card data are extracted from the SQL database by specifying the study period and types of card usage (only the underground and some rail trips). Considering the limitation of computing power, the integrity of data, and there were no significant holidays/events during the period, the study period of this project is set to range for four

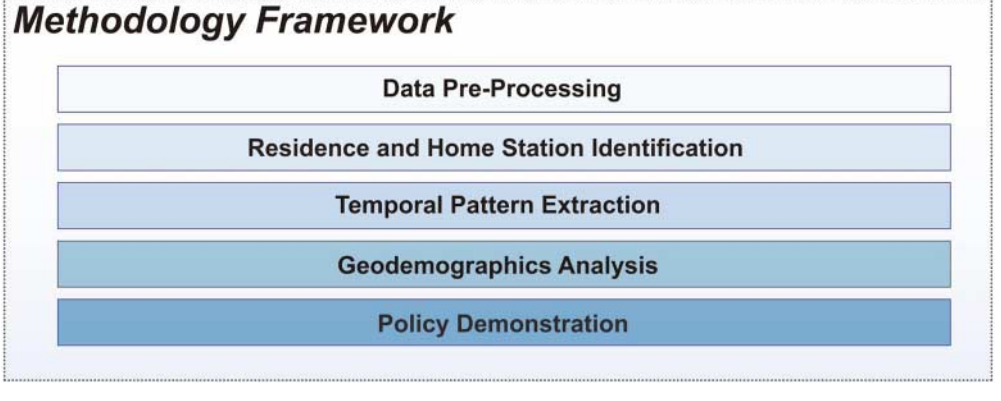

Figure 1. Overview of methodology framework.

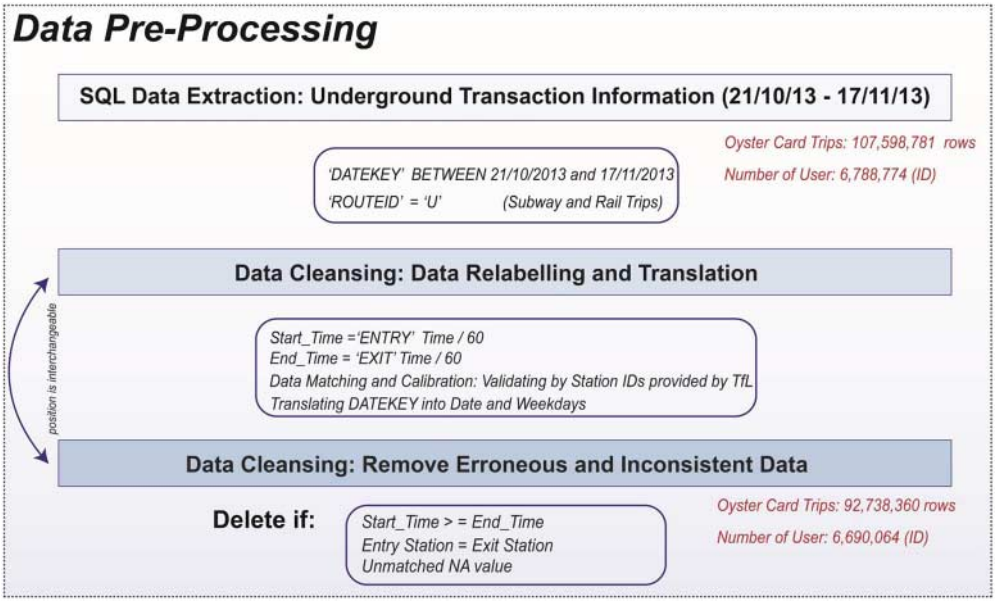

Figure 2. Data pre-processing workflow.

Table 1. Example dataset derived from Oyster card data (after translation).

Oyster ID	Date	Entry	Exit	Boarding	Alighting	Card type
15184207	21/10/2013	Westminster	Putney	1265	1290	Retail
15987462	21/10/2013	Acton Town	Twickenham	651	681	Photocard
16982142	21/10/2013	Baker Street	Kennington	1161	1191	Staff
...	...	...	...	...	...	
53126021	17/11/2013	South Quay	Tower Hill	1335	1369	Retail

weeks in 2013 (between 21 October 2013 and 17 November 2013). The SQL data extraction ends up with an approximately seven Gigabytes CSV file which involves 107,598,781 Oyster card usage information that was made by 6,788,774 passengers within our study period. Table 1 demonstrates the conceptual representation of the structure of the Oyster card data extracted and translated from the SQL database.

Although some unstructured data are immediately filtered out through the initial data extraction, for example, the bus trips whose destination information is unknown are not involved in this study, the remaining data do exhibit some noises. The noises are mainly configured by the erroneous and inconsistent data. For instance, in some cases, the destination information is missing or unmatched with the officially published underground station ID checklist (i.e. the NLC code); the exit time is earlier than the entry time; the destination is same as the origin. These aforementioned noises data are accordingly removed from the dataset.

After the pre-cleaning process, the data are ready to be imported to the formal data processing stage. The cleansed dataset remains 92,738,360 trips made by more than 6,500,000 passengers (IDs).

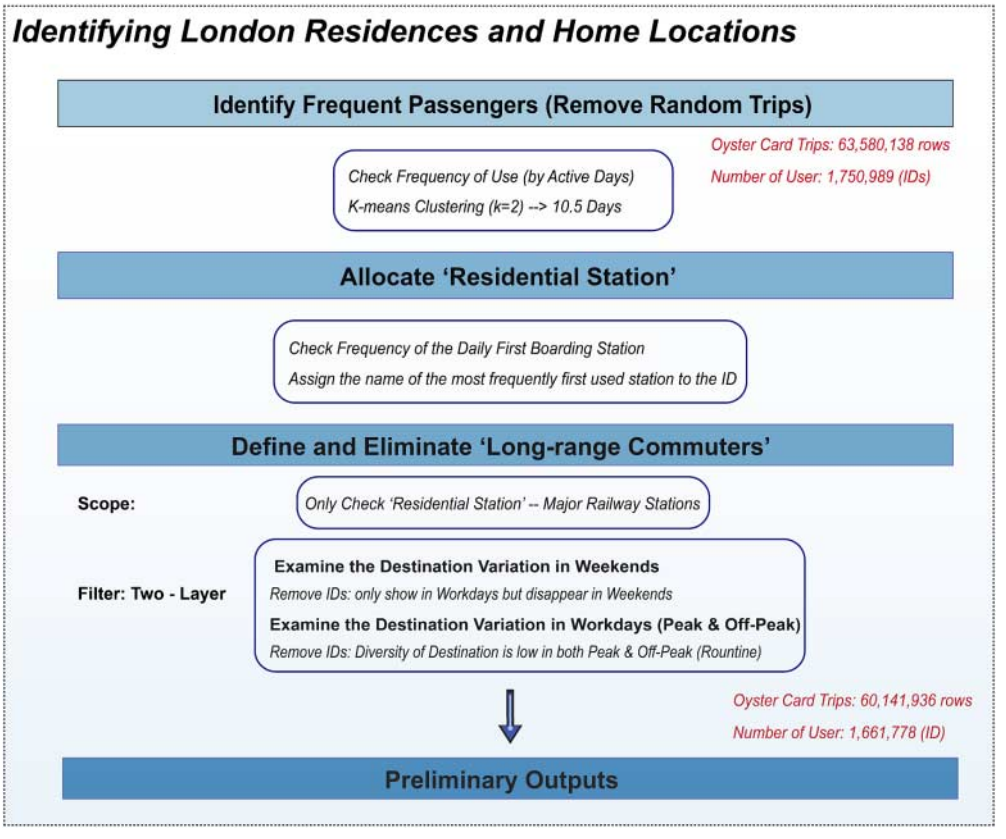

Figure 3. Identifying London residences and home locations.

4. Identifying residences and home locations

To make a better link to the contextual information (i.e. geodemographics, which will be discussed later), London residences are needed to be identified. In this study, we define Londoner as London's residences – people who only live in the Greater London. This given definition of 'Londoner' clearly distinguishes London residences from the tourists and long-distance commuters (people who only work but not necessarily live in London). Figure 3 presents the workflow showing procedures of residence identification.

4.1. Identify frequent travellers

There are two major steps to find potential 'Londoner'. The first step is to find frequent passengers based on their travel regularity. The travel regularity can be viewed containing two aspects: firstly, the total travel frequency and secondly, the periodicity that can be indicated by the 'active day'. Extensive empirical studies have mentioned the threshold to differentiate frequent and infrequent travellers. However, because different studies have a different data structure and study period in different cities, the threshold values set by the specific condition of the study are also various. For example, Kieu, Bhaskar, and Chung (2015) differentiate passengers by utilising *k*-means clustering to the one-dimensional array (i.e. total

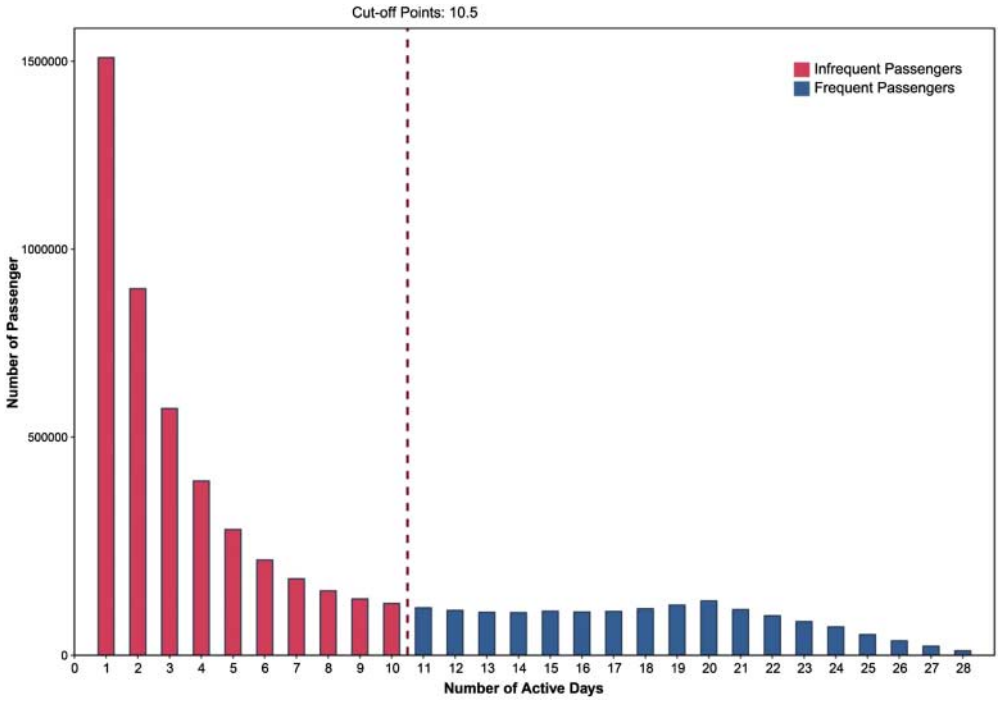

Figure 4. Distribution of passengers by the number of active days; cut-off point generated from *k*-means.

travel frequency) and pick the value of 53 trips as the cut-off point. Even the study time periods are quite similar (approximately one month), Lathia, Quercia, and Crowcroft (2012) and El Mahrsi et al. (2014), respectively, employ 2 out of 30 days and 10 out of 30 days as their frequency threshold to screen the frequent traveller without giving reasons for their ad hoc choices.

