2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8310

Inferring Social-Demographics of Travellers based on Smart Card Data

Zhang, Yang and Cheng, Tao
SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Abstract

With the wide application of the smart card technology in public transit system, traveller's daily travel behaviours can be possibly obtained. This study devotes to investigating the pattern of individual mobility patterns and its relationship with social-demographics. We first extract travel features from the raw smart card data, including spatial, temporal and travel mode features, which capture the travel variability of travellers. Then, travel features are fed to various supervised machine learning models to predict individual's demographic attributes, such as age group, gender, income level and car ownership. Finally, a case study based on London's Oyster Card data is presented and results show it is a promising opportunity for demographic study based on people's mobility behaviour.

Keywords: *social-demographics; smart card data; travel variability.*

1. Introduction

Recently, social-demographic prediction based on individual's behaviour has become an emerging topic both in industry and academia. However, researchers mainly concentrate on people's online behaviour, such as web browsing (Hu, Zeng, Li, Niu, & Chen, 2007; Saste, Bedekar, & Kosamkar, 2017) and social network (Rao, Yarowsky, Shreevats, & Gupta, 2010; Vijayaraghavan, Vosoughi, & Roy, 2017). The discriminative power of people's mobility in the physical world has been overlooked, especially the travel behaviour via public transit.

Nowadays, public transit (PT) system plays a significant role in people's daily life. The modern PT network has widely equipped with the automatic ticketing system, called Automated Fare Collection (AFC) systems. As the popularity of AFC system, big data collecting through AFC records provide an opportunity to reveal hidden travel patterns by segmenting users to improve public transportation service quality and provide information for customers (Goulet Langlois, Koutsopoulos, & Zhao, 2016; J. Zhao, Qu, Zhang, Xu, & Liu, 2017). Traffic smart card has amassed a large amount of data to profile users' travel pattern, including travel mode, travel routes and time, and so much more. However, it lacks the social-demographic attributes of passengers to further explore 'who are the card carriers' and 'why they behave differently', which are crucial to better understand the users' requirement and travel patterns, offering a full picture of travel in urban area, which can help government make better transport planning, supply passengers with more personalized PT services and enhance the PT experience.

In this work, we devote to developing a framework for social-demographic inference based on smart card data (SCD). We first establish a feature extraction process to profile passengers' travel behaviours by using SCD. Then, several supervised machine learning algorithms are adopted to infer individual social-demographics, such as age level, gender, income level and car ownership. This study can also help us better understand the relationships between passengers' travel behaviours and social-demographics.

This paper is organized as follows. Section 2 briefly reviews the related works. Section 3 illustrates the framework and methodologies. Then, a case study based on London's Oyster Card dataset is carried out in Section 4 and results are presented and discussed. Finally, we summarise the conclusions, limitations, and future work in Section 5.

2. Related Works

A considerable number of existing works have demonstrated the impact of demographic factors on passengers' travel patterns. Early corresponding studies (Hanson & Hanson, 1981) suggest that the attribute of the individual or household are instrumental in shaping

daily travel decisions. As suggested in the recent literature (Yang et al., 2017), generally, an in-depth understanding of mobility patterns can be obtained by clustering people into distinct groups according to their demographic characteristics.

With regard to the analysis of the association between travel patterns and social-demographics, some works statistically describe the social-demographic features among diverse travel patterns (Goulet Langlois et al., 2016; Ortega-Tong, 2013), others illuminated the travel behaviour across social-demographic groups (Shobeiri Nejad, Sipe, & Burke, 2013). In recent year, some researchers have sought to illustrate the linkages between travel patterns and social-demographic variables on specific groups of travellers. For example, Siren and Hakamies-Blomqvist (2004) examine the association between selected demographics variables and mobility of elderly citizens in Finland. Results show that older persons experience reduced mobility mainly for leisure-related trips and their mobility was strongly associated with driving behaviour, education level and home location. van den Berg, Arentze, and Timmermans (2013) analysed the relationship between socio-demographic and social activity-travel patterns.

Most of the existing studies focus on the qualitative analysis of the relationship between the social-demographics and the individual travel behaviours. Alternatively, social-demographic inference or prediction has attracted increasing attention in the big data era. Social-demographic inference aims to characterise travellers, which can aid in transport planning, land use improvement and business settlement. To the best of our knowledge, there is no existing literature investigating the predictability of traffic smart card data for passengers' social-demographics inference. To fill this research gap, in this paper, we further study on what extent travellers' social-demographics can be inferred from their PT transaction records.

