Characterising Passengers' Travel Patterns in London Public Transit

Yunzhe Liu* , Tao Cheng[1]

[1]SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College London, Chadwick Building, Gower Street, London, WC1E 6BT
{ Yunzhe.Liu@liverpool.ac.uk; tao.cheng@ucl.ac.uk}

November 10, 2016

KEYWORDS: Personalised Smart Card Data, Public Transport Planning, Latent Dirichlet Allocation Modelling, Travel Pattern Analysis

1. Introduction

Since the invention of the Smart Card in the late 1960s, this technology itself and its derivative applications have been increasingly benefiting and maturing in several industrial sectors, such as healthcare, banking, and government. In the public transit field, extensive researches that adopt data from the SCAFC system have been conducting accordingly, most of which are popularly categorised in the strategic-level study that aims primarily to understanding passengers' travel pattern via characterisation and classification (Pelletier, Trépanier, and Morency, 2011). Understanding transit patterns can be beneficial to assess the performance of the transit network, more accurately forecast travel demand, adjust service arrangement which copes with the ridership varying from both space and time.

Recent studies, e.g. Lathia et al. (2013), have identified the considerable differences of travel pattern existing between individuals, which contextually proves the critical role of personalisation. However, only a few studies are using truly personalised card data, which recognise each trip affixing to each passenger as an individual observation to appreciate the diversity of individuals. Furthermore, most of the studies characterising passengers utilise similarity-based clustering algorithms such as k-means, hierarchical agglomerative clustering, unsupervised Bayes (Strehl, Ghosh, and Mooney, 2000). These Euclidean distances based clustering algorithms often generate poor cluster results for high dimensional clustering (Strehl, Ghosh, and Mooney, 2000). In this context, this paper aims at seeking a new way to characterise the various patterns of individual passenger in public transit derived from the personalised smart card data. It creatively adopts a text mining technique, the generative model-based clustering technique, to conduct the temporal patterns clustering for London Oyster card passengers.

2. Data Description

The Oyster Card system, "the world's most popular" SCAFC system operating in London introduced by Transport for London (TfL) in 2003. According to the Mayor of London (2015), over 80% of the public journeys in London are taken by the Oyster holders who make approximately 30 million daily journeys on the public transit network. The Oyster Card data containing transaction information for the London Underground and Overground are extracted from the SQL database according to the "DATEKEY", "ROUTEDID", which are subsequently written into CSV table. Taking the computational capability into consideration, the study period is set to range for four weeks in 2013 (between 20/Oct/2013 and 17/Nov/2013). The size of raw dataset (CSV table) involves 60,889,787 Oyster card usage information that is made by nearly 4,248,774 passengers during the study period. It should be noticed that, in this case, each unique ID number is counted as an individual passenger, while some people do hold more than one Oyster card. Beside the card identifier (i.e. the "ID" number) and their time stamp (the "DATEKEY"), each travel history recorded by the smart card system contains the Start Time (i.e. the boarding time), End Time (i.e. the alighting time), Entry (the origin), Exit (the destination), and the Product Types (PPT). Table 1 presents the basic of the structure of Oyster Card data downloaded from the SQL database provided by TfL. For the simplicity, the data in the table are adapted from the original one and some columns are omitted.

* Current address: Department of Geography and Planning, University of Liverpool, Roxby Building, Liverpool, L69 7ZT

Table 1 Example of Subtracted Oyster Card Data

ID	Date	Entry	Exit	Product Type	Start Time	End Time
15184207	21/10/2013	Westminster	Putney	Travelcard	1265	1290
15987462	21/10/2013	Acton Town	Twickenham	No Ticket	651	681
15312602	17/11/2013	South Quay	Tower Hill	Staff Pass	1335	1369

3. Methodology and Results

The main objective of this section is to identify clusters of passengers exhibiting similar travel behaviour in the temporal dimension. To do so, the raw data containing every passenger's boarding records are firstly counted as accumulated frequency within the one-hour temporal intervals and subsequently processed into "weekly travel profile" for each smart card user, which is exemplified in Figure 1. Basically, different frequency of the trip indicates different usage/demand of public transit between individuals. For example, the first passenger (ID 4498) exhibits a peak-time travel behaviour as most of his travels are accumulated within the peak-time hours; While passenger (ID 6342) holds an off-peak and night travel behaviour.