In this paper, taking a comprehensive consideration, we combined the methods from the aforementioned examples together to fit a relatively objective standard that fits our data structure. The concept of 'active day' from Lathia, Quercia, and Crowcroft (2012) and El Mahrsi et al. (2014) is firstly imported; and secondly, we employ *k*-means clustering ($k = 2$; i.e. frequent/infrequent) to assist us to define the cut-off point defining frequent/infrequent traveller. After implementing this method, 10.5 (days) is finally adopted as the cut-off value, defining that the frequent passenger is the passenger that should use his/her Oyster card at least 37.5% of days within the four-week study period. Figure 4 shows the histogram and cut-off point selected in this paper. Among the 6,690,064 card ID imported from the pre-processing phase, 73.8% of them (4,939,075) is classified as infrequent users. The remaining 26.2% (1,750,989) frequent passengers will be imported to the next step.

4.2. Allocation of Residential Stations

Once an individual is classified as a frequent passenger, his/her most frequently used 'daily first boarding' station is identified simultaneously. The daily first boarding underground

station is generated by examining the maximum frequency of the boarding station during the whole four weeks. Just as its name implies, no matter how many times did this person travel through the London Underground within a single day, only his/her first boarding station after the opening time will be operant and counted. This most frequently used daily first boarding underground station will be assigned as 'Residential Station' to that frequent passenger. In other words, that passenger can be assumed to live in the neighbourhood that is located in proximity to his 'Residential Station'. This assumption is also a key linkage between passenger and geodemographics, which will be discussed in the later section.

4.3. Identify and eliminate long-distance commuters

After allocating frequent passengers with their 'Residential Station', the second screening step is to identify longer-distance commuters. The commuting flow data, from the 2011 Census, clearly illustrate that long-distance commuters ('the distance workers') do commonly exist in many major cities across the UK, which are also reported in LTDS documentation (TfL 2015). Because the long-distance commuters are employed in London, they are also likely to be classified as frequent passengers. Hence, they are also assigned to their own 'Residential Station' that are more likely to be national or international railway stations (e.g. King's Cross and Victoria), which means that they only use these stations as gateways to enter London but not necessarily live near these stations. Lathia, Quercia, and Crowcroft (2012) also identify the existence of long-distance commuters and their 'Familiar Locations' (which is the term they adopted and can be interchangeable with 'Residential Station' used in this study). They decided to selectively 'close down' these major entries to ensure long-distance commuters are minimised in their study. The approach they employed is seemingly feasible in terms of both effectively and efficiently filtering out the distance workers; however, residences who actually live near these stations and assigned to these railway stations as their 'Residential Stations' are also eliminated, which inevitably lose a considerable number of passengers and their valuable travel information.

In this context, this paper aims to provide a heuristic approach attempting to minimise the effects from the long-distance commuters group rather than arbitrarily delete all the passengers whose 'Residential Stations' are those major railway stations. The first phase is to 'zoom in' the scope of the target station, i.e. only passengers whose 'Residential Station' is categorised as the railway stations of London (validated by using the data provided by TfL) are affected by the further filter. According to the data, 22.2% passengers (387,875) are assigned to the major railway stations in Greater London. These passengers, therefore, have relatively higher possibility to be long-range commuters.

The filter is twofold; firstly, the weekend activity is examined: passengers who only have Oyster card records at workdays but no trip at weekends during the whole study period at all are filtered out. Because London residences are more likely to go out and travel if they are free on the weekends at least one time among the four weeks, whereas long-distance commuters are less likely to go London during their day-offs. Secondly, this filter examines the variability/diversity of passenger's destination at workdays by peak and off-peak time. The frequency of passengers' destination during peak time and off-peak can be calculated, respectively, forming a matrix which is partially shown in Table 2. In this case, Cluster 1 will remain in the dataset, but cluster 2 will be filtered out. The basic rationale of this layer filter is to find the low diversity of destination used during both peak and off-peak time. It

Table 2. Destination Diversity and Clusters

Oyster Card ID	Number of Destination (Peak)	Number of Destination (Off-Peak)	Cluster
10005707	1	4	2
10016470	16	12	1
10019440	2	5	2
10164918	14	12	1
12253114	10	1	1

should be mentioned that the destination checked should not be the 'Residential Station' to avoid counting return trips. Similar to the aforementioned method we employed to define frequent traveller, the *k*-means clustering is also implemented here to find the cluster that exhibits a low frequency of travel during both peak and off-peak.

For example, commuters (including long-range commuters) will travel from his/her residential station to his/her workplace (destination) during the peak time and will probably work around there for a whole day (until the evening peak). So, the diversity of destination during the peak time is quite low; as for the off-peak, commuters (including long-range commuters) will not travel so frequently during the working time; so, this filter is to identify the commuter group; combining the previous layer, i.e. people will not do any travel during the weekend, the remaining group has a high probability to be long-range commuter (need to be eliminated from the dataset).

The filter we adopted in this study does contain limitations, i.e. it certainly removes some potential Londoners. For example, pupils would likely to be neglected if their trips are not various enough during the off-peak time, e.g. some of them are likely to go back home directly after school. However, it should be mentioned that to accurately identify residences/commuters from smart card data is not the main focus of this paper. By importing the pre-processed data through the filter, the final output remains 60,141,936 Oyster transactions contributed by 1,661,778 cardholders. Our finding is approximately close to the estimation of Londoners. According to the LTDS (2013–2014), 882,576 (male) and 677,714 (female) (1,560,290 in total) generated underground/DLR trips (Tfl 2015).

Figure 5 shows the graduated symbol map inferring the quantity of Londoners assigned to each TfL's NLC stations. Obviously, the central London witnesses a larger number of tube traveller due to the higher population as well as station density, whereas stations situated in the outskirts of metropolitan are allocated by less population.

5. Travel patterns of residence

Figure 6 demonstrates the workflow of extraction of travel patterns, which consists of two sub-steps, namely data reformatting and temporal clustering.

5.1. Data reformatting

In this step, pre-processed data are processed and formatted into 'weekly travel profile' that is prepared for the pattern analysis. Figure 7 contains example demonstrating the 'weekly travel profile' for three travellers generated from the processed data. The number shown in the heat map indicates the cumulative frequency of Oyster card use during study periods.

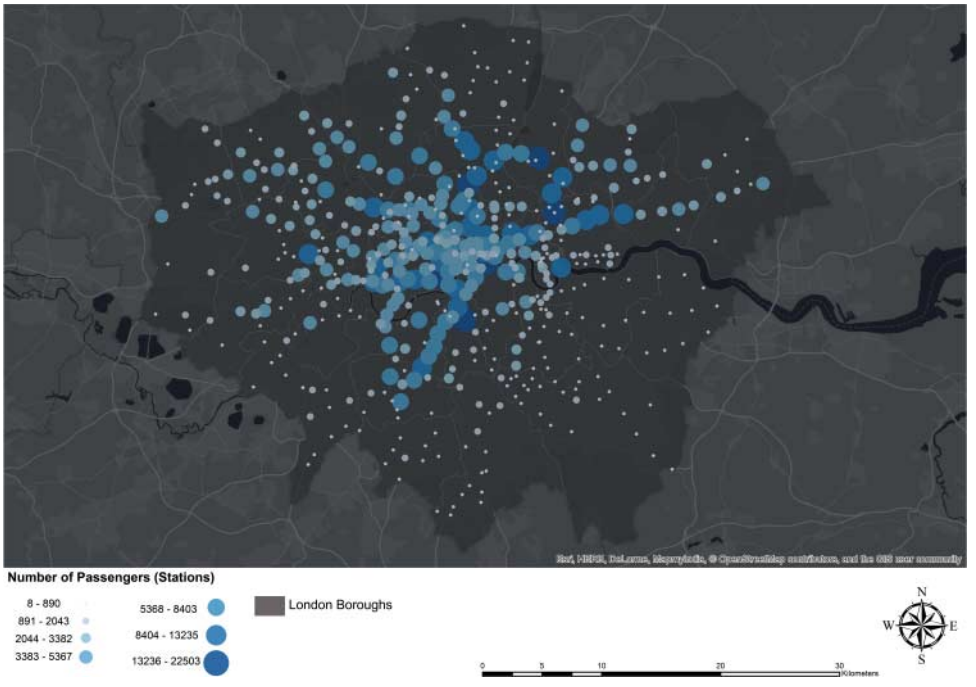

Figure 5. Residential Stations and number of Londoners allocation.

The one-hour temporal interval that acts as a bin to count the accumulated frequency is set, which totally ends up with 28 temporal intervals (based on the TfL's standard, 'an operational day' contains 29 hours) for each of the weekly profile. Furthermore, peak hour and weekend indicators, respectively, presented by a red and green rectangle are also added to the heatmaps shown in Figure 7. Take one passenger (ID $=$ 4498) for instance, this passenger mainly gets on the tube during the middle of peak times (7.00 to 8.00 in the morning and 17:00 to 18.00 in the afternoon) and has sporadic trips at night during workdays, with some flexible trips are witnessed at noon during the weekends.