3. Dataset

3.1. London's Oyster Card Data

The dataset used in this study is a compilation of Oyster Card transaction records in London, UK, during the full year of 2013. There are two types of SCD, one from the tube system and the other from the bus system. Each 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 around 2.18 million journeys made by 9708 passengers, made up of 33.7% tube journeys and 66.3% bus journeys. Each transaction record contains the following fields: (1) unique ID, (2) boarding time, (3) alighting time (tube journey only), (4) boarding station, (5) alighting station (tube station only), (6) journey mode (bus or tube).

3.2. London Travel Survey Data

Transport for London (TfL) carried out the London Travel Demand Survey (LTDS), a continuous household survey of the London area, covering all London boroughs and the City of London. The LTDS is conducted based on the household for collecting individual or household demographic, social-economic and travel-related information. Around 8000 randomly selected households undertake the LTDS annually. All household members aged 5 and over need to complete the questionnaire. The unique Oyster card ID voluntarily provided by interviewed individuals in households for linking LTDS to Oyster card transaction records. The social-demographics data used in this study are provided in Table 1.

Table 1. Demographic attributes and corresponding categories

Attribute	Num. of labels	Categories and fraction
Age	3	Young (<30): 20.79% Adults (30 – 65): 58.04% Elder (>65): 21.17%
Gender	2	Male: 42.95% Female: 57.05%
Car ownership	3	Have no cars: 44.98% Have one car: 40.40% Have more than one car: 14.62%
Income level	3	Low income: 31.26% Middle income: 39.82% High income: 29.93%

4. Framework and Methodologies

The framework of social-demographic prediction is shown in Figure 1. The framework includes four main steps: (1) raw SCD preprocessing, (2) travel feature extraction, (3) social-demographic prediction, (4) performance evaluation. Details are given below.

Figure 1. Methodology framework

4.1. Data preprocessing

The first step in the analysis was linking the two datasets by using smart card ID. After that, We take 6354 frequent users' SCD as the primary sample. To deal with the missing alighting time and station of the bus journey, we refer to the method proposed by Jinhua Zhao, Rahbee, and Wilson (2007).

4.2. Feature extraction

A key issue in passenger segmentation is to construct accurate and comprehensive passenger profiles from SCD. In this study, various travel features are defined as to calibrate passenger profiles in order to differentiate the individual travel patterns. All features are categorized into four types, related to temporal variability (When), spatial variability (Where) and travel mode preference (How), respectively. The feature extraction process and explanation has been demonstrated in our previous work (Zhang & Cheng, 2017). Here, we just simply list the features generated from SCD in Table *2*.

Table 2. Travel features for passenger profiling

Subgroup	Feature	Description
Temporal features	AFTI	The average first start time on weekdays
	LFTI	The average last start time on weekdays
	MPT_NUM	the number of trips during morning peak (7:00am-10:00am)
	EPT_ NUM	the number of trips during evening peak (4:00pm-7:00pm)
	AVG_TRIP	The average number of trips per day
	ACTI_DAY	Active days in the whole year
	ACTI_DUR	Active duration in the whole year
Spatial features	AVG_ TIME	The average of tube trip time
	VAR_ TIME	The variance of tube trip time
	MAX_TD	The average radius travelled by tube per day
	AVG_TS	The average of the number of different tube stations used per day
	VAR_TS	The variance of the number of different tube stations used per day
	AVG_BL	The average of the number of different bus lines used per day
	VAR_BL	The variance of the number of different bus lines used per day
	AVG_INNER	The mean value of the inner zone number
	AVG_OUTER	The mean value of the outer zone number
Mode preference	TUBE_NUM	The total number of the tube journeys
	BUS_NUM	The total number of the bus journey
	MODE_T	How often a passenger changes the transport mode per day? (average)

4.3. Demographic prediction

After generating individual's travel features, with the social-demographics as the ground truth data, we utilize several popular supervised machine learning approaches to predict demographics. First, logistic regression (LR) is a natural choice for this type of a task. We also adopt random forest (RF), naïve Bayesian (NB) and multi-layer perceptron (MLP). We formulate the gender inference problem as a binary classification task and the others as the multi-class classification problems. The details of all these algorithms will not be discussed here.

5. Experiment Results

We report the social-demographic attributes inference for different classification methods, containing LR, RF, NB and MLP. The performance of prediction is evaluated by Accuracy (*Acc*), Precision (*Prec*), Recall (*Rec*) and F1 value (*F*1) (Qin et al., 2017). We conduct a 5-fold cross-validation and calculate the four performance metrics. Performance comparison results are shown in Table 3. Results show the best prediction accuracy of 'Age group', 'Gender', 'Income level' and 'Car ownership' can achieve 66.68%, 61.33%, 55.76% and 61.28%, respectively. Obviously, the prediction results can achieve a relatively high accuracy, but in some tasks, the scores of *Prec* and *Rec* are not very satisfied, since the class-imbalance problem has not been considered in these standard models.

