Sustainable Cities and Society 53 (2020) 101927

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data

Tianli Tang, Ronghui Liu*, Charisma Choudhury

Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Smart card data
Machine learning
Gradient boosting decision tree
Alighting bus stop

ABSTRACT

Origin-destination flow of passengers in bus networks is a crucial input to the public transport planning and operational decisions. Smart card systems in many cities, however, record only the bus boarding information (namely an *open* system), which makes it challenging to use smart card data for origin-destination estimations and subsequent analyses. This study addresses this research gap by proposing a machine learning approach and applying the gradient boosting decision tree (GBDT) algorithm to estimate the alighting stops of bus trips from open smart card data. It advances the state-of-the-art by including, for the first time, weather variables and travel history of individuals in the GBDT algorithm alongside the network characteristics. The method is applied to six-month smart card data from the City of Changsha, China, with more than 17 million trip-records from 700 thousand card users. The model prediction results show that, compared to classic machine learning methods, GBDT not only yields higher prediction accuracy but more importantly is also able to rank the influencing factors on bus ridership. The results demonstrate that incorporation of weather variables and travel history further improves the prediction capability of the models. The proposed GBDT-based framework is flexible and scalable: it can be readily trained with smart card data from other cities to be used for predicting bus origin-destination flow. The results can contribute to improved transport sustainability of a city by enabling smart bus planning and operational decisions.

1. Introduction

'*By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, notably by expanding public transport* (UN, 2015).'

The smart public transport system is an irreplaceable part of the 'Smart City' agenda (Ma et al., 2019). A well-planned and efficient bus system is a critical component of sustainable transport eco-system. The benefits of buses can be viewed from a range of different angles: (i) compared to cars, buses offer high capacity and low emission travel (Kwan & Hashim, 2016); (ii) buses are low-cost and quick to implement, relative to rail-based urban public transport systems such as metro; and (iii) bus operations have the flexibility to penetrate and respond to where and when the passenger demand is (Pei, Lin, Liu, & Ma, 2019). However, many of the urban bus systems suffer from poor images of unreliability, crowding, bus bunching, and generally low level of services (Berrebi, Watkins, & Laval, 2015; Bordagaray, dell'Olio, Ibeas, & Cecín, 2013). One of the important factors affecting their level of services and reliability is the temporal and spatial variability in the bus ridership distributions (Liu & Sinha, 2007; Sorratini, Liu, & Sinha, 2008). Understanding the factors driving the bus passenger behaviour and accounting for them to accurately estimate bus ridership are therefore the basic foundation for planning and operating a good public transport system (Hollander & Liu, 2008; Ibarra-Rojas, Delgado, Giesen, & Muñoz, 2015; Wu, Liu, & Jin, 2016; Wu, Liu, & Jin, 2017; Wu, Liu, Jin, & Ma, 2019).

Bus ridership, or the origin-destination matrix of bus travel demand, is affected by many factors. Existing studies in the literature have tended to focus on the population density and bus service provision of the area (Johnson, 2003; Xie, Jiao, An, Zheng, & Li, 2019), the socio-economic-employment characteristics of the traveller such as their car ownership, income, etc. (Paulley et al., 2006; Xie, An, Zheng, & Li, 2019). Bus passengers are exposed to outdoor weather environment during their travel, much more possible than car drivers and metro train users are. As a result, people may choose destinations and routes dif-

* Corresponding author.
E-mail addresses: ml15tt@leeds.ac.uk (T. Tang), R.Liu@its.leeds.ac.uk (R. Liu), C.F.Choudhury@leeds.ac.uk (C. Choudhury).

https://doi.org/10.1016/j.scs.2019.101927
Received 24 July 2019; Received in revised form 21 October 2019; Accepted 2 November 2019
Available online 07 November 2019

ferently under different weather conditions (e.g. small but closer shop versus larger but farther supermarket; going straight home versus stopping at an intermediate location to run an errand; route 'without transfer' but long walk versus 'with transfer' but no walking, etc.). In terms of empirical evidence, there have been recent interests in the weather impact on bus ridership on the demand side, and how bus operating strategies should respond to weather conditions on the supply side (see the review by Böcker, Dijst, & Prillwitz, 2013). For example, adverse weather is found to reduce the level of services of the bus system, while extreme weather (such as rainstorm and flood) could cause significant disruption to bus service (Hofmann & O'Mahony, 2005; Yin, Yu, Yin, Liu, & He, 2016). Similarly, passengers' travel behaviour, in terms of whether to travel, trip timing, route, and destination, could also be influenced by the different weather conditions. Arana, Cabezudo, and Peñalba (2014) show that wind and rain reduce trip-making, while mild temperature encourages passengers to travel. Aaheim and Hauge (2005) report that heavier precipitation and lower temperature shorten the distance people travel. Sabir (2011) points out that weather may change people's decision in the travel destination, especially for leisure travel. Liu, Susilo, and Karlström (2015) find that, in Sweden, both commuters and non-commuters are more willing to choose a closer destination in heavier rain. Hereby, we speculate that the passengers may change their alighting stops due to the different weather conditions, and we consider the ambient weather variables in our estimation.

