Behavioural Analysis of Smart Card Data

N. Sari Aslam[a], T. Cheng [b], J. Cheshire [c]

[a b c] University College London(UCL), Department of Civil, Environmental and Geomatic Engineering, Gower St, London, UK

12 January 2018

Summary

Smart card data captured by automated fare collection (AFC) systems are a valuable resource for the analysis of human behaviour. The paper presents an approach of processing transit data for clustering analysis to identify user activities with similar characteristics. The effectiveness of the methods was evaluated using performance evaluation metrics. An external evaluation was used to compare the results with the ground truth. The results demonstrate that simple methods can produce good results when the input dataset used in the model is prepared and enriched with the most relevant features set.

KEYWORDS: Smart card data, trip purposes, machine learning, clustering methods

1. Introduction

Recent decades have seen an immense increase in the availability of digital traces of user data collected from sources such as GPS devices and mobile phones (Kong et al. 2009). These data sources have led to the emergence of new opportunities in the field of user mobility and behaviour research. Harnessing the potential of these data can lead the way to solving problems such as traffic and air pollution in big cities (Zheng et al. 2014). In this context, the data collected via AFC systems in transportation networks are a valuable resource that can be used to achieve a better understanding of human mobility and sustainable transportation.

Several studies have made use of the smart card transit data to extract the user segments and the behavioural contexts of the journeys. Different methods have been proposed to identify behavioural patterns and show the variability of travellers' activity patterns from smart card data (Kusakabe & Asakura 2014; Morency et al. 2007; Agard et al. 2006).

This study aims to carry out the preliminary work to build a framework of behavioural analysis for the identification of the purpose of the trip. Unsupervised machine learning methods such as clustering, along with a limited amount of expert labelled data is used to understand the meaning of the segments which are related to the individuals' activities.

[a]n.aslam.11@ucl.ac.uk,
[b]tao.cheng@ucl.ac.uk
[c]james.cheshire@ucl.ac.uk

2. Methodology

The analysis in this study makes use of the clustering methods, an important part of a wider area of Unsupervised Machine Learning. It is an effective form of analysis in the absence of a labelled dataset. The identification of the 'purpose of a trip' from journeys is a problem of conceptual clustering. Two or more activities belong to the same cluster if the cluster defines a concept to them, e.g. entertainment, pleasure or shopping. Figure 1 illustrates the workflow of the study in more detail.

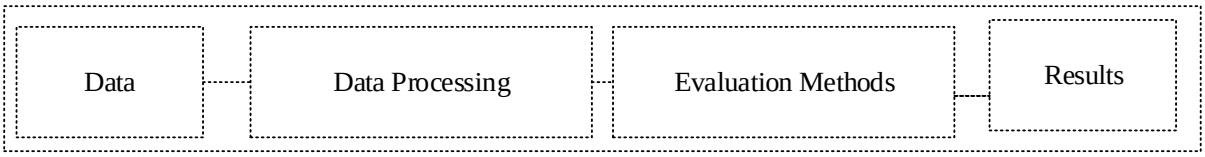

Figure 1: Workflow of the study

2.1. Data Description

The volume of data on the TfL network is extremely high with approximately 3 million journeys each day (TfL 2016). The data sample contains the transit record of completed journeys by 50 (randomly selected) individuals for October and November 2013. In addition to the unlabelled dataset from the TfL network, labelled data were collected from volunteers. Journeys data include attributes such as entry date/time, entry station, exit date/time, exit station and transport modes such as a London Underground, train, London Overground train, tram and bus. Bus journeys were excluded from the analysis as they do not contain the complete spatial and temporal information of the journey due to a single tap-in of the Oyster card.

2.2. Data Processing

Data processing aims to improve the accuracy of the clustering methods. As a first step, *activities were extracted* from the transit data for each TfL user (Bouman et al. 2013). Activities are represented by the time spent (duration) at a specific location (identified by the station) between two consecutive journeys.

The activity extraction step was followed by the identification of additional features *(feature extraction)* such as 'home location' and 'work location'. For the majority of the users, the day to day activities revolve around these key locations. The location has been identified using a heuristic approach (Hasan et al. 2012; Chakirov & Erath 2012). First and last journeys of the day present the necessary information in the identification of home, and the most time-consuming activity during the day is often the work location for the majority of users.