Table 3. social-demographic inference performance comparison by four evaluation metrics

Model	Age				Gender			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF	0.6331	0.6317	0.6730	0.6661	**0.6133**	0.6242	0.5832	0.6030
LR	0.6647	0.5645	0.5598	0.6325	0.5950	0.5526	0.5630	0.5797
NB	0.4578	0.4432	0.5292	0.4915	0.5703	0.5660	0.5686	0.5677
MLP	**0.6668**	0.5743	0.5614	0.5936	0.6029	0.5456	0.5642	0.5936
Model	Income level				Car ownership			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF	0.5444	0.5502	0.5478	0.5536	0.6033	0.5753	0.6241	0.6439
LR	**0.5576**	0.5630	0.5633	0.5638	0.6001	0.4882	0.4924	0.5716
NB	0.5035	0.4683	0.5435	0.5133	0.3942	0.3808	0.4666	0.4346
MLP	0.5535	0.5593	0.5609	0.5586	**0.6128**	0.4825	0.4940	0.5864

6. Conclusion and Discussion

This study mainly explores the possibility of using smart card data to predict an individual's social-demographics. This framework helps to obtain people's social-demographic data without traditional travel surveys. The results presented in this paper can be baselines for further research.

However, there is still large room for improvement. First, for many demographic inference tasks, the dataset is category-imbalanced. More advanced techniques should be used to solve the class-imbalance classification problem. Second, some inference tasks are correlated. For example, the average income level of the elder is usually lower than that of the middle age people. Thus, it is necessary to investigate the correlations and use the multi-task learning method in the modelling to improve the prediction accuracy.

References

Goulet Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, *64*, 1-16. doi:http://dx.doi.org/10.1016/j.trc.2015.12.012

Hanson, S., & Hanson, P. (1981). The Travel-Activity Patterns of Urban Residents: Dimensions and Relationships to Sociodemographic Characteristics. *Economic Geography, 57*(4), 332-347. doi:10.2307/144213

Hu, J., Zeng, H.-J., Li, H., Niu, C., & Chen, Z. (2007). *Demographic prediction based on user's browsing behavior.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.

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

Qin, Z., Wang, Y., Cheng, H., Zhou, Y., Sheng, Z., & Leung, V. (2017). Demographic Information Prediction: A Portrait of Smartphone Application Users. *IEEE Transactions on Emerging Topics in Computing*, *PP*(99), 1-1. doi:10.1109/TETC.2016.2570603

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying latent user attributes in twitter.* Paper presented at the Proceedings of the 2nd international workshop on Search and mining user-generated contents.

Saste, A., Bedekar, M., & Kosamkar, P. (2017, 10-11 Feb. 2017). *Predicting demographic attributes from web usage: Purpose and methodologies.* Paper presented at the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

Shobeiri Nejad, S. M., Sipe, N. G., & Burke, M. I. (2013). *Retail travel behavior across socio-economic groups: a cluster analysis of Brisbane household travel survey data*.

Siren, A., & Hakamies-Blomqvist, L. (2004). Private car as the grand equaliser? Demographic factors and mobility in Finnish men and women aged 65+. *Transportation*

Research Part F: Traffic Psychology and Behaviour, *7*(2), 107-118. doi:http://dx.doi.org/10.1016/j.trf.2004.02.003

van den Berg, P., Arentze, T., & Timmermans, H. (2013). A path analysis of social networks, telecommunication and social activity–travel patterns. *Transportation Research Part C: Emerging Technologies*, *26*, 256-268. doi:http://dx.doi.org/10.1016/j.trc.2012.10.002

Vijayaraghavan, P., Vosoughi, S., & Roy, D. (2017). *Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning*. Paper presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Yang, Z., Lian, D., Yuan, N. J., Xie, X., Rui, Y., & Zhou, T. (2017). Indigenization of urban mobility. *Physica A: Statistical Mechanics and its Applications*, *469*, 232-243. doi:https://doi.org/10.1016/j.physa.2016.11.101

Zhang, Y., & Cheng, T. (2017). *Feature Extraction for Long-term Travel Pattern Analysis*. Paper presented at the GISRUK 2017.

Zhao, J., Qu, Q., Zhang, F., Xu, C., & Liu, S. (2017). Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems, PP*(99), 1-12. doi:10.1109/TITS.2017.2679179

Zhao, J., Rahbee, A., & Wilson, N. H. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5), 376-387.