Transportation Research Part C 79 (2017) 274–289

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Analyzing year-to-year changes in public transport passenger behaviour using smart card data ☆

Anne-Sarah Briand *, Etienne Côme, Martin Trépanier, Latifa Oukhellou

Université Paris-Est, IFSTTAR, COSYS-GRETTIA, 14-20 Boulevard Newton, F-77447 Marne-la-Vallée, France
CIRRELT/Polytechnique Montréal, Department of Mathematics and Industrial Engineering, P.O. Box 6079, Station Centre-Ville, Montréal, Québec H3C 3A7, Canada

ARTICLE INFO

Article history:
Received 14 March 2016
Received in revised form 27 January 2017
Accepted 28 March 2017
Available online 6 April 2017

Keywords:
Smart card
Passenger clustering
Mixture model
Public transit
Longitudinal analysis

ABSTRACT

In recent years, there has been increased interest in using completely anonymized data from smart card collection systems to better understand the behavioural habits of public transport passengers. Such an understanding can benefit urban transport planners as well as urban modelling by providing simulation models with realistic mobility patterns of transit networks. In particular, the study of temporal activities has elicited substantial interest. In this regard, a number of methods have been developed in the literature for this type of analysis, most using clustering approaches. This paper presents a two-level generative model that applies the Gaussian mixture model to regroup passengers based on their temporal habits in their public transportation usage. The strength of the proposed methodology is that it can model a continuous representation of time instead of having to employ discrete time bins. For each cluster, the approach provides typical temporal patterns that enable easy interpretation. The experiments are performed on five years of data collected by the Société de transport de l'Outaouais. The results demonstrate the efficiency of the proposed approach in identifying a reduced set of passenger clusters linked to their fare types. A five-year longitudinal analysis also shows the relative stability of public transport usage.

1. Introduction

The notion of mobility was developed near the end of the 1980s and gradually introduced individual practices and lifestyles into the analysis of transport demand, which made it necessary to redefine the meaning of the term. The aim of technically optimizing the straightforward spatial movement of goods and individuals (i.e., planning, flow, traffic, vehicle technology, etc.) has been supplemented, or even replaced, by the objective of obtaining a detailed understanding of the variation in individuals' ability to travel (accessibility), individuals' experience of daily travel conditions (comfort, sustainability), and/or even the role that mobility plays in individuals' lifestyles in terms of both actual and possible interactions. This paradigm shift should be understood as recasting the focus to concentrate more on individuals and less on vehicles and technologies. Consequently, mobility is now studied by economists, sociologists, urban planners, geographers and data scientists.

Traditionally, the analysis of mobility is based on travel surveys (household travel surveys (HTSs), origin-destination surveys, and cordon line surveys). These surveys have a number of advantages – for example, they cover all transport modes and trip purposes and generate meta-data regarding the respondents (gender, socio-occupational group, etc.). However, they are also expensive and consequently are undertaken fairly infrequently (typically, every five or ten years), which means that current developments and the public policies that aim to influence them are not closely monitored. In addition, HTSs are

☆ This article belongs to the Virtual Special Issue on "Smart cards, big data and travel behaviour".
* Corresponding author at: Université Paris-Est, IFSTTAR, COSYS-GRETTIA, 14-20 Boulevard Newton, F-77447 Marne-la-Vallée, France.

http://dx.doi.org/10.1016/j.trc.2017.03.021
0968-090X/

typically subject to statistical biases (due to sample size, representativeness, and non-responses), some of which can be partly compensated for by sampling theory techniques (such as correction and imputation). Another type of bias that is more difficult to quantify has been identified recently by Lathia and Capra (2011), who have shown that there are differences between the trips listed by participants in survey responses (when providing information about all their trips on the previous day) and the trips they actually made (as objectively shown by ticketing data).