Additional features *(feature extraction)* include 'Activity From' and 'Activity To', which leverages the knowledge of home and work location along with information of the journeys preceding the activity location. The value range of these features is [0 – Home, 1-Work and 2-Other]. Additional features used in the analysis are 'Weekend Flag', 'Start Hour' and 'End Hour' of the activity. As a last step in the processing phase, features were scaled to normalise the range of independent input variables.

2.3. Evaluation Methods

Data points represent individual activities such as work, shopping and entertainment, defined by the features. *The clustering* is performed to the processed dataset of activities to isolate them into clusters that contain similar data points (activities), while the dissimilarity between groups is as high as possible. The clustering will enable the data points to be grouped by similar activities, and each cluster would be representative of the certain type of activity.

The aim is to identify the most suitable clustering technique for user activities dataset. The results are evaluated based on the Calinski-Harabaz and Silhouette criteria that gauge how dense and well separated clusters are. For Calinski-Harabaz, a higher value is indicative of a good score, whereas Silhouette ranges between 1 and -1 and values of 0 suggest overlapping clusters while values closer to 1 are considered a good score for highly dense clusters (Desgraupes 2013).

In addition to determining the density and separation of the clusters, Fowlkes Mallows Index (FMI) was calculated to measure the performance of clustering against the labelled data prepared by human experts. FMI values close to 0 specify that label assignments and clustering are largely independent whereas a value close to 1 indicates significant agreement (Desgraupes 2013).

Table 1: Clustering Indices applied in this study

Calinski-Harabaz Index	**Silhouette Index**	**Fowlkes Mallows Index**
$\text{CHI} = \frac{SSB}{SSW} * \frac{N-k}{k-1}$	$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)}$	$FMI = \sqrt{\frac{tp}{tp+fp} + \frac{tp}{tp+fn}}$
SSB is the overall between-cluster variance, SSW is the overall within-cluster variance, k is the number of clusters and N is the total number of observations (data points)	-1≤s(i)≤1 i is each datum, a(i) is the average distance of I and b(i) is the lowest average distance of i to all points	tp is the number of true positives, fp is the number of false positives and fn is the number of false negatives

An important consideration is to establish what a good number of components k to avoid the problem of overfitting. Bayesian Inference Criterion (BIC) is used in this study for model selection to establish the optimal number of components in the dataset.

3. Results

The results section is divided into the evaluation of the clustering methods using clustering evaluation metrics and external benchmark measures.

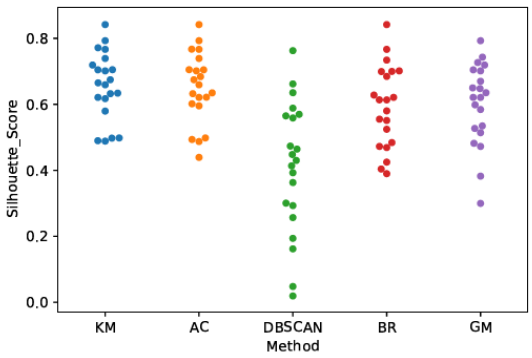

Figure 2: Silhouette score comparison of five clustering methods

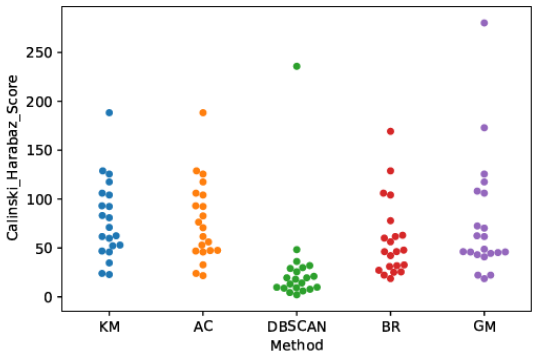

Figure 3: Calinski Harabaz score comparison of five clustering methods

Figure 2 plots the results of the Silhouette score, where the score ranges from 0-1. Each dot represents the results of the selected users' score. All of the clustering methods demonstrate a good score apart from DBSCAN. Similarly, Calinski-Harabaz (Figure 3) demonstrates a similar score range for all clustering methods other than DBSCAN.