The same profiling approach is applied to each of the unique card ID within the dataset, and eventually, 1,661,778 travel profiles are generated. Each profile denotes a cardholder's trip distribution over each operational hour of each day of the week. Equivalently, each card user is viewed as an observation over 203 temporal variables (configured by 29×7 hours): the first variable is the frequency of journeys he made on Sunday 0.00 to 1.00 am, the second is the number of his trips from 1.00 to 2.00 on Sunday, and so forth. At this stage, these variables can be subsequently translated into a series of 'temporal words' (words multiplied by the number of accumulated frequency), which can provide a foundation for the LDA clustering process.

5.2. Latent Dirichlet allocation

In this step, clustering analysis is utilised for the identified groups of passengers who exhibit similar travel behaviours based on their 'weekly travel profiles'. As the clustering process is

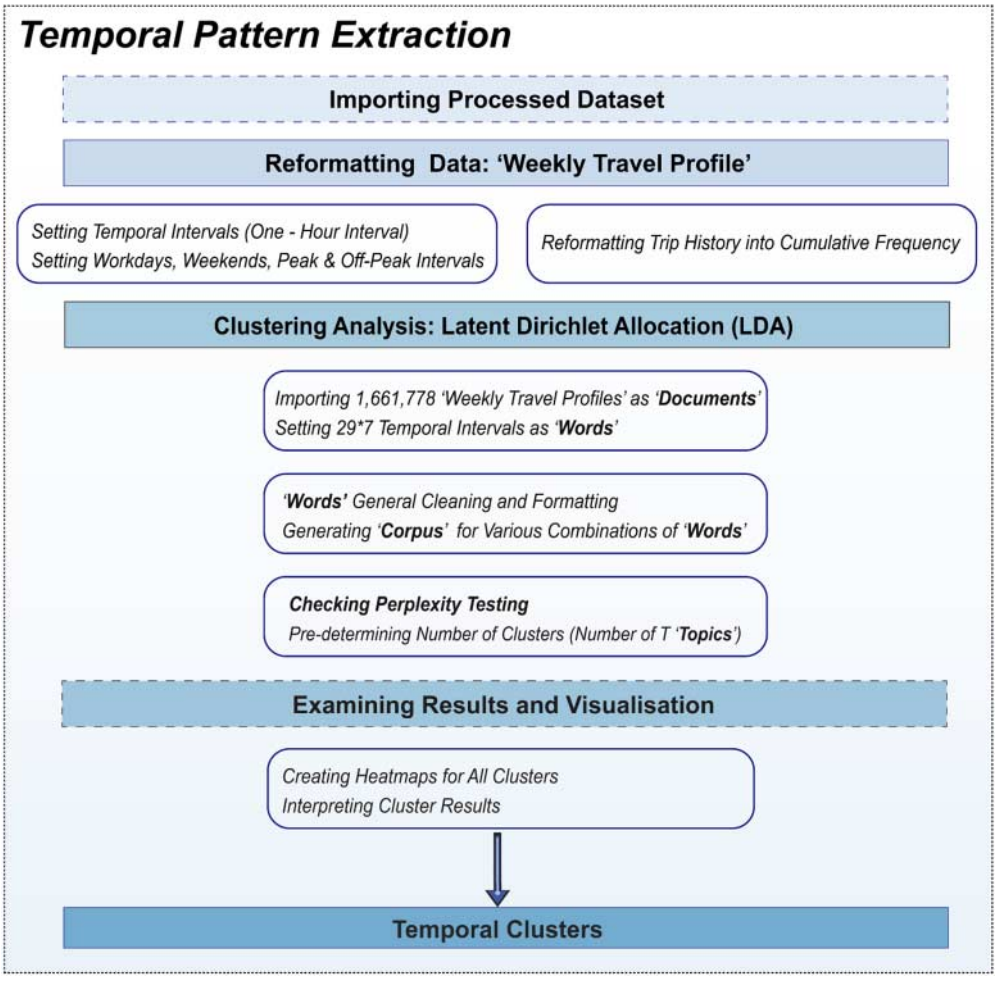

Figure 6. Temporal pattern extraction workflow.

solely based on the temporal profiles (boarding time), possible clusters generated through the clustering are hence named as 'Temporal Clusters' which also in accordance with their counterparts (i.e. 'Socioeconomic Clusters'; geodemographic classification). In this study, the LDA algorithm is adopted to conduct clustering analysis, and the reason to select this kind of cluster technique is presented below.

LDA is one of the most commonly used unsupervised topic modelling methods, as 'a generative probabilistic model for collections of discrete data such as text corpora' (Blei, Ng, and Jordan 2003, 993). The fundamental theory of LDA is that 'documents are represented as random mixtures over latent topics, where each topic is characterised by a distribution over words' (Blei, Ng, and Jordan 2003, 996). The aim of LDA is to infer that maximise the likelihood (the posterior probability) of the collection of electronic text (Blei and McAuliffe 2007). More specifically, LDA is defined as a three-level hierarchical Bayesian model which models each entity of a collection of documents as a finite mixture over an underlying set

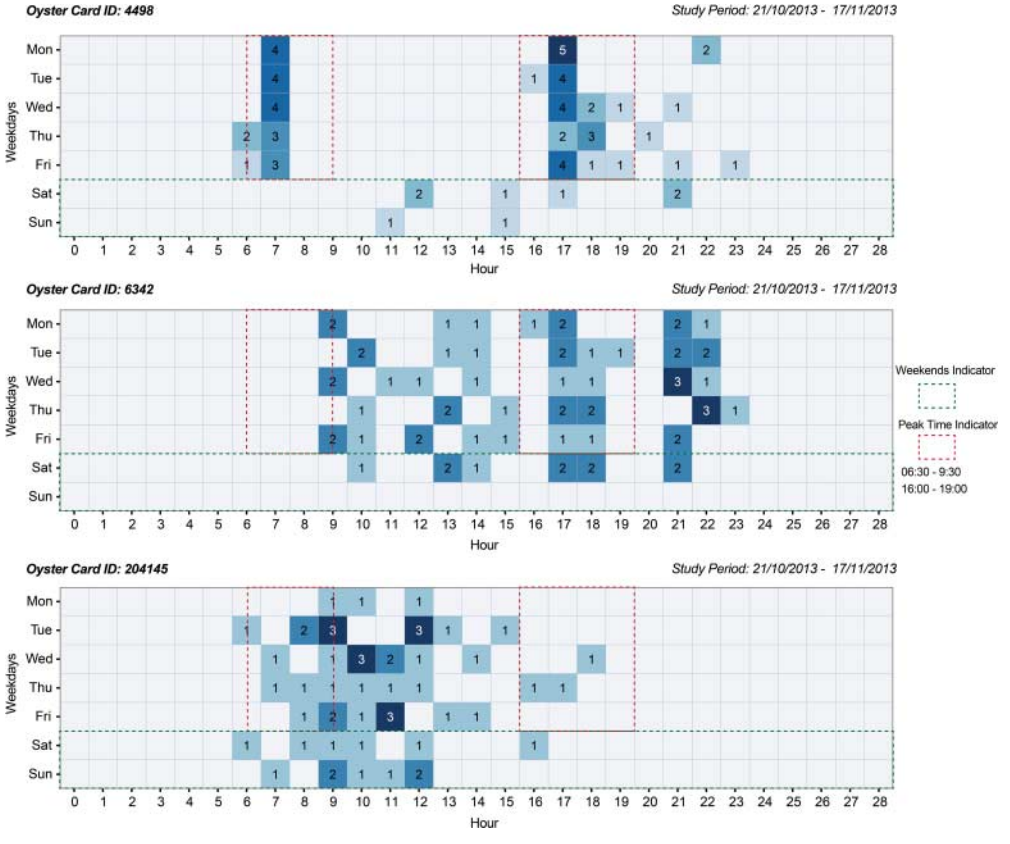

Figure 7. Examples of 'Weekly Travel Profile'.

of topics. And each topic is modelled as an infinite mixture over an underlying set of topic probabilities which offer an explicit representation of a document.

LDA has been successfully adopted to analyse text information generated from various sources, for instance, journal articles (Wu et al. 2014), 'We the Media', and social media (such as microbloggers, e.g. Twitter) (Cha and Cho 2012; Lai, Cheng, and Lansley 2017), and contextual photos from Flickr (Awadi, Khemakhem, and Jemaa 2012). Moreover, LDA and LDA-based models can also assist to solve problems in many domains such as bioinformatics and collaborative filtering, content-based image retrieval. For instance, Perina et al. (2010) used an LDA algorithm to obtain highly informative representation for microarray experiments in gene clustering and sample classification. More recently, He et al. (2015) develop an LDA-based technique that automatically recommends tourist routes and creates a user profile model to improve the current tourist routes recommendation system. Although El Mahrsi et al. (2014) eventually choose to utilise mixture of unigrams model to analyse passengers' travel pattern due to the concern of complexity reduction, they did point out that their methodology can be analogous to the LDA model as each passenger is viewed as a document containing multiple words (the words are their travel date and time). Their work is the first and the only attempt to use text mining to analyse the smart card data.

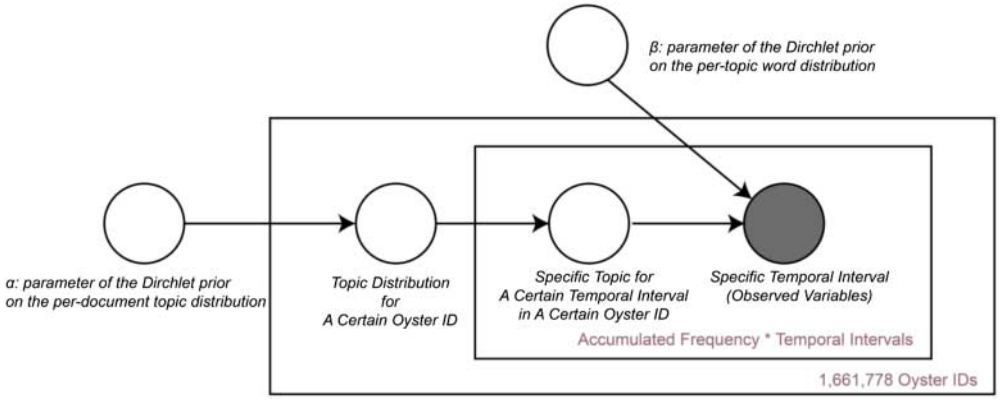

Figure 8. LDA representations in Oyster card data.

