
Article:
Tang, T, Liu, R orcid.org/0000-0003-0627-3184 and Choudhury, C orcid.org/0000-0002-8886-8976 (2020) Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. Sustainable Cities and Society, 53. 101927. ISSN 2210-6707

https://doi.org/10.1016/j.scs.2019.101927

Please cite the paper as:
Tang, T., Liu, R. and Choudhury, C., 2019. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. Sustainable Cities and Society. In press. https://doi.org/10.1016/j.scs.2019.101927

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

Tianli Tang [a], Ronghui Liu [a*], Charisma Choudhury [a]

[a] Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, United Kingdom
[*] Corresponding author
E-mail address: ml15tt@leeds.ac.uk (Tianli Tang); R.Liu@its.leeds.ac.uk (Ronghui Liu); C.F.Choudhury@leeds.ac.uk (Charisma Choudhury)

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.

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

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 and 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 et al., 2019). However, many of the urban bus systems suffer from poor images of unreliability, crowding, bus bunching, and generally low level of services (Berrebi et al., 2015; Bordagaray et al., 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 and Sinha, 2007; Sorratini et al., 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 and Liu, 2008; Ibarra-Rojas et al., 2015; Wu et al., 2016, 2017; Wu et al., 2019b).

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 et al., 2019b), the socio-economic-employment characteristics of the traveller such as their car ownership, income, etc. (Paulley et al., 2006; Xie et al., 2019a). 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 differently 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 et al., 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 and O'Mahony, 2005; Yin et al., 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 et al. (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 et al. (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 and Choudhury, 2019; Zhang et al., 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 and 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 et al., 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 et al., 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 et al. (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 et al., 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 et al., 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 et al., 2007) and London (Gordon et al., 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 et al., 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 and Trépanier, 2015; Trépanier and 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 et al. (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 et al. (2007) search the possible alighting stops by minimising the distance to the boarding stop of the next trip. Nunes et al. (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 et al. (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. Studies using the attraction rate and trip-chaining models have so far been mainly based only on the smart card data, with some incorporating the network characteristics into the studies. In reality, there are many other factors that can affect the ridership choices made by the passengers, such as the effect of weather on passengers' habitual travel behaviour (see section 2.3 for details), special events, etc. This paper attempts to incorporate such weather-related factors in the estimation of bus ridership to address this research gap.

2.2 Data mining using machine learning technique

Although the development of automatic data collection system offers detailed data on various aspects of the public transport system, the abundance of available data challenges the traditional data mining methods such as classification, clustering, and regression analysis. Machine learning as a data mining method is shown to be able to handle high-dimensional and multivariate data in a complex, dynamic and even chaotic system, and to identify the patterns in the data and the relevant influential factors (Witten et al., 2016; Wu et al., 2019a; Wu et al., 2016).

Recently, there has been an increase in the number of studies trying to bring machine learning to the analysis of public transport data (see examples listed in Table 1). For example, Yu et al. (2011) apply several machine learning models: support vector machine, artificial neural network, k nearest neighbours algorithm and linear regression to predict the bus arrival time from the bus running time on different routes. Corman and Kecman (2018) build Bayesian networks to predict train delays in real-time from a live data stream. Meanwhile, how to use the data in the automatic fare collection system is also an interesting topic for many studies.

There are two types of automatic fare collection systems: the closed automatic fare collection system, which records both the boarding and alighting information, and the open automatic fare collection which records only the boarding information. For the closed (mostly metro) automatic fare collection system, there have been extensive studies applying machine learning to forecast the metro passenger flow from smart card data via the networks of hybrid empirical mode decomposition and back-propagation neural network (Wei and Chen, 2012), multiscale radial basis function network (Li et al., 2017) and long short-term memory neural networks (Liu et al., 2019).