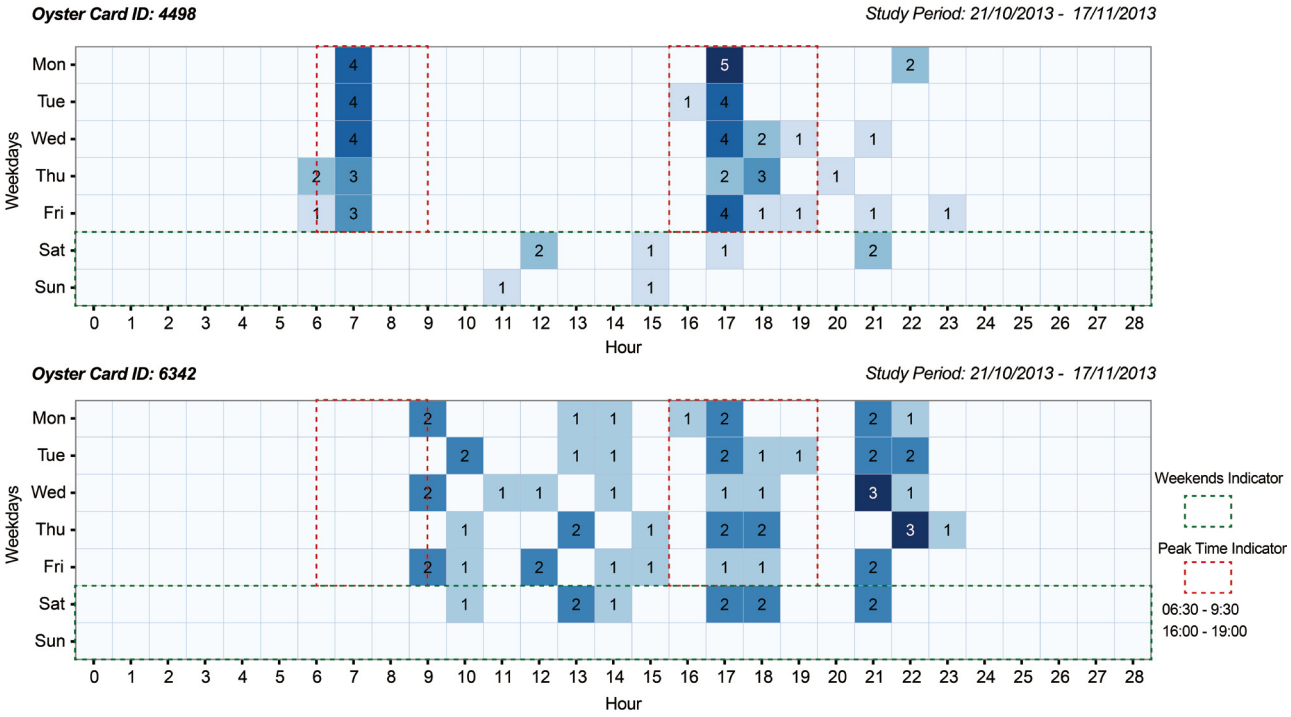

Figure 1. Weekly Travel Profile Example for Passenger ID 4498 and 6342

After creating all passengers with their own "weekly travel profile", LDA (Latent Dirchlet Allocation) text mining technique is proposed to be adopted to analyse the temporal travel pattern. LDA is one of the most commonly applied unsupervised topic modelling methods, which is described as "a generative probabilistic model for collections of discrete data such as text corpora" (Blei, Ng, and Jordan, 2003, p.993). LDA has been successfully adopted to analyse text information generated from various sources, for instance, journal articles (Wu et al., 2014), social media (Cha and Cho, 2012; Lai, Cheng, and Lansley, 2015), and contextual photos from Flickr (Awadi, Khemakhem, and Jemaa, 2012). Although the most commonly applied field for LDA model is text-based, LDA and LDA-based model can also assist to solve problems involving "data from domains such as collaborative filtering, content-based image retrieval and bioinformatics" (Blei, Ng, and Jordan, 2003, p.995). The combination of the "weekly profile" matrixes can be analogous to the content-based image retrieval (CBIR) problem. Under this perspective, an Oyster Card user is viewed as a "document", in which the "word" can be compared to a combination of their travel time period (reformatted as "Weekday_ Time", e.g. Monday_9.00) multiplied by the accumulated frequency.