Table 3. Processed dataset – LDA-based 'Text' data.

Oyster ID	Temporal words (intervals)
15184207	Monday_9_10, Monday_9_10 ...
15987462	Monday_9_10, Monday_9_10 ...
16982142	Monday_12_13, Monday_12_13 ...
...	...
53126021	Sunday_15_16, Sunday_15_16 ...

5.3. Temporal pattern extraction by LDA

Figure 8 illustrates the graphical model representation of LDA specified for the Oyster card data, which can be compared to the original LDA graphical representation introduced by Blei and McAuliffe (2007). In this context, each Oyster card users are viewed as a 'document', in which the 'word' is a combination of their travel time period (reformatted as 'Weekday_Time', e.g. Monday_9.00) multiplied by the accumulated frequency. We used the default settings ($\alpha = \beta = 0.1$) that are widely used.

Table 3 displays the example of the reformatted dataset. The first column contains the Oyster card ID, and the second column involves the 'temporal intervals' multiplied by the number of their frequency. Additionally, the number of topics (T) that has been pre-defined by using perplexity will be discussed below.

Reed (2012) claims that the key objective of topic modelling is to automatically identify the topics from a series of given documents. Accordingly, the assumption that the quantity of topic (i.e. T or K) is known is not very precise. In other words, this parameter is required to be defined in advance. Strictly, there is no correct answer to this issue, however, by utilising statistical interference to evaluate the model analyst can generate a more reasonable frequency of cluster based on the feedback of the clustering quality. Perplexity is one of the most popular evaluations of LDA (Blei, Ng, and Jordan 2003), which measures the modelling power (i.e. how well a probability distribution) through computing the inverse log-likelihood of unobserved documents. Generally, the better model exhibits lower perplexity indicating fewer uncertainties about the unobserved document. In this study, we tested different numbers of topics (ranging from 2 to 25) and recorded the perplexity from each of the selection. As can be seen from the testing result presented in Figure 9, there are

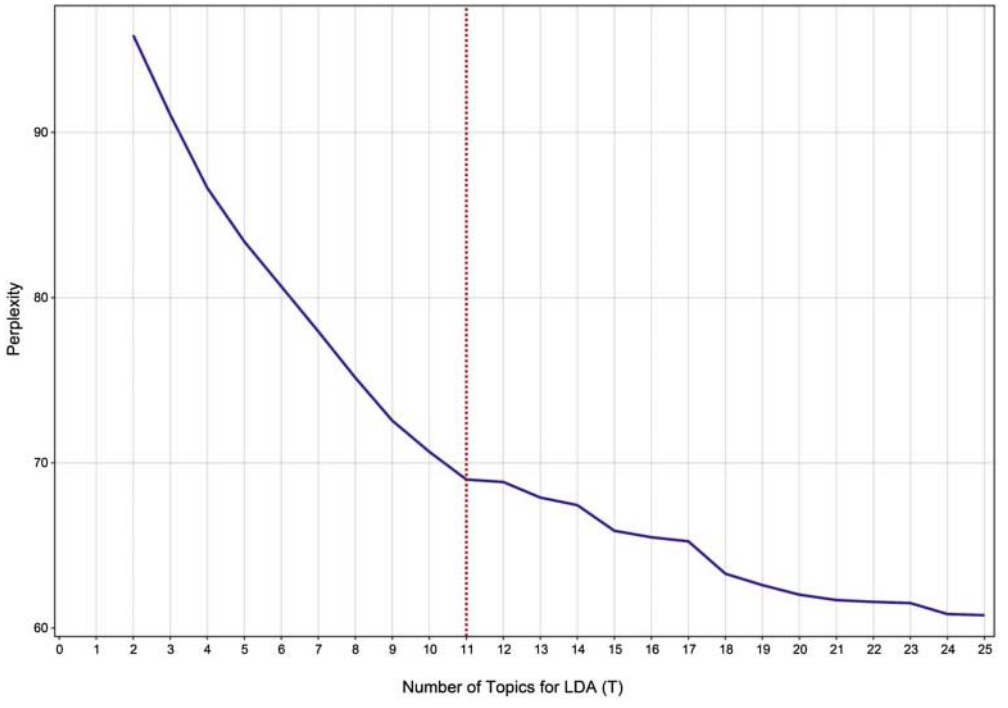

Figure 9. Perplexity test for LDA number of topics selection (2–25 topics).

some clear 'elbows' in between 10 and 12, 14 and 16, as well as 17 and 19 topics. Taking the convenience of result interpretation and clustering size into consideration, 11 topics were finally selected as the optimal number of topics (T).

Figure 10 demonstrates the 11 Temporal Clusters generated through LDA modelling. Basically, each of the documents (i.e. passengers) from the LDA model is categorised into a vector of proportions, namely a mixture of words (i.e. Temporal Intervals). These 'word-clouds' are presented in a 'weekly travel profile' manner so as to coordinate with the one presented in the previous section, forming the 11 'Temporal Clusters'. The dark colour of a certain temporal interval indicates a high probability of the appearance of the interval, whereas the lighter colour represents a lower probability. In other word, a series of heatmaps reveal the possibility of travel time, which will be the core basis of cluster interpretation. It should be clarified that although the cluster resulted in a situation where a card user is described as a vector of probability among the 11 clusters since they may exhibit different types of travel pattern, each card user is assigned to a unique Temporal Cluster based on the highest probability (i.e. most preferable pattern).

5.4. Temporal Cluster interpretation

Overall, three categories broadly depicting different travel patterns can be identified in Figure 10, namely regular peak-time pattern, off-peak noon travel pattern, and randomly evening/weekend travel pattern. Firstly, more than half number of Temporal Clusters can be characterised into the same category based on their regular peak-time travel pattern, which partly portrays a typical home-to-work commute. The heat maps clearly show that

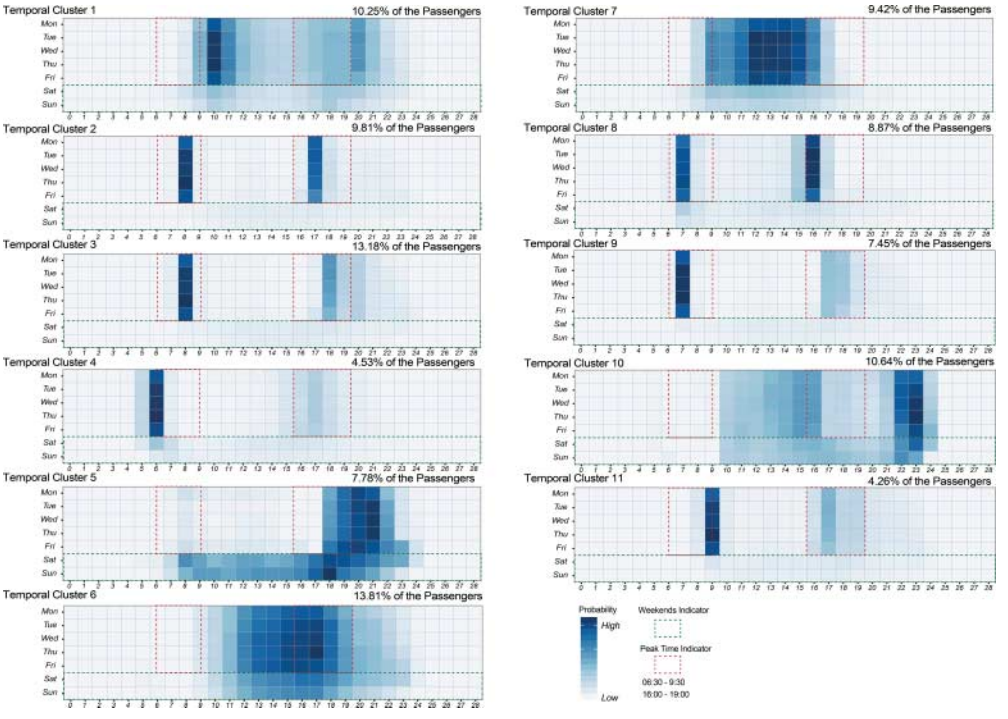

Figure 10. Temporal clustering results.

passengers from these six clusters are more likely to take the subway in the rush hours of five workdays, especially during the morning peak. Secondly, besides the clusters depicting a routine peak-time pattern, Temporal Clusters 1 and 7 can configure another classification portraying an off-peak noon travel behaviour. In this category, passengers from these temporal clusters predominately board the metro during the off-peak period between two peak times. Clusters 5 and 10 collectively construct the category demonstrating an evening/weekend travel pattern. Generally, passengers within these two groups are more likely to travel during the late night (from 22.00 to 24.00) in all weekdays and also have some random trips during the daytime at the weekends. The last cluster that has not been described yet is the Temporal Cluster 6, whose travel pattern depicts a random travel behaviour since the suggestions presented by a relatively balanced diffusion of high travel possibility across the whole week (the late-night pattern also identified at the weekends).

It should be mentioned that, although temporal clusters in the same category exhibit broadly similar travel pattern, some nuances do exist between them. Moreover, accompanying with the information offered by the card type, the travel pattern for each of the temporal clusters can be examined individually. Here, however, these detailed patterns will not be interpreted in further since it goes beyond the main concentration of this paper.