There are relatively limited studies of machine learning application to open automatic fare collection systems. Toqué et al. (2016) infer the alighting stops using the trip-chaining model to predict the origin-destination matrices at stop level in 15-minute windows using long short-term memory neural networks. Jung and Sohn (2017) develop a deep learning model to predict the alighting stops for each transaction, taking account of variables on the land-use near the boarding and candidate alighting stops. The key literature on machine learning applications on public transport research is summarised in Table 1.

Table 1 Selected literature on the applications of machine learning on public transport research.

Literatures	Public transport modes	Targets	Machine learning models	Data resources
Yu et al. (2011)	bus	arrival time	SVM, ANN, k-NN, LR	real-time traffic data
Corman and Kecman (2018)	train	delay	Bayesian network	scheduled and real timetable
Liu et al. (2019)	metro	passenger flow	LSTM	closed smart card data; weather and holiday events
Li et al. (2017)	metro	passenger flow	MSRBF	closed smart card data
Wei and Chen (2012)	metro	passenger flow	EMD - BPN	closed smart card data; (holiday events)

Literatures	Public transport modes	Targets	Machine learning models	Data resources
Toqué et al. (2016)	bus	origin-destination matrix	LSTM	open smart card data
Jung and Sohn (2017)	bus	destination	Deep learning	open smart card data; land-use
This study	bus	alighting stops	GBDT	open smart card data; travel history; weather conditions

Abbreviations: support vector machine (SVM), artificial neural network (ANN), k nearest neighbours (k-NN) algorithm, linear regression (LR), long short-term memory neural networks (LSTM), multiscale radial basis function network (MSRBF), empirical mode decomposition (EMD), back-propagation neural network (BPN), gradient boosting decision tree (GBDT)

Our study proposes a machine learning method to estimate bus passengers' alighting stops from open smart card data where only the boarding stops are observed. Compared to the existing literature on the subject (Table 1), we employ an innovative new data mining approach, the gradient boosting decision tree (GBDT) model, to estimate the alighting stop for every bus trip recorded in the smart card data. Furthermore, we incorporate passengers travel history and their travelling environment – in terms of the ambient weather conditions, into the estimation. We examine the impact of these additional variables on the performance of the estimation.

2.3 Weather impacts on bus ridership

As noted in the Introduction, there has been existing research that established relationships between the varying weather conditions and overall bus ridership. Table 2 summarises the key literature that examines the weather impact on public transport ridership. Generally, precipitation is found to be one of the most important factors affecting bus ridership. Hofmann and O'Mahony (2005) show that the number of smart card trip records decreases with rainfall. Similar conclusion, drawn by Saneinejad et al. (2012), is that commuting trips of all modes, including buses and private cars, is negatively affected by precipitation. In contrast, Singhal et al. (2014) find that urban transit ridership increases on snowy days as the poor driving conditions tend to shift people from private cars. Additionally, Guo et al. (2007) investigate how rain and snow affect the public transport ridership and report that rain and snow tend to reduce the ridership for both bus and rail, but heavy snow might actually increase the rail ridership. The temperature and wind are other important sensory weather variables. Stover and McCormack (2012) report that in a cold climate, the ridership decreases as the temperature dropped and increases on warmer days. Similarly, Guo et al. (2007) discover a significant

positive impact of temperature on transit ridership in warm weather but find no correlation between ridership and temperature in cold weather conditions. They speculate that the impacts are not caused by the prevailing temperature but the temperature changes and human perceptions. Kashfi et al. (2015) report that temperature has an insignificant effect on the daily bus ridership. For the impacts of wind, Arana et al. (2014) show that the wind, together with rain, leads to a reduction in transit ridership. Singhal et al. (2014) also find a negative effect of wind on hourly subway ridership, but the effect is not significant on daily ridership. Guo et al. (2007) note that increasing wind speed reduces bus ridership, but it has a negligible impact on rail ridership. Besides these three weather variables (temperature, rainfall and wind), Zhou et al. (2017) analyse the impact of weather condition on bus and metro system together with relative humidity and air pressure. Their study reports that increase in humidity, wind and rainfall is generally associated with a certain degree of transit ridership decrease while their degree and the significance of the impact vary from one weather variable to another, and between weekday and weekend. Wei et al. (2019) consider the weather impact on weekday and weekend travel by bus, train and ferry. They find that, unsurprisingly, ferry is mostly affected by bad weather. Poor weather conditions do not appear to affect train journeys during the weekdays, but to reduce train journeys made during weekends. They find that bus trips are negatively affected by rainfall, but not affected by temperature.

