Transportation Research Part C 64 (2016) 1–16

Contents lists available at ScienceDirect

Transportation Research Part C

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

Inferring patterns in the multi-week activity sequences of public transport users

Gabriel Goulet Langlois [a], Haris N. Koutsopoulos [b], Jinhua Zhao [c,*]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115, United States
[c] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 7 September 2015
Received in revised form 15 December 2015
Accepted 23 December 2015
Available online 2 February 2016

Keywords:
Travel behavior
Smart card data
Activity sequence
User clustering
Public transportation
Data mining

ABSTRACT

The public transport networks of dense cities such as London serve passengers with widely different travel patterns. In line with the diverse lives of urban dwellers, activities and journeys are combined within days and across days in diverse sequences. From personalized customer information, to improved travel demand models, understanding this type of heterogeneity among transit users is relevant to a number of applications core to public transport agencies' function. In this study, passenger heterogeneity is investigated based on a longitudinal representation of each user's multi-week activity sequence derived from smart card data. We propose a methodology leveraging this representation to identify clusters of users with similar activity sequence structure. The methodology is applied to a large sample ($n = 33{,}026$) from London's public transport network, in which each passenger is represented by a continuous 4-week activity sequence. The application reveals 11 clusters, each characterized by a distinct sequence structure. Socio-demographic information available for a small sample of users ($n = 1973$) is combined to smart card transactions to analyze associations between the identified patterns and demographic attributes including passenger age, occupation, household composition and income, and vehicle ownership. The analysis reveals that significant connections exist between the demographic attributes of users and activity patterns identified exclusively from fare transactions.

1. Introduction

Diverse cities and the varied opportunities they foster are reflected in the heterogeneous travel patterns of the passengers of large urban transit networks. Beyond conventional 9-to-5 commuters, a variety of non-working routines and non-conventional work routines (for example driven by shift work, multi-employment, or self-employment) structure the activity patterns of public transport (PT) users. While these diverse activity patterns are typically considered on a daily basis, considering activity sequences across multiple days and weeks may reveal important differences among users. Segmenting users based on these differences is useful to gain a better understanding of the PT passenger population. From the provision of passenger information customized to specific user segments, to targeted travel demand management campaigns (Halvorsen, 2015), to service planning informed by the types of passengers traveling along different portions of the network, knowledge of the diversity among transit users provides opportunities to improve passenger experience and service provision.

* Corresponding author.

http://dx.doi.org/10.1016/j.trc.2015.12.012

Exploring heterogeneity in multi-week activity and journey sequences requires longitudinal observations of users. While conventional survey data contain detailed information about most aspects of a user's activity pattern (purpose, location, etc.), their costs typically proscribe large samples of users from being observed over long time periods. In contrast, smart card data provides a continuous stream of information about the PT travel of a large number of users. This information can be used to partially infer, and hence analyze, certain components of each user's general activity pattern (Lee and Hickman, 2014; Kusakabe and Asakura, 2014). Pelletier et al. (2011) present a review of research leveraging smart card data for such analysis.

Specifically related to this research, some studies focus on segmenting the travel patterns of PT users using smart card data. Ortega-Tong (2013) defined 20 different clustering variables related to travel frequency, journey times, origin–destination pairs, activity duration, fare type and public transport mode choice to identify 8 different user segments using the K-medoids algorithm. The resulting 8 groups were aggregated into four categories: non-exclusive commuters, exclusive commuters, non-commuter residents, and leisure travelers. Focusing on travel regularity, Ma et al. (2013) identified journey characteristics, including journey boarding time, bus route sequence, and bus stop sequences, frequently observed for the same user over a 1-week period in Beijing. From the number of days traveled and the number of frequent journey characteristics identified for each user, they define 5 clusters of varying regularity levels using the *k*-means++ clustering algorithm. Similarly to Ma et al. (2013) and Kieu et al. (2014) defined measures of temporal regularity and spatial regularity focused on weekday travel to segment public transport users in South East Queensland, Australia. They subjectively define segment boundaries for the resulting distribution and identify four groups: irregular passengers, regular OD pair passengers, habitual time passengers, and routine OD and time passengers. Finally, the early work of Morency et al. (2007) used *k*-means to identify typical patterns of bus boarding time in Gatineau, Canada.

While the work of these authors highlights the potential of smart card data to classify travel patterns, the approaches are limited in capturing the sequence within which each journey occurs. The clustering variables used by these studies are all derived from a scalar aggregation of a passenger's journeys which ignores the organization of multiple journeys over time.

Moving away from a fully scalar representation of user's travel patterns, El Mahrsi et al. (2014) use a vector of hour periods to represent the times at which each user is observed traveling. They identify 16 clusters of weekly temporal patterns by comparing the times at which users start journeys on each day of the week. While their approach preserves the order of hours within the week, it relies on aggregating multiple weeks of data to compute an average number of journeys for each hour. As such, it also ignores the sequence in which journeys are completed, and disregards all geographical information about journeys.

Important information about passenger's activity pattern is lost through such aggregation. As described by Hagerstraand (1970), and in line with the precepts of activity based travel theory, certain activity patterns include activities and journeys arranged in 'non-permutable' sequences. Activity patterns are defined not only by the attributes of the activities and journeys they are composed of, but also by the order in which these activities are organized. Abstracting this order may obscure sequence structure specific to certain passenger segments.

The research is organized around two objectives. First, we aim to develop a methodology leveraging smart card data to identify clusters of users sharing similar multi-week activity sequences. This methodology should provide an approach to investigate heterogeneity among passengers which can be applied systematically over time using continuously collected fare transactions. Second, we aim to provide empirical analysis of the heterogeneity among users of an extensive transit network through a large scale application of this methodology in London's transit system. This aim focuses on describing the underlying structure of activity sequences contained in each cluster and on exploring the socio-demographic attributes associated with each pattern.

In line with these objectives, the contribution of this work is twofold. From a methodological perspective, we provide a novel representation of travel patterns based on the longitudinal activity sequence of each user, and synthesize pervasive computing and data mining methodologies to identify trends from these sequences. From an empirical perspective, we analyze and expose the nature of heterogeneity among London's public transport users. We also provide evidence of significant associations between the patterns identified from traces of travel alone and socio-demographic attributes, by combining socio-demographic data about individual users to smart card records.

The remainder of this paper is organized as follows. Section 2 provides an overview of the methodology, and Section 3 describes the application of this methodology to London's user population. Finally, the conclusions and limitations of the work are discussed in Section 4.

2. Methodology

2.1. Representing longitudinal activity sequences

Central to the approach implemented in this research is the representation of each individual's travel pattern. In order to preserve the relationships between journeys and activities organized over multiple days, each user is represented as a time-ordered sequence of activities inferred from smart card data. Fig. 1 illustrates two such sequences, each associated with a different individual. Each column along the *x*-axis shows a day, covering a 4-week analysis period. Time of day is indicated