There are eleven temporal clusters generated through the LDA modelling, which are also presented in a "weekly travel profile" manner in Figure 2. The dark colour indicates a high probability of the appearance of a "Temporal Interval", whereas the lighter colour represents a lower probability. More than half number of the temporal clusters (Cluster 2, 3, 4, 8, 9, and 11) exhibit a similar travel pattern that is the peak-time pattern, while some nuances do exist between these groups. For example, passengers characterised in Cluster 4 experience a very early travel pattern among the others within this general classification. Beside the clusters depicting a regular peak-time pattern, some other travel behaviours can also be identified according to the heatmaps. For instance, Cluster 1 and 7 jointly configure an off-peak noon travel pattern; members of Cluster 5 and 10 are more likely travel flexibly during the weekends and late night; Cluster 6 shows a quite erratic travel behaviour after the morning peak. Additional information that helps to interpret these temporal clusters can be extracted through inspecting the configuration of the product types applied by the passengers.

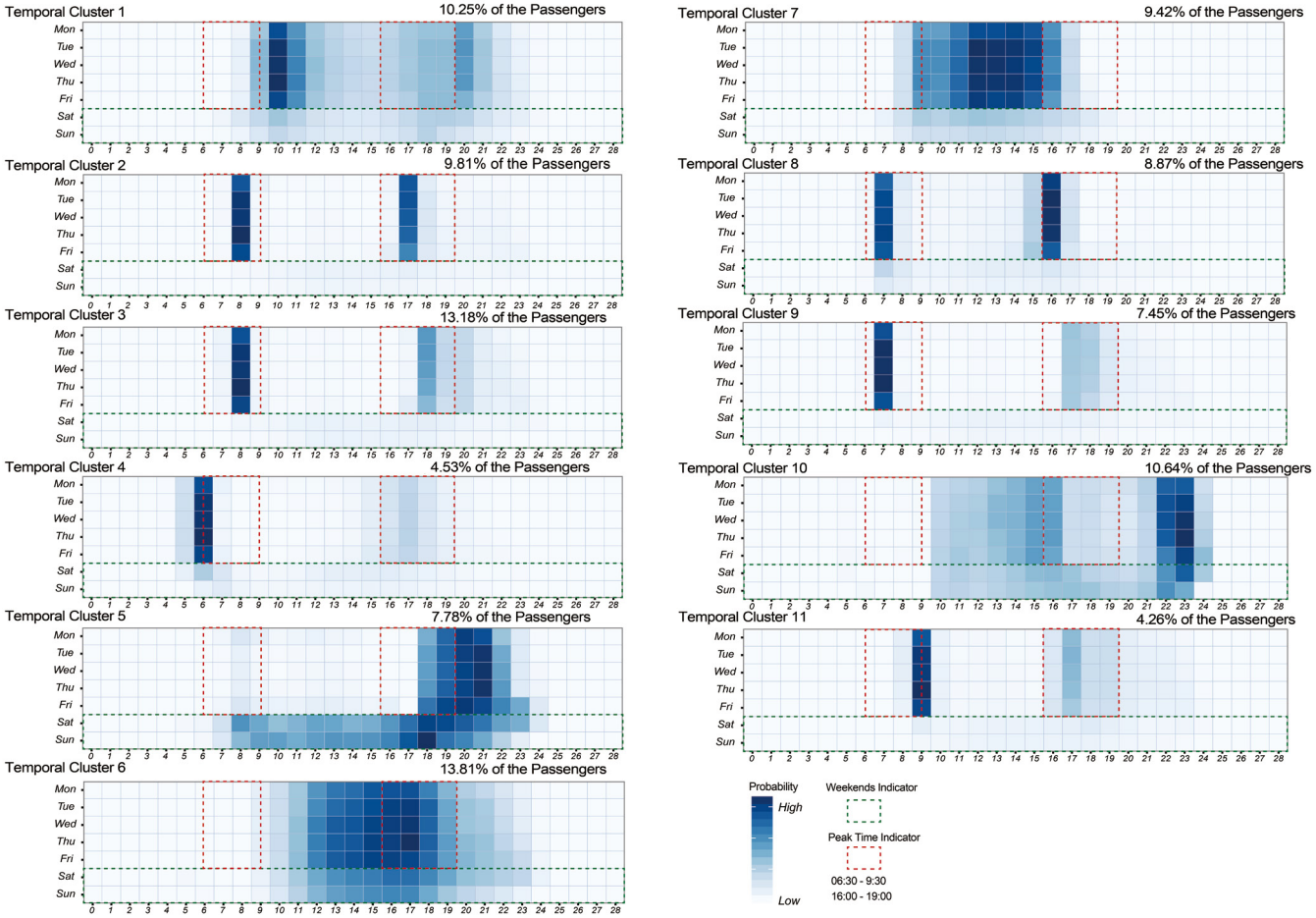

Figure 2. Heat Maps for the Eleven Temporal Clusters from LDA Modelling

4. Conclusions

By illustrating the application of LDA modelling in London's Oyster card data, this paper presents a novel approach of conducting personalised smart card analysis. The temporal clusters generated through the LDA do show some promising results, which can also be applied in some practical situations (e.g. assessing the Night Tube services). However, due to ethnic concerns, e.g. privacy, data are inherently anonymous and some key information depicting the purpose of travel are not included. To get better interpretation of the clusters through only relying on the information embedded in the smart card is certainly not enough. Accordingly, we expect to link the smart card with contextual information so as to improve the cluster interpretability.

5. References

Awadi, H., Khemakhem, T., and Jemaa, M. (2012). *Applying LDA in contextual image retrieval ReDCAD participation at ImageCLEF Flickr Photo Retrieval 2012*. Available: http://ceur-ws.org/Vol-1178/CLEF2012wn-ImageCLEF-AwadiEt2012.pdf