The most common method to quantify the impact of weather on public transport ridership in these studies is regression modelling, and the most commonly used independent weather variables are temperature and precipitation. Guo et al. (2007) take into account discretised weather variables, such as heavy or light precipitation, and warm or cool temperatures. Zhou et al. (2017) replace the absolute value of weather variables by its deviation to the average condition. Wei et al. (2019) consider the interplay of different weather variables in the model to formulate the complex impact of weather conditions on the ridership.

Table 2 Correlations between public transport travel demand and weather conditions.

Literatures	Public transport modes	Precipitation		Temperature		Wind
		Warm	Cold	Warm	Cold	
Hofmann and O'Mahony (2005)	Bus	↓		×		×
Saneinejad et al. (2012)	All	↓		↑		×
Singhal et al. (2014)	Metro	↑		− (daily) ↑ (hourly)		− (daily) ↓ (hourly)
Stover and McCormack (2012)	Bus	↓		↑ (winter)		↓ (except summer)

Literatures	Public transport modes	Precipitation		Temperature		Wind
		Warm	Cold	Warm	Cold	
Guo et al. (2007)	Bus	↓	↓	↑	–	↓
	Rail		↑ (heavy snow)			–
Kashfi et al. (2015)	Bus	–		–		×
Arana et al. (2014)	Bus and train	↓		↑		↓
Zhou et al. (2017)	Bus and metro	↓		↓		↓
Wei et al. (2019)	Bus	↓		–		–
	Train	– (weekday) ↓ (weekend)		– (weekday) ↓(weekend)		–
	Ferry	↓		↑ (weekday) – (weekend)		–

Note: ↓: the negative correlation; ↑: the positive correlation; –: no correlation; ×: not discussed.

In this paper, we employ a machine learning method to independently evaluate the relative importance of different weather variables on bus ridership and consequently incorporate the important weather effects in the bus ridership prediction.

3 Data sources

3.1 Network description and data sources

In this study, we estimate the alighting stops of individual bus trips made using the smart card in the city of Changsha. The city, in the central south of China, is separated by the Xiangjiang River: the city's Central Business District (CBD) lies east of the river, while the west is principally residential zone. There are more than 200 bus lines in Changsha, operated by three bus companies. The study network, shown in Figure 1, is a subset of the Changsha bus network and includes seven bus lines, all operated by the same bus company. Despite it being only a sub-network of the city, the case network covers the key public transport interchanges in the city: the three bus terminals and two rail stations, as well as the three river crossings that connect the major geo-economical centres of the city. The seven bus lines are also representative in service characteristics, including long-distance and sparse-stop lines (Line 063 and 168), long-distance and dense-stop lines (Line 147), short-distance and sparse-stop lines (Line 006 and 007) and short-distance and dense-stop lines (Line 123 and 150).