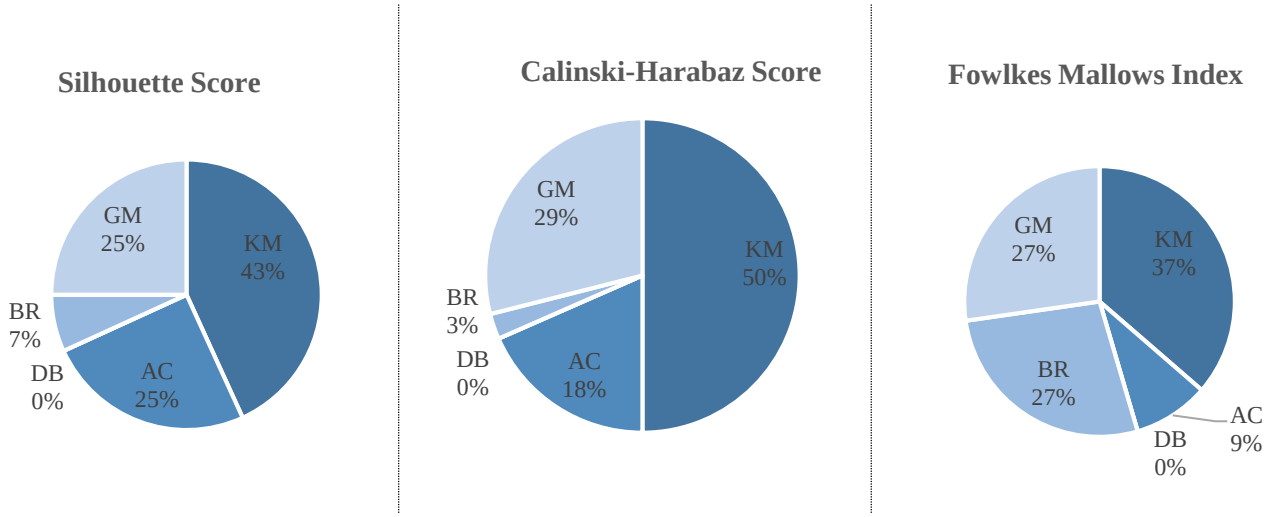

Figure 4: Percentage of the highest scoring method (KM = K means, AC= Agglomerative Clustering, DB= DBSCAN, BR= Birch and GM= Gaussian Mixture)

Figure 4 compares the results of clustering criteria of Silhouette, Calinski Harabaz Score and Fowlkes Mallows Index (FDI), highlighting the methods with the best results. In most cases, K-Mean and Gaussian-mixture produce the most optimum scores.

4. Conclusion and Future Work

The paper demonstrates that the smart card data is a good source of information for the study of human behaviour and mobility analysis. This paper presents a comparative analysis of clustering techniques. The study establishes that good results can be obtained even with simple learning methods if the input dataset is engineered to extract the most relevant features that highlight the distinguishing characteristics of the data.

The future work will aim to build on this study by incorporating the expert labelled data into a unified model for the identification of the purpose of the trip under a semi-supervised learning methodology. The study will benefit from integrating additional sources of user data such as GPS traces into the learning process. Additionally the location-specific features such as Point of Interest (POIs) can be incorporated into the dataset to provide a better understanding of the user behaviour.

5. Acknowledgements

I am grateful to the Economic and Social Research Council for funding my studentship at UCL.

6. Biography

Nilufer Sari Aslam is currently PhD student at Department of Civil, Environmental and Geomatic Engineering at UCL. Nilufer's research interests are big data analysis, spatial-temporal analysis and machine learning.

7. References

Agard, B., Morency, C. & Trépanier, M., 2006. Mining Public Transport User Behaviour From Smart Card Data. *IFAC Proceedings Volumes*, 39(3), pp.399–404. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1474667015359310.

Bouman, P., Kroon, L. & Vervest, P., 2013. Detecting Activity Patterns from Smart Card Data. *Bnaic*. Available at: http://bnaic2013.tudelft.nl/proceedings/papers/paper_90.pdf.

Chakirov, A. & Erath, A., 2012. Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport Smart Card Payment Data for Public Transport. *International Conference on Travel Behaviour Research*, (July).

Desgraupes, B., 2013. Clustering Indices. *CRAN Package*, (April), pp.1–10. Available at: cran.r-

project.org/web/packages/clusterCrit.
Hasan, S. et al., 2012. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151, pp.304–318.
Kong, Q.-J. et al., 2009. An approach to Urban traffic state estimation by fusing multisource information. *IEEE Transactions on Intelligent Transportation Systems*, 10(3), pp.499–511.
Kusakabe, T. & Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, 46, pp.179–191. Available at: http://dx.doi.org/10.1016/j.trc.2014.05.012.
Morency, C., Trépanier, M. & Agard, B., 2007. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), pp.193–203.
Zheng, Y.U. et al., 2014. Urban Computing : Concepts , Methodologies , and Applications. , 5(3).