Analyses of daily travel can rely on the growing number of digital records that are generated during trips. In many transportation networks, boarding logs are registered for all cards, which generates large-scale data. Anonymized ticketing data can indeed be used to implement new modelling approaches for urban mobility analyses even if they were not initially designed for this purpose. These data have a number of valuable intrinsic advantages, i.e., they are exhaustive (i.e., the data are distributed along the network on all transport modes), they are spatially and temporally precise, and they lack the response bias that plague survey data. However, to protect personal privacy, these data provide no socio-economic data on the user. Likewise, the data provide no information about trip purposes or the reasons behind the choice of a given travel mode or route. The use of smart card data to analyse urban mobility thus raises a number of issues and challenges that researchers in this field are working to address.

A great number of studies have been undertaken to examine passenger behaviour on transit networks, and a detailed review of related studies on this topic is provided in the next section. Transportation users' habits are of great interest to transit network operators because knowing these habits can help to predict affluence and enhance our understanding of the demand for transportation. Mobility patterns may also be used in urban modelling to provide simulation models with realistic public transport passenger behaviours. Many methods have been developed that identify such mobility patterns, most of which are based on clustering approaches. However, very few aims to analyse the evolution of cluster partitioning over a span of years (e.g. longitudinal analysis has been done in Morency et al. (2007)). The aim of this paper is to analyse the variability of passenger behaviour over time – in addition to passenger clustering – based on smart card data that spans a five-year period. Indeed, a transit network with few passenger activity changes from year to year will require less attention in terms of operator adjustments than a network with many passenger activity changes.

This paper makes the following contributions:

- It proposes that passenger cards should be clustered using temporal activities and should seek to partition passengers into small sets of clusters based on their usage habits of the transportation network, i.e., passengers with the same habits in terms of the hours in which they use the transport network will belong to the same cluster. In addition, this paper will demonstrate the replicability of the methodology initially proposed by Briand et al. (2016). The strength of this approach centres on its ability to take a continuous representation of time into account instead of having to employ the time binning used in most of the previous literature.
- In addition to offering a simple interpretation of cluster patterns, the aim here is to exploit the potential of the clustering methodology to perform a longitudinal analysis and, in particular, to study the evolution of passenger behavioural changes using their membership in different clusters over time. Thus, this study details the experimental results obtained using a real dataset covering a period of five years.
- Spatial characterization is also performed on the clusters identified. Shannon entropy is used to perform this analysis notwithstanding that socio-economic data on passengers are not available, except for fare type.

The remainder of this paper is organized as follows. Section 2 reviews related studies on ticketing data analysis and clustering. The real datasets used in our study and preliminary statistics are then described in Section 3. The approach to clustering passengers based on their temporal behaviour is introduced in Section 4. The experimental results are then presented and discussed in Section 5. Finally, concluding remarks and a discussion of future work are presented in Section 6.

2. Related studies

In recent years, the increased number of digital footprints that are collected reflecting our daily mobility has led to unprecedented approaches to innovative mobility studies (Zheng et al., 2014). From the use of GPS (Pang et al., 2013) to bike sharing data (Ahillen et al., 2015), growing amounts of information are available. Our study takes place in the public transportation field and focuses on mobility behaviours using smart card data collected via fare card collection systems. Although the information potential of smart card data has been attested to in previous studies (Bagchi and White, 2004; Park et al., 2008; Utsunomiya et al., 2006), their incomplete nature (lack of alighting locations, socioeconomic data, etc.) continues to motivate a substantial amount of research (Pelletier et al., 2011). Moreover, myriad topics have been addressed with respect to mining smart card data. We choose to categorize them three main topics, i.e., data completion and enrichment, prediction and passenger behaviour analysis.

2.1. Data completion and enrichment

Alighting data are of great interest for purposes of network analysis. Notably, the first estimations on Origin-Destination (OD) matrices based on assumptions of passenger mobility were undertaken in (Barry et al., 2002). Other studies have also focused on this topic by expanding prior hypotheses used to estimate destinations (Trépanier et al., 2007; Wang et al., 2011)