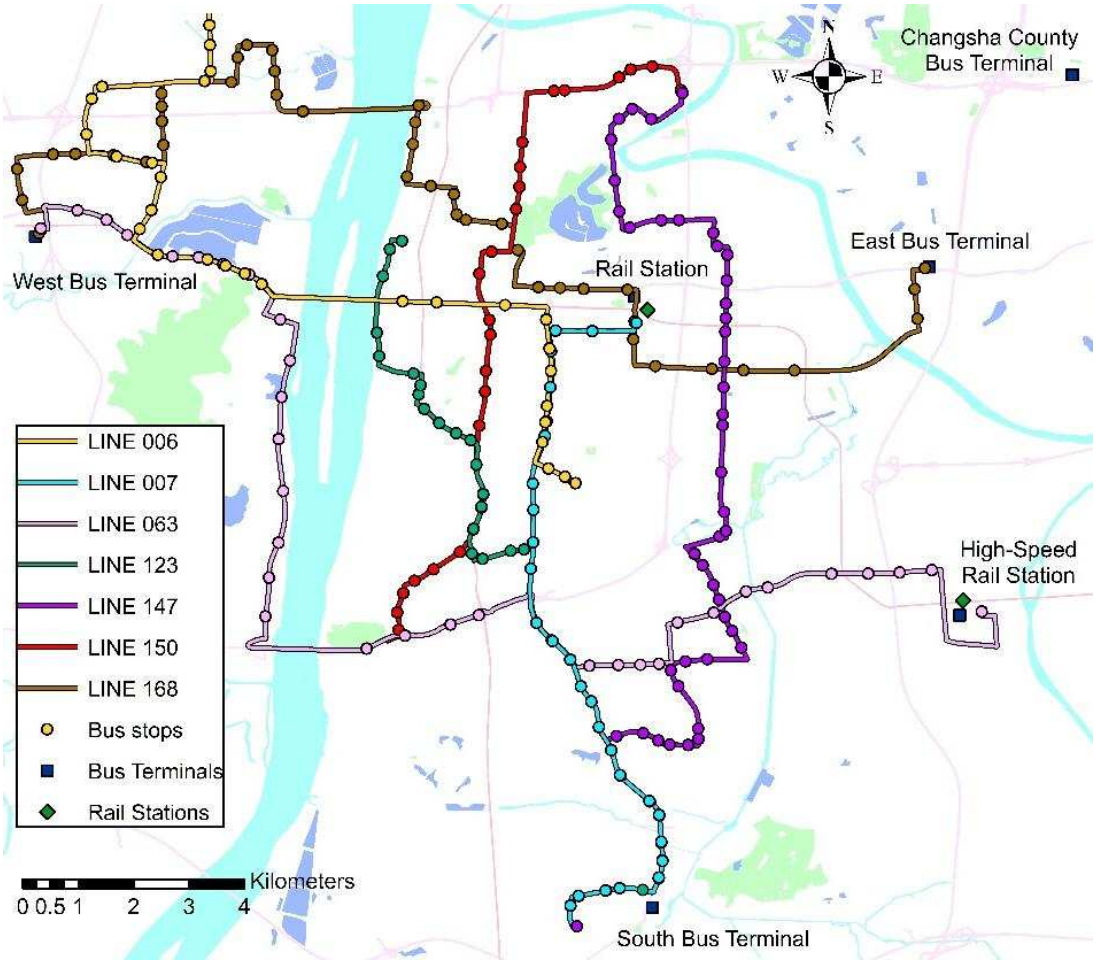

Figure 1 The case study bus network in Changsha, China.

Changsha's smart card system is a typical open automatic fare collection system, where passengers swipe cards only at boarding. The system records information on smart card ID, boarding time, line and the bus vehicle boarded. All buses in Changsha are fitted with GPS trackers, which record the vehicle ID, the longitude and latitude of the vehicle location every 10 sec.

Six months (April to September 2016) of smart card data is made available to this study, in which 12 days' data was missing. There are 17,159,076 smart card records in total or roughly 80000 records per day on average. The number of daily smart card records for the study period is shown in Figure 2. It shows that the ridership in holidays is markedly lower than that in working days. It is also noticeable that the four weeks between the end of July and mid-August have low ridership when the city typically experiences heatwaves with an average temperature around 35ºC and maximum temperature of 40 ºC. Besides the high temperature, July and August are also the school summer holiday, which may also contribute to a decrease in bus ridership.