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*. 3 (3), p993-1022

Cha, Y. and Cho, J. (2012). *Social-Network Analysis Using Topic Models*. Available: http://oak.cs.ucla.edu/~cho/papers/SIGIR12.pdf

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

Lai, J., Cheng, T., and Lansley, G. (2015). *Spatio-Temporal Patterns of Passengers' Interests at London Tube Stations*. Available: http://leeds.gisruk.org/abstracts/GISRUK2015_submission_26.pdf.

Lathia, N., Smith, C., Froehlich, J., and Capra, L. (2013). Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*. 9 (5), p643- 664.

Mayor of London. (2015). *Annual Report and Statement of Accounts 2014/15.* Available: http://content.tfl.gov.uk/annual-report-2014-15.pdf

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), p557-568.

Strehl, A., Ghosh, J., and Mooney, R. (2000) Impact of Similarity Measures on Web-page Clustering. *In Workshop on Artificial Intelligence for Web Search. AAAI-2000*, p58-64. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.2110

Wu, Q., Zhang, C., Hong, Q., and Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science*. 40 (5), p611-620.

Biography

Tao Cheng is a Professor in GeoInformatics at UCL whose research interests span network complexity, Geocomputation, space-time analytics and Big data mining (modelling, prediction, clustering, visualisation and simulation) with applications in transport, crime, health, social media, and natural hazards. She is the founder and Director of SpaceTimeLab for Big Data Analytics (http://www.ucl.ac.uk/spacetimelab), a multi-disciplinary research centre at UCL that use integrated space-time thinking to gain actionable insight for government, public and industry.

Yunzhe Liu is currently a first year PhD student in the School of Environmental Sciences at University of Liverpool, who was graduated from the MSc Geographic Information Sciences at UCL with Tao Cheng's supervision. He is interested in researching about Big Data mining, Geodemographics, and urban geography/planning. He is now in the Geographic Data Science lab at University of Liverpool.