6. Linking Temporal Clusters to open geodemographics

Figure 11 illustrates the procedures contained in the third phase. For the purpose of improving the interpretability for created temporal clusters, it is helpful to add some contextual

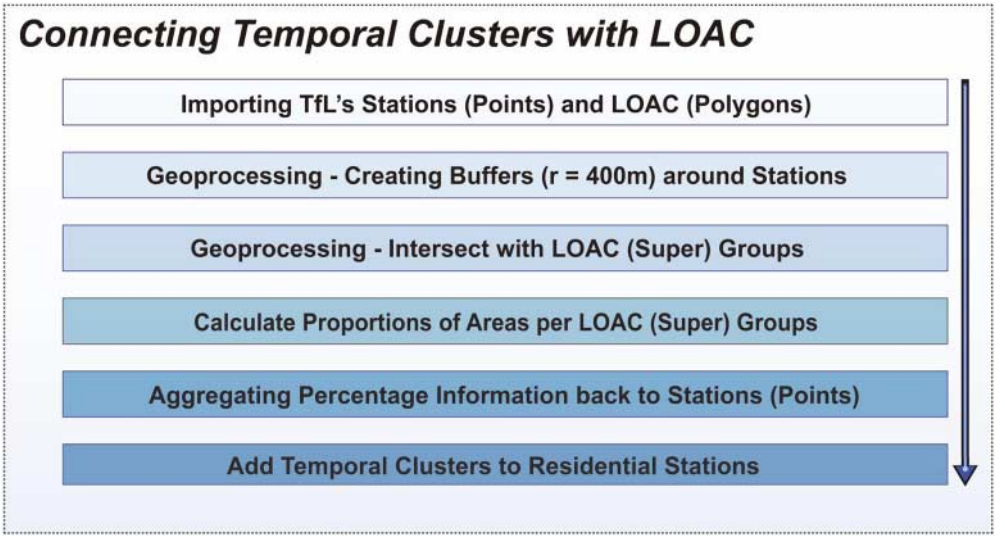

Figure 11. Connecting temporal clusters LOAC workflow.

information (i.e. socioeconomic data) available at a fine-grained spatial level to the temporal dataset (El Mahrsi et al. 2014). In this study, the geodemographic analysis is utilised to reinforce the contextual richness of the Temporal Cluster in order to improve the cluster interpretability.

6.1. London output area classification

The geodemographic analysis is a methodological framework aiming at a contextual summary of salient multidimensional socioeconomic and built environment characteristics for small area zonal geography, producing what are often shorthanded as 'neighbourhood' classification. The advantages of geodemographics are well documented (see Harris, Sleight, and Webber 2005; Leventhal 2016). This analysis has been developed for decades internationally, whose implementations have been ranging from both private and public sectors (Singleton and Spielman 2014).

The London Output Area Classification (LOAC) is an open, purely census-based, the general-purpose geodemographic classification created specifically for Greater London at the Output Area geography (Longley and Singleton 2014). The methodological information to build this classification is detailed in Longley and Singleton (2014). Basically, the LOAC partitions 25,053 OAs in Greater London into eight Super Groups (21 groups). Each LOAC Super Groups and Groups has its own name and a brief 'pen portrait' that describes the most possible multidimensional characteristics of the member characterised in this classification. Figure 12 shows the geographic distribution of the LOAC Groups within the Greater London.

6.2. Linking underground stations to LOAC

The fundamental assumption of this study is that passenger lives in the neighbourhood that is proximity to their most frequently used first boarding station (i.e. the residential station).

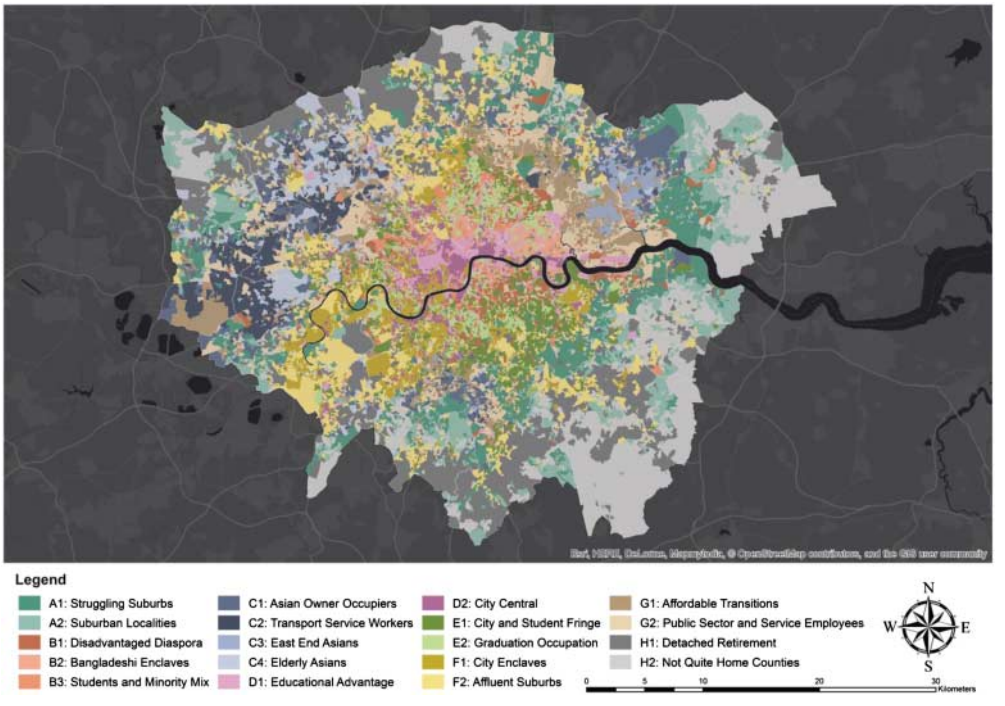

Figure 12. LOAC Groups distribution.

In this step, all TfL's stations in Greater London are matched with the OAs occupied by LOAC based on their geographical proximity. To achieve this aim, a walkable distance buffer (400 m, or five-minute walking distance) is created for each of the stations. The catchments created around the stations represent the assumed community near the station. Figure 13 depicts the catchment areas of all the underground stations among LOAC Groups. Each buffer involves several numbers of LOAC, which is presented into proportion based on the area occupying within the catchment domain. For instance, the catchment area of Kensington Olympia station is from by approximately 10% of 'B: High Density and High Rise Flats', 53% of 'D: Urban Elites', 13% of 'E: City Vibe', and 24% of 'F: London Life-Cycle', which further can be subdivided by the LOAC Groups.

By utilising the method mentioned above, LOAC and Temporal Clusters are merged through the spatial location of their corresponding 'Residential Stations'. Accordingly, the linkages between LOAC and the Temporal Cluster can be inspected through the proportion of LOAC for each Temporal Cluster. This proportion can be calculated by multiplying the proportion of total passengers, respectively, occupied by Temporal Cluster and the proportion of LOAC area. The result shown in Figures 14 and 15, respectively, illustrates the proportional distribution of LOAC Groups and LOAC Super Groups for each Temporal Cluster. Overall, all the LOAC (both Super Groups and Groups) can be found in 11 Temporal Clusters. The demand for public transportation gradually decreases from the densely populated city centre to the outskirts of the city, which can be proved by circumstances in which the lowest proportions for each Temporal Cluster are predominately contributed by the

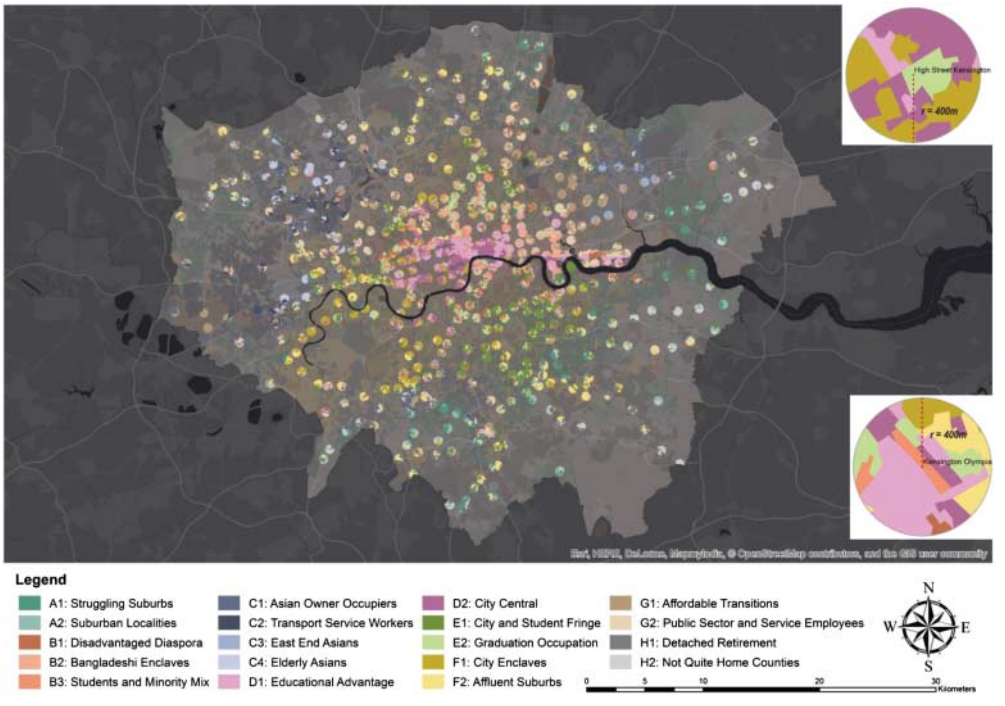

Figure 13. Residential Stations with 400 m Buffer and LOAC Group.

members categorised in 'A Intermediate Lifestyle' and 'H Ageing City Fringe', whose living areas are located far from the city centre. Conversely, 'D Urban Elite' and 'E City Vibe', LOAC Super Groups that are mainly located within Zone 1 and Zone 2 jointly employ more than half of the compositions for almost all Temporal Clusters (apart from Temporal Cluster 4), which indicates the constantly high demand for public transportation and briefly highlights the young generation are the majority occupying the tube travel. The remaining geodemographic classifications, especially 'B High Density and High Rise Flats' and 'F London Life-Cycle', whose need for public transport are also constant but more moderate, can accordingly be the target for the new transport initiatives since the choice for individuals from these social classes to either employ private or public as their travel mode is fuzzier. Again, although some other valuable information can be retrieved in more detail, calling back to the main concentration of this study, the next few sections will mainly focus on the interpretation of Temporal Clusters that exhibits potential 'Night Tube' travel pattern in the next section.