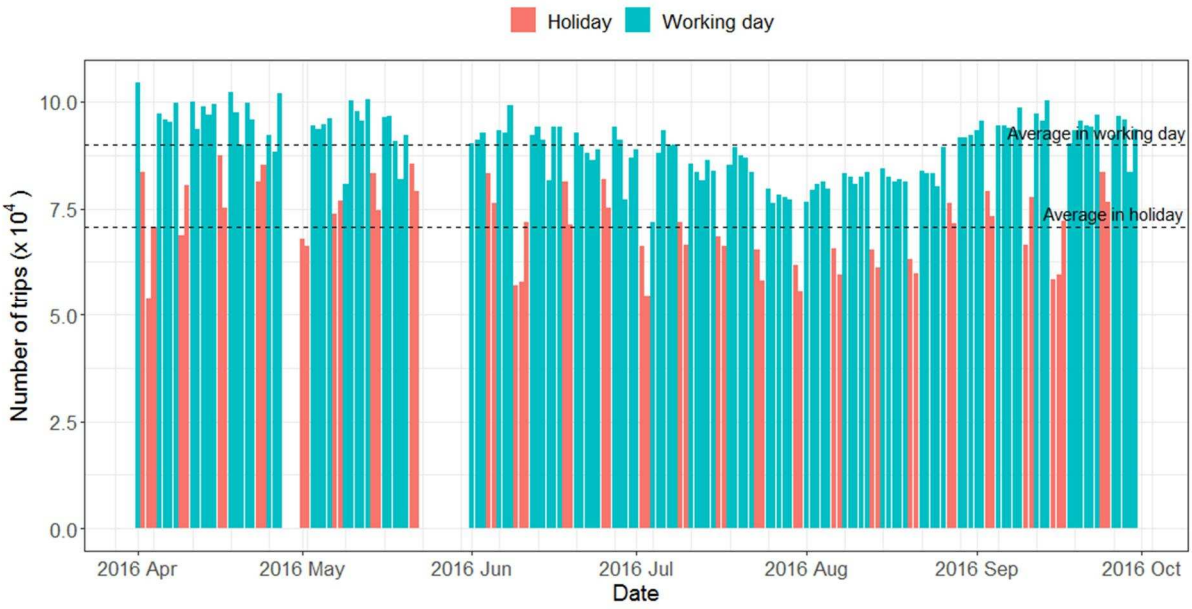

Figure 2 The number of smart card trips by day.

As part of a wider study of bus patronage, we are also interested in the effect of weather on passengers' travel demand. We acquired weather data for those six months. The weather data includes hourly measurements of temperature, precipitation, pressure, humidity, visibility, wind speed, and an indicator/register of the type of weather event of the hour. The weather events registered include, for example, clear, cloudy, thunderstorm, rain, etc. Figure 3 illustrates the hourly changes in weather events, overlaid with the number of smart card trips during that day (and separated for weekdays and holidays). What is also overlaid is the 'global' daily bus ridership, averaged over all the six months of the study period. It can be seen that, on clear days, the number of smart card trips is smaller than the global average. Combining the pattern in Figure 2, the high temperature and blistering sun may reduce trip-making. Before sunrise and after sunset when the temperature is lower, the ridership on those individual days is similar to the global average. On cloudy days, the ridership during the morning and afternoon peaks of the weekday is lower than the global average, while the ridership on the holiday day is delayed by two hours. Fog and haze have little or no impact on the bus passengers' travel behaviour in the working day, while the ridership on holiday is consistently reduced. Rain appears to have reduced bus ridership on both holiday and working days. In Section 3.2, we present a statistical analysis on the significance of the different weather events on ridership.