Big data sources from the automatic data collection system can be utilised to support public transport planning and operation (Zannat & Choudhury, 2019; Zhang, Zhang, Sun, Zou, & Chen, 2018). For example, the automatic fare collection and automatic vehicle location systems offer new opportunity to understand the behaviour and patterns of bus ridership. With automatic data collection, the methods to estimate the ridership have been gradually shifted from the traditional manual survey, such as point check and ride check (Ceder, 2007), to data mining using readily available and large automatically collected data. There have been remarkable research interests recently in ways to extract the relevant and useful information from automatically collected data. Public transport users' smart card data from the automatic data collection system has been widely used as the most attractive resource to estimate bus ridership (Bagchi & White, 2005). Many of the bus systems, however, operate as a *single-tap* or *open* system, where passengers tap/swipe smart cards only at boarding, and thus we do not have information about their alighting. This raises challenges in using smart card data to directly derive bus origin-destination demand information, more specifically bus passengers alighting stops. Most of the existing research on this topic has so far only been able to estimate the alighting stops of regular commuter bus passengers, by approximating the alighting stops of their morning commuting bus journey as being the boarding stops of their evening return bus trip. In this paper, we attempt to provide a machine-learning-based framework to estimate the alighting stops for general bus trips, including regular and non-regular bus journeys.

The remainder of this paper is structured as follows. Section 2 reviews the methods in estimating the bus ridership and introduces machine learning techniques used in mining automatically collected data. A review of the weather factors affecting bus ridership is also presented. Section 3 introduces the case study network and the open smart card data used in this paper and highlights the limitation of applying the existing methods (trip chaining, for example) to our case. A machine learning approach based on the recently developed gradient boosting decision tree (GBDT) algorithm is proposed in Section 4 to solve the multi-class classification problem of estimating the alighting stops for the trips. Section 5 describes the trip features used in the model and designs the experiments whose results are presented in Section 6. Finally, Section 7 summarises our findings and suggests future research interests.

2. Literature review, research gaps and proposed improvements

2.1. Bus ridership and alighting stop estimation using open smart card data

Passengers' travel history can be tracked by the smart card data and then used for inferring their travel behaviour and ridership (Pelletier, Trépanier, & Morency, 2011). In the literature, there are two main approaches to estimate bus ridership from the open smart card data: attraction rate and trip-chaining model (see the review by Li, Sun, Jing, & Yang, 2018).

Briefly speaking, the attraction rate modelling estimates the attractiveness of a bus stop to the passenger, considering its boarding stop, the bus line of travel, and other relevant factors. Dou, Liu, and Yang (2007) propose a method to calculate the alighting probability at bus stops from the travel distance and passenger numbers. Another method in the attraction rate model is the reverse ridership method (Hou, He, & Zhang, 2012), which proposes that the proportion of the boarding passengers is equal to the proportion of the alighting passengers at the same stop in the reverse bus service. The attraction rate model can hence approximate the total bus passenger origin-destination ridership over a day, which is useful for long-term bus planning purposes. It is not, however, suitable to estimate the within-day (such as hourly) ridership which is critical for short-term or real-time bus operation and management. It is also not suitable for application at the individual smart card user level, which can be useful for policy testing purposes (e.g. testing the implication of a policy to provide fare discount for frequent travellers).

The second approach, trip-chaining model (Barry, Newhouser, Rahbee, & Sayeda, 2002), uses open smart card data to estimate linked trips and uses the results to establish the associated alighting stops. This method has been applied in extensive studies in New York (Barry et al., 2002), Chicago (Zhao, Rahbee, & Wilson, 2007) and London (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013). The trip-chaining model makes two strong assumptions: (i) each passenger gets on-board at the station where he/she alighted at the last trip; and (ii) each passenger's daily final alighting stop is the same as his/her first boarding stop of the day (Barry, Freimer, & Slavin, 2009). These assumptions put a limit on the applicability of the method. As summarised by Li et al. (2018), such a naïve trip-chaining model is not applicable to the following groups of passengers: (i) who use an untraceable mode of transport, for example taking a taxi on a leg of the journey; and (ii) who do not return to their origin stops. Since then, various studies have been making improvements to this naïve trip-chaining model. For the unlinked trips (e.g. those which involve a different untraced mode of transport in between bus trips), Trépanier and colleagues (He & Trépanier, 2015; Trépanier & Chapleau, 2006) suggest using passengers' historic travel pattern, and they propose a density-based method using arrival time and distances corresponding to each potential stops to identify the probability of alighting at that stop. For the daily trips which do not go back to the first boarding stop, Munizaga, Devillaine, Navarrete, and Silva (2014) find that many midnight trips (between 0–2 am) belong to trip chains on the previous day, and they suggest distinguishing the day at 4 am to reduce missed trips in recognising the trip chains.

One of the key processes in trip-chaining based models is to identify the most likely alighting stop among possible stops in close proximation. Trépanier, Tranchant, and Chapleau (2007) search the possible alighting stops by minimising the distance to the boarding stop of the next trip. Nunes, Dias, and Falcão e Cunha (2016) define a threshold of distance by the transaction fares system with distance-based fare structures. Munizaga and Palma (2012) replace the distance by a generalised time, while Nassir, Khani, Lee, Noh, and Hickman (2011) combine smart card records with a range of additional data sources, including bus timetable, automatic passenger counter and automatic vehicle location system, to identify the alighting stop of the last trip.

A common feature in these improved trip-chain models is that they rely on historical data to find the next boarding (alighting) stops.