7. Transport planning – Night Tube

As mentioned before, there is a research gap linking empirical results to the practical transport planning, which therefore need to be bridged. In this section, the results can be implemented to assess the existing or planning transport policy. Here, specifically, the results are used to monitor the Night Tube Campaign run by TfL.

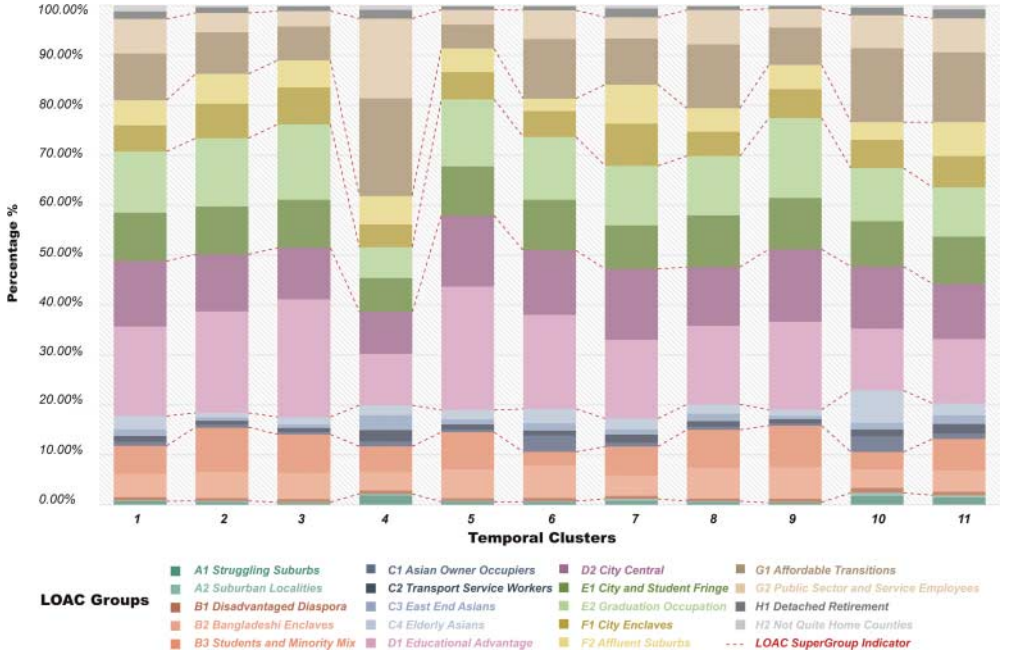

Figure 14. Proportions of LOAC Groups per Temporal Clusters.

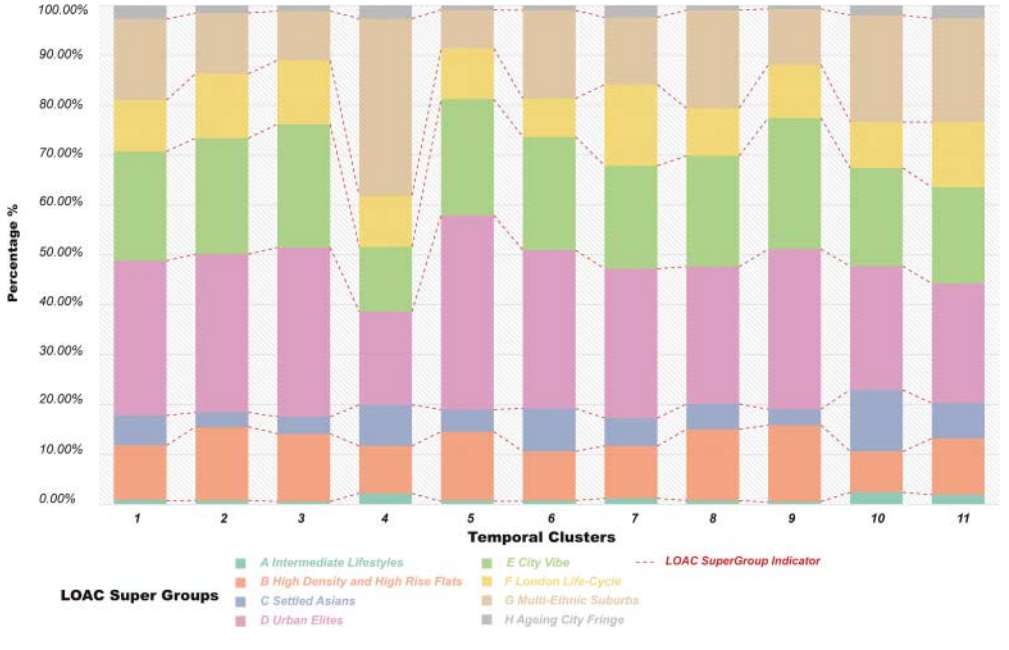

Figure 15. Proportions of LOAC Super Groups per Temporal Clusters.

7.1. Night Tube introduction

The Night Tube is one of the recent and promising campaigns launched by TfL which provides night-time services on the London Underground and London Overground systems to travellers on Friday and Saturday nights starting from the middle of 2015. TfL (2014) explicitly points out that the demand for Night Tube is growing more significantly than the daytime travels. TfL reports that there are already more than 50,0000 users of the Tube after 22:00 on Fridays and Saturdays. Implementing the Night Tube could significantly assist the London Underground service to meet the constantly growing demand for night travel. More fundamentally, the main purpose of Night Tube is to stimulate the development of London's night-time economy, reinforce the attractiveness, as well as enrich the existed transit services provided by TfL (e.g. Night Bus) (TfL 2014). According to TfL (2014), the Night Tube is estimated to have provided more than 2000 permanent jobs and significantly increase the Benefit Cost Ratio to about 3.9:1, up to now (middle of 2016). Currently (i.e. 2016), two London underground lines are offering the Night Tube service, namely Central and Victoria lines. In terms of the scheme, 'trains running on average every 10 minutes across the entire Victoria line', as for the Central Line, 'Trains running approximately every 10 minutes between White City and Leytonstone and approximately every 20 minutes between Ealing Broadway to White City and Leytonstone to Loughton/Hainault' (TfL 2014).

7.2. Who needs the Night Tubes?

As mentioned in Section 4.2, passengers categorised as Temporal Clusters 5 and 10 are more likely to have a weekend night travel pattern. Additionally, members of Temporal Cluster 6 that represents a quite random travel pattern also show some interests in weekend nigh travel. Based up Figures 14 and 15, it looks like three types of LOAC could benefit from night tube service.

Firstly, students and young workers who mainly live in the neighbourhoods that are categorised as 'D: Urban Elite', 'E: City Vibe', and some of 'B3: Students and minority mix' are gaining the most benefits from the Night Tube service. Particularly, the highest proportion of 'D: Urban Elite' can be identified in Temporal Cluster 5, within which more than half of the percentage is taken by the 'D1 Educational Advantage' group. This indicates that the major travel groups with this travel pattern are very likely to be the full-time students who are living in a centrally located communal establishment. This result also approves the weekend/night travel demand from the student groups, reflecting the demand for the Night Tube service, since the card holders are very likely to have student lifestyles.

Secondly, passengers classified as 'G Multi-Ethnic Suburbs' are also likely to be beneficial from the Night Tube services. Although the highest proportion taken by this LOAC Group is detected in Temporal Cluster 4, indicating the very early routine travel pattern, the proportions existing in Temporal Clusters 5, 6, and 10 are still evident. Summarised from the LOAC description, members of this group are from multiple ethnic backgrounds and can be characterised as hard-press living with children of school age.

Thirdly, some of the elderly people whose primary language is not English, e.g. 'C4: Elderly Asian', are also likely to be positively affected by the Night Tube service. The constancy of demand for Night Tube from these groups is quite typical, although the demand is not as high as the younger generations. Also, considering the fact that the elderly

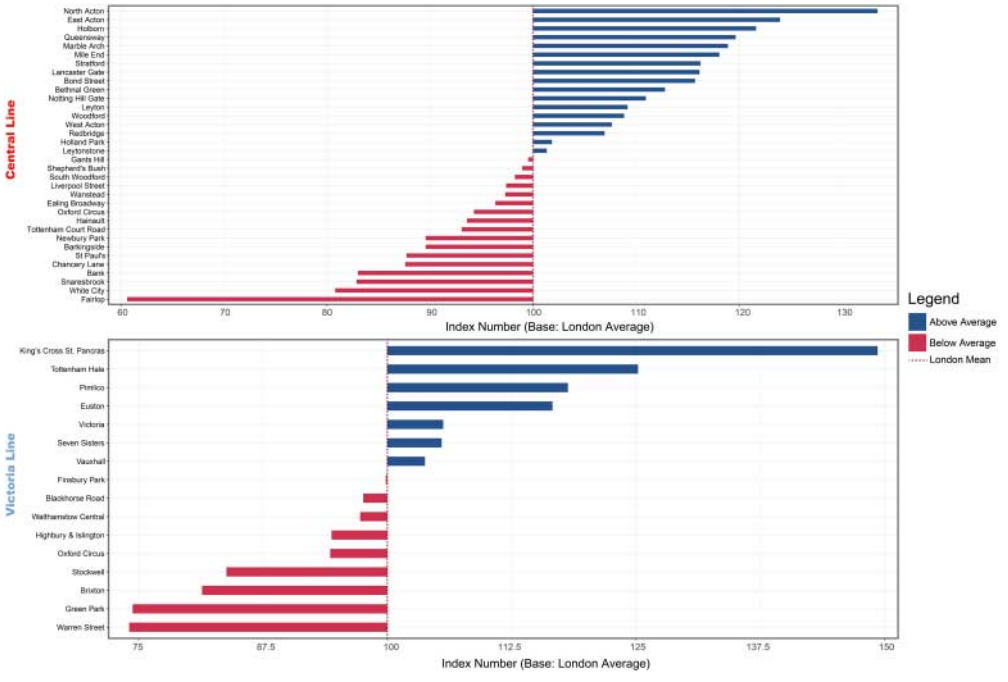

Figure 16. Index analysis for Central and Victoria Lines (London average = 100).

generation from these groups, the Night Tube service may significantly improve their mobility in terms of bridging social inequality.