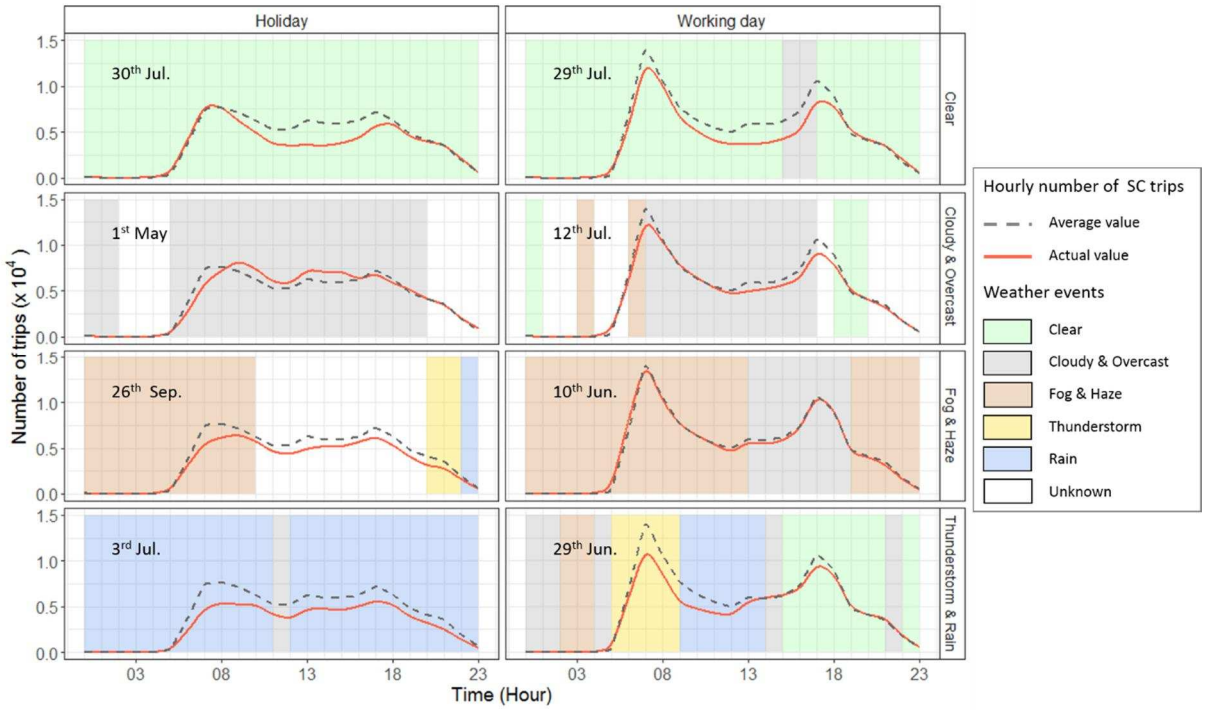

Figure 3 Hourly ridership in a week for the typical weather events.

3.2 Data pre-processing

We consider a passenger's travel from origin to destination as a journey, and each leg of their journey as a trip. To simplify the problem and clarify the data analysis process, the following assumptions are made: (A1) each passenger owns only one smart card, and each card can be only used by its owner; (A2) a journey that requires transfer among different lines is regarded as separate trips, each with its boarding and alighting stops. Although in practice, the same smart card may be used by family members or friends, by assuming (A1), we take each smart card user's travel history into consideration in estimating his/her alighting stops. With the above definition and assumptions, a trip is composed of a single pair of boarding and alighting stops.

The smart card records are firstly cleaned up based on the assumption (A1). If there are two or more records appear in the same vehicle at the same station in a very short time interval (defined as within 1 minute), the data is registered as repetitive records and counted only once (the first record) in this study. 6.3% of the data is recorded as repetitive. The remaining smart card records are combined with the GPS tracking of bus vehicles to obtain the passengers' boarding bus stops. 9.9% of trips cannot be matched with the GPS record of the vehicle number, perhaps due to poor quality of the data. Then, we capture the timestamps when a bus enters and exists a bus stop, and match the boarding time with these timestamps to find out the boarding stops. Based on this boarding-stop inference method, an additional 15.5% of trips whose

boarding stops could not be inferred. In total, 31.7% of the original smart card records are not useful, and the remaining 68.3% of the data (with 11.7 million smart card records) is applied in this study.

In our proposed machine learning approach, the training set utilises the alighting stops as the label to teach the model how to do the classification, and the testing set also requires their alighting stops to validate the accuracy and performance of the model. However, with the open automatic fare collection system, the smart card records we have do not contain any alighting information. Here, we use the trip-chaining method (Wang et al., 2011) to synthetically generate the alighting stops for the trips in training and testing datasets and assume these as the 'real' alighting stops for our proposed method. Following this naïve trip-chaining method, the trips are categorised into the following four types:

- Trips in a chain (X1);
- Segments in transfer journey excluding last one (X2);
- Last segment in the transfer journey (X3); and
- Other trips (X4).

As presented in Table 3, types X1 and X2 trips account for 26.7% of the overall records. This percentage is much lower than the cases in previous studies, for example, 75% in London (Gordon et al., 2013), 70% in Chicago (Zhao et al., 2007) and 90% in New York (Barry et al., 2002). One possible reason for the low share (of X1 and X2 trips) could be that our study network is a subset of the Changsha bus network, covering only seven bus lines. It is quite possible that there are more trips of X1 and X2 types made using other bus lines (operated by different bus companies) that are not counted in our sample. This reflects the practical constraints imposed by the bus operating framework in Changsha, as well as in many other cities, where there is more than one bus company operating different bus lines in the city, and there is no central governing body (such as Transport for London) to combine and share the smart card data generated by the different companies. A consequence of not having full access to all smart card data in a city would lead to breaks during the trips chains and lower percentage of X1 and X2 types of trips.

Table 3 The smart card data records for the study network.

Type		The number of trips	Percentage
Invalid data	Repetitive trips	1085500	6.3%
	No GPS data	1698515	9.9%
	No boarding stop	2661630	15.5%

Type		The number of trips	Percentage
Cannot infer the alighting stops	X4	6580433	38.3%
	X3	548182	3.2%
Database for this study	X2	746202	4.3%
	X1	3838614	22.4%
Total		17159076	100.0%

Earlier, we saw in Figure 3 illustrations of the different weather events and their effect on overall bus ridership. Here, we examine statistically the significance of weather events on ridership of each type of trip (chain). We use the one-way analysis of variance to examine the significant differences in the number of trips made under different weather events; the results of the statistical analysis are presented in Table 13 in Appendix A. We find that the all the p-values are less than 0.05, except those for the X3 and X4 types of trips, which proves that passengers, regardless of taking transfer trips or chaining trips, have a significantly different travel behaviour under different weather events. The results further support the hypothesis that the weather has a significant impact on trip chaining (X1 trips), and on trips that involve transfers (X2+X3, and X1+X2+X3).

Following the trip-chaining method by Wang et al. (2011), the alighting trips can be only inferred for trips of X1 and X2 types. It may be noted that some of the types X3 and X4 trips can also be used by the method proposed by He and Trépanier (2015). For consistency, however, only the trips in X1 and X2 are used in our study with the machine learning model.

4 A GBDT-based machine learning approach for alighting stop estimation

In this section, we propose a machine learning classification approach to identify the alighting stop of the trip from an open automatic fare collection system, where the alighting information of the passenger is not recorded. We incorporate each smart card user's travel history and the weather conditions into the machine learning estimation framework.

4.1 Notations

The following notations are adopted in this paper (Table 4).

Table 4 Table of notations.