7.3. Evaluating the Night Tube routes

Three Temporal Clusters (i.e. 5, 6, and 10) are used as our indicators for evaluating the Night Tube service. Since we have known the 'Residential Station' of passengers from these clusters, we can inspect the passenger composition of each of the stations alongside the two Night Tube routes.

An index analysis is conducted to examine the performance of these stations. Clusters 5, 6, and 10 together averagely occupy around 32.2% of the total subway and railway passengers in all stations located within Greater London, which is set as the base (i.e. index score = 100) of the index analysis. Hence, index scores for the night tube stations are calculated based on the London average, namely the percentage of passenger taken by Clusters 5, 6, and 10 divided by the base (i.e. 32.2%) and multiply 100. An index score of 200 is therefore double the average, and 50 would be half. Figure 16 shows the index scores for both Central Line and Victoria Line by stations. Given the London Average in the middle (represented by the red dashed line), stations can be divided into two groups, namely above average and below average. In this assessment, station whose index score is above the average (coloured in dark blue) can be viewed as 'Performed Well' in terms of coping with the demand of Night Tube travel; also, as the Temporal Clusters are purely derived from passenger's boarding transaction, station above the London average can be interpreted as a popular origin for Night Tube travel.

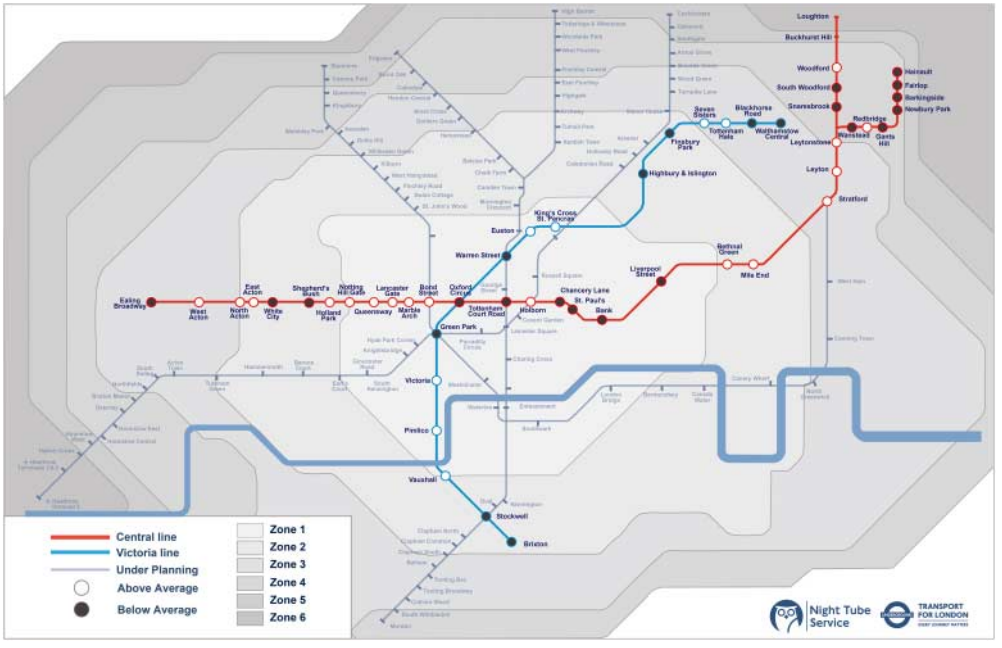

Figure 17. Night Tube map and performance evaluation.

Figure 17 illustrates the existing (in 2014) and under-planning Night Tube lines and the monitoring results garnered from the index analysis. The Night Tube service operating on the Central Line effectively copes with the night travel demand in 17 out of 34 stations. Particularly, North Acton, East Acton, and Holborn are the top three stations containing the highest demand for Night Tube services, whose index scores are more than 120, meaning that 20% higher than the London Average. Moreover, two continuous parts of 'Well-Performed' stations can be identified alongside the Central Line, namely between Holland Park and Bond Street; between Bethnal Green and Leytonstone. As for the negative part, a cluster of the station located in Zone 4 (i.e. the majority of stations located between Wanstead and Hainault) shows a below average pattern, indicating a relatively low demand for Night Tube. Additionally, index scores of stations between Chancery Lane and Liverpool Street, located in Zone 1, are below the average value.

The Victoria Line copes with the night travel demand in 7 out of 16 stations. Comparing to the proportion of 'Well-Performed' stations on Central Line, generally, the Night Tube service on the Victoria Line is relatively less successful. Stations located in Zone 1, such as Warren Street, Oxford Circus, and Green Park, are less attractive for late night travel. However, the major national railway stations, such as Euston, Victoria, and King's Cross, are categorised above the London average, illustrating a high demand for night travel.

It is important to notice that this implementation can be utilised either retrospectively or prospectively. As the Night Tube Service scheme was started in 2014, while the Temporal Clusters are produced by adopting the 2013 Oyster card data, the examination of the demand for night travel at each station can be seen as an estimation/preparation for the incoming campaign. Moreover, if the Temporal Clusters are created by using the

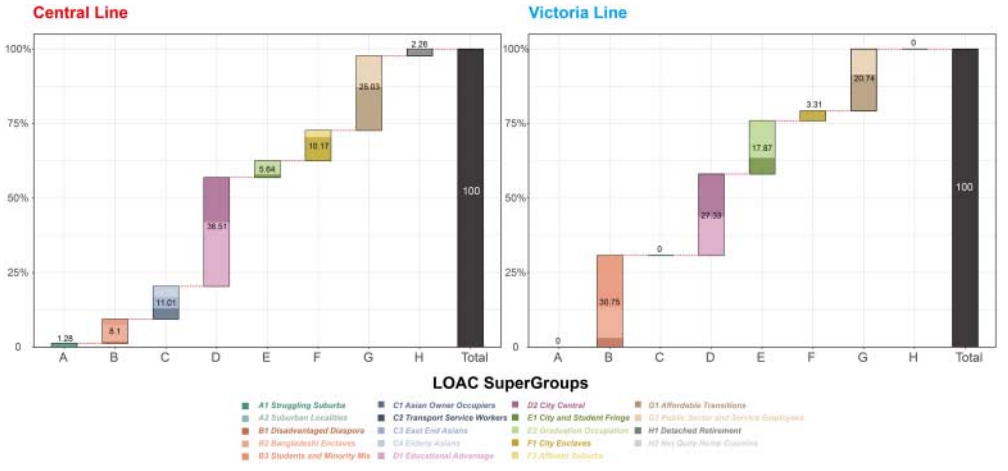

Figure 18. LOAC Super Groups and Groups composition in 'Well-Performed' stations on Central and Victoria Night Tubes.

latest dataset, the application could accordingly be seen as an assessment/a monitor of the existing transport service.

7.4. Who is benefitting from Night Tube service by Central and Victoria lines

As the 'Well-Performed' stations are identified, subsequently, we can examine the LOAC (Super Group) neighbourhood composition of each of the identified stations in order to analyse the multidimensional characteristics of residences who are more likely to exposure to the benefits provided by the Night Tube service. Figure 18, respectively, shows the LOAC composition for the 'Well-Performed' stations on Central and Victoria Lines. Generally, altogether, all of the LOAC Super Groups can benefit from the Night Tube service, however, in terms of LOAC Groups level, residences from 'A2: Suburban Localities' are not covered by the service. Service on the Central Line has a more comprehensive coverage as all of the LOAC Super Groups are involved, while the Victoria Line misses three Super Groups (i.e. A, C, and H). Both of subway lines, especially the Central Line, exhibit a high proportion of residences categorised as 'D: Urban Elites' and 'G: Multi-Ethnic Suburbs', which jointly occupies around 50% of the LOAC composition (Central Line: 61.5%; Victoria Line: 48.1%). Comparing to the Central Line, the Night Tube service operating on the Victoria Line potentially benefits more residences who come from a mix of ethnic backgrounds and not use English as their main spoken language, manifested by more than 30% of neighbourhoods located within the catchments of 'Well-Performed' stations on the Victoria Line are classified as 'B: High Density and High Rise Flats'.

The monitoring results broadly coincide with the estimations we mentioned in Section 7.1, i.e. D (D1), E, B3, G and C4 are the groups need night tube services.

8. Conclusions and future works

The advantages offered by the SCAFC system are not merely tied to the domain of physical transport infrastructures, and transport planning is also beneficial from the intelligence

of the big data from the SCAFC system, ranging from strategic, tactical, and operational levels. This paper focuses on the strategic level of study, aiming to create a methodological framework that systematically integrates truly personalised smart card data with open geodemographics together so as to improve the understanding of the traveller's behaviours in terms of interpretability and contextuality. Additionally, the empirical results from the data mining task are subsequently implemented to a real transport planning case study, i.e. the Night Tube, bringing the practical significance of our research.

Some limitations, however, do involve in this paper, which is therefore required to be concerned and overcome in the future works. First and foremost, the study period of this project is set as 28 days (ranging between 20 October 2013 and 17 November 2013), which was mainly driven by the consideration of balancing the size of data and the computational capacity. As the SCAFC system can capture transaction data passively and continuously, longer study period can accordingly be set, which is more likely to result in more representative and informative results. Moreover, this project only focuses on analysing the smart card data retrieved from the underground and some of the TfL's rail system, leading to the omission of considerable trip information generated from London bus. The reason for selecting transaction history only from the underground system is that the Oyster card data are better structured than the data from the bus trips. As the destination location information is not recorded by the Oyster card system in the bus journey, errors, inconsistent travels, and other uncertainties such as problems related to traffic mode interchanging are more likely to negatively affect the analysis. Although the clustering process is merely taking the boarding information of the underground system into account, the data (pre-) processing stages are evaluated by using the destination information, such as filtering out the long-distance commuters.