Notations	Description
trip	Vector containing the features and alighting stop of a trip
r	A feature representing a characteristic of the trip
r	The vector including all the features associated with the trip
V	The number of features employed in the model
d	Alighting stop of a trip
k	Index of an alighting stop
K	The number of alighting stops in the network
merror	Estimation error
M_{wrong}	The number of trips that estimate to the wrong alighting stops
M_{total}	The total number of trips in the dataset
t	Index of iteration
T	Maximum iteration
S_a	Training set
m	The number of data in the training set
h(·)	Probability function of alighting at stops
R	Disjoint region that collectively covers all the trips
j	Index of the region
J	The number of regions
c	Coefficient corresponding to regions and defining the boundaries of regions
f(·)	Boost tree model
L(·)	Loss function
p(·)	Symmetric multiple logistic transform of the probability of alighting at stops
g	Decent direction

4.2 The machine learning estimation framework

Machine learning approach works by training the algorithm to optimise a certain performance criterion using large data samples (Alpaydin, 2014). In machine learning languages, one data record is called an 'instance'. An instance contains many observed (or model) 'features', and one 'target label' (or a class) to be estimated. The set of the observed features is called a 'feature vector'. The observed features considered in our estimations are described in Section 5.1. In our study, the target label is the alighting stop we want to estimate. The instance, which in our study represents a bus trip, is mathematically represented as:

$$\text{trip} = \left(\mathbf{r}, \text{d}\right) \tag{1}$$

where **r** denotes the vector of V observed features of the trip:

$$\mathbf{r} = \{\text{r}_1, \text{r}_2, \cdots, \text{r}_\text{V}\} \tag{2}$$

and d denotes the target label, i.e. the alighting stop of the trip:

$$d \in \left\{ d_k \middle| k = 1,2,\cdots,K \right\} \tag{3}$$

where the set of d_k represents all the possible alighting stops (or classes), and K is the total number of stops (which is 306 in our case study). Thus, our problem of predicting alighting stops is a multi-class classification problem.

The first step in machine learning is to separate all the trips in smart card data into three datasets: training, verification and testing datasets. The training set is used to obtain a trained model; verification set is used to evaluate our trained model in the training process, while the trained and verified model is applied to the testing dataset to predict the alighting stops of the trips in the dataset. The performance of a trained model is measured in terms of an estimation error (merror), defined as:

$$\text{merror} = \frac{\text{M}_\text{wrong}}{\text{M}_\text{total}} \tag{4}$$

where M_wrong is the number of trips that is estimated to the wrong alighting stop and M_total is the total number of trips in the dataset. We calculate the merrors for each of the training and verification dataset, which are used as the stopping criterion after each iteration of the training process.

Each model is trained with a set of hyper-parameters of the machine learning model. A range of initial hyper-parameter values is tested, resulting in a range of different trained models and model estimation errors. The final selected trained model is the one with the minimum estimation error, which is then applied to the testing dataset for estimating the alighting bus stops for the individual trips in that dataset.

4.3 A multi-class GDBT algorithm

4.3.1 General framework

The training algorithm introduced in this study is the gradient boosting decision tree (GBDT) algorithm. Firstly proposed by Friedman (2001), the algorithm is based on the integration of statistical and machine learning methods. More specifically, since a single tree (as used in Friedman, 2001) is too weak to lead to an accurate result, GBDT uses a set of simple trees, in

the form of a classification and regression tree, to calculate the results and draws the conclusion (i.e. to estimate the alighting stop of each trip in our model) together.

Unlike most of the applications of GBDT which have only a binary choice, the problem of alighting stop estimation belongs to the multi-class classification problem (MCCP). We outline in Figure 4 the main processes for such a multi-class GDBT algorithm. Since there are 306 possible stops in the label, we consider the alighting stops one by one. The multi-class classification problem of estimating the alighting stops is then transformed to a set of regression problems and used to build the classification and regression tree which calculates the alighting probability at each stop. The bus stop with the highest alighting probability is chosen as the final estimation result, i.e. the most probable alighting stop.