The second limitation is related to the criteria settings for identifying Londoner. The criteria used in this project are based on the heuristic method that primarily considers data structure (i.e. travel frequency) and the theoretical experience from similar studies. Moreover, in terms of finding their residing neighbourhood, a broad assumption was made in this paper. However, it should be admitted that passengers' home location sometimes can be further away (at least not within the 400 m walkable distance) from their most frequently used station. Several factors that required to be considered, for instance, they might travel to the station by bus or other modes of transports from their home. Therefore, to systematically achieve this goal, one of the alternatives is that to utilise the machine learning technique that can automate the workflow of residence identification, which might improve the quality and provide more accuracy in pairing local residence with contextual information (e.g. geodemographics) and accordingly improve the result interpretability. Furthermore, assuming one unique card ID as one passenger is not strict, to some extent, since passengers may have several cards to cope with different kinds of travel situations, e.g. purchasing different travel bundles according to how far they are going to travel. Additionally, it should be re-emphasised that due to lacking of some of the important attributes, e.g. travel propose, merely using the smart card transaction data without supplementary datasets is seemingly not adequate for a more comprehensive analysis. Although it is beyond the scope of the present study, one of the potential improvements for this perspective is to effectively estimate the travel purpose for the smart card data from the existing travel survey through the machine learning technique is one of the potential improvements.

The final limitation obviously exhibiting in this project is the timeliness. The Oyster card data obtained in the year of 2013, whereas the LOAC is purely based on variables captured by the 2011 census. Some scholars, such as Leventhal (2016), argue that the census-based geodemographics are currently still the mainstream of the geodemographic analysis because many of variables in the socioeconomic, demographic, and physical environmental domains are changing very slowly and therefore still effective for constantly using. However, London, one of the most dynamic cities all over the world, some of the census variables (e.g. occupation) do alter more swiftly than in other cities. Given the particularity of London, the input data need to be timely or coordinate with each other. However, it should be noticed that one of the innovative attempts of this study is to examine the feasibility of using open data to engage with practical urban planning question. Therefore, in other word, choosing LOAC data is one of the optimal examples to achieve this target rather than an irreplaceable action. In fact, many other datasets come from both public and private sectors can be used to enrich the information extracted from the smart card data. For instance, the newly released IMD 2015 can replace the position taken by LOAC in this project, which also can provide contextual socioeconomic information for the passengers, as IMD 2007 has already been successfully connected with the Oyster card data in Lathia, Quercia, and Crowcroft (2012). Also, the position of LOAC can be replaced by the classifications created with a specific purpose, such as the classification created by El Mahrsi et al. (2014).

Acknowledgements

The authors acknowledge the Transport for London (TfL) for the provision of the Oyster card data. The results presented and views expressed in this manuscript are the responsibility of the authors alone and do not represent the views of TfL.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is part of the Consumer Data Research Centre (CDRC) project supported by the UK Economic and Social Research Council (ES/L011840/1).

ORCID

Yunzhe Liu http://orcid.org/0000-0002-7189-3323
Tao Cheng http://orcid.org/0000-0002-5503-9813

References

Agard, B., C. Morency, and M. Trépanier. 2006. "Mining Public Transport User Behaviour From Smart Card Data." *IFAC Proceedings Volumes* 39 (3): 399–404.

Ali, A., J. Kim, and S. Lee. 2016. "Travel Behaviour Analysis Using Smart Card Data." *KSCE Journal of Civil Engineering* 20 (4): 1532–1539.

Awadi, H., T. Khemakhem, and M. Jemaa. 2012. *Applying LDA in Contextual Image Retrieval ReDCAD Participation at ImageCLEF Flickr Photo Retrieval* 2012. Accessed September 7, 2017. https://www.semanticscholar.org/paper/Applying-LDA-in-Contextual-Image-Retrieval-ReDCAD-Awadi-Khemakhem/5c6232065b89520cf7fe0c1314a18d0b774cf687.

Bagchi, M., S. Gleave, and P. White. 2003. *Use of Public Transport Smart Card Data for Understanding Travel Behaviour*. Accessed September 7, 2017. abstracts.aetransport.org/paper/download/id/1707

Blei, D., and J. McAuliffe. 2007. *Supervised Topic Models*. Accessed September 7, 2017. https://www.cs.princeton.edu/~blei/ papers/BleiMcAuliffe2007.pdf

Blei, D., A. Ng, and M. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3 (3): 993–1022.

Cha, Y., and J. Cho. 2012. *Social-Network Analysis Using Topic Models*. Accessed September 7, 2017. oak.cs.ucla.edu/ ~ cho/ papers/SIGIR12.pdf

Daraio, C., M. Diana, F. Di Costa, C. Leporelli, G. Matteucci, and A. Nastasi. 2016. "Efficiency and Effectiveness in the Urban Public Transport Sector: a Critical Review with Directions for Future Research." *European Journal of Operational Research* 248 (1): 1–20.

El Mahrsi, M., E. Come, J. Baro, and L. Oukhellou. 2014. *Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data*. Accessed September 7, 2017. http://www.comeetie.fr/pdfrepos/ urbcomp2014.pdf

Harris, R., P. Sleight, and R. Webber. 2005. *Geodemographic, GIS and Neighbourhood Targeting*. Chichester: John Wiley & Sons Ltd.

He, Z., Z. Wu, B. Zhou, L. Xu, and W. Zhang. 2015. "Tourist Routs Recommendation Based on Latent Dirichlet Allocation Model." *2015 12th Web Information System and Application Conference (WISE)*: 201–206.

Kieu, L., A. Bhaskar, and E. Chung. 2015. "Passenger Segmentation Using Smart Card Data." *IEEE Transactions on Intelligent Transportation Systems* 16 (3): 1537–1548.

Kusakabe, T., and Y. Asakura. 2014. "Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach." *Transportation Research Part C* 46: 179–191.

Lai, J., T. Cheng, and G. Lansley. 2017. "Improved Targeted Outdoor Advertising Based on Geotagged Social Media Data." *Annals of GIS* 23 (4): 237–250.

Lathia, N., and L. Capra. 2011. *Mining Mobility Data to Minimise Travellers' Spending on Public Transport*. Accessed September 7, 2017. http://www0.cs.ucl.ac.uk/staff/n.lathia/papers/lathia_kdd2011.pdf

Lathia, N., D. Quercia, and J. Crowcroft. 2012. *The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility*. Accessed September 7, 2017. http://researchswinger.org/publications/lathia12hidden.pdf

Lathia, N., C. Smith, J. Froehlich, and L. Capra. 2013. "Individuals among Commuters: Building Personalised Transport Information Services From Fare Collection Systems." *Pervasive and Mobile Computing* 9 (5): 643–664.

Lee, I., S. Oh, and J. Min. 2011. *Prospect of Technology for Public Transit Planning using Smart Card Data*. Accessed September 7, 2017. http://www.railway-research.org/IMG/pdf/poster_lee_inmook.pdf

Leventhal, B. 2016. *Geodemographics for Marketers: Using Location Analysis for Research and Marketing*. London: Kogan Page.

Longley, P., and A. Singleton. 2014. *London Output Area Classification (LOAC) Final Report*. Accessed September 7, 2017. https://files.datapress.com/london/dataset/london-area-classification/2011%20LOAC%20Report.pdf

Pelletier, M. P., M. Trépanier, and C. Morency. 2011. "Smart Card Data use in Public Transit: A Literature Review." *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568.

Perina, A., P. Lovato, V. Murino, and M. Bicego. 2010. "Biologically-aware Latent Dirichlet Allocation (BaLDA) for the Classification of Expression Microarry." In *Pattern Recognition in Bioinformatics*, edited by T. Dijkstra, E. Tsivtsivadzed, E. Marchiori, and T. Heskes, 230–241. San Diego: Springer.

Reed, C. 2012. *Latent Dirichlet Allocation: Towards a Deeper Understanding*. August 14 2016. http://citeseerx.ist.psu.edu/viewdoc/summary?doi{\mathsurround=\opskip$=$}10.1.1.399.2859

Seaborn, C. 2009. *Smart Card Data for Multi-Modal Network Planning in London: Five Case Studies*. Accessed September 7, 2017. https://trid.trb.org/view/1107481http://abstracts.aetransport.org/conference/index/id/15

Singleton, A., and S. Spielman. 2014. "The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom." *The Professional Geographer* 66 (4): 558–567.

Transport for London. 2014. *TfL 90993 – Impact of the Night Tube on London's Night-Time Economy*. Accessed September 7, 2017. http://content.tfl.gov.uk/night-time-economy.pdf

Transport for London. 2015. *London Travel Demand Survey (LTDS) Summary Report 2005/06 – 2013/14*. Accessed September 7, 2017. https://tfl.gov.uk/corporate/publications-and-reports/london-travel-demand-survey

Trépanier, M., and C. Morency. 2010. *Assessing Transit Loyalty with Smart Card Data*. Accessed September 7, 2017. https://www.researchgate.net/publication/234129247_Assessing_Transit_Loyalty_with_Smart_Card_Data

Trépanier, M., N. Tranchant, and R. Chapleaub. 2007. "Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System." *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 11 (1): 1–14.

Utsunomiya, M., J. Attanucci, and N. Wilson. 2006. "Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning." *Transportation Research Record: Journal of the Transportation Research Board*, 119–126.

Wang, Z., X. Li, and F. Chen. 2015. "Impact Evaluation of a Mass Transit Fare Change on Demand and Revenue Utilizing Smart Card Data." *Transportation Research Part A* 77 (1): 213–224.

Webb, V. 2010. "*Customer Loyalty in the Public Transit Context*." Master's Thesis, Massachusetts Institute of Technology.

Wu, Q., C. Zhang, Q. Hong, and L. Chen. 2014. "Topic Evolution Based on LDA and HMM and its Application in Stem Cell Research." *Journal of Information Science* 40 (5): 611–620.

Yu, C., and Z. He. 2016. "Travel Pattern Recognition Using Smart Card Data in Public Transit." *International Journal of Emerging Engineering Research and Technology* 4 (7): 6–13.

Zhang, F., N. Yuan, Y. Wang, and X. Xie. 2015. "Reconstructing Individual Mobility From Smart Card Transactions: a Collaborative Space Alignment Approach." *Knowledge and Information Systems* 44 (2): 299–323.

Zhao, J., V. Webb, and P. Shah. 2014. "Customer Loyalty Differences Between Captive and Choice Transit Riders." *Transportation Research Record: Journal of the Transportation Research Board* 2415: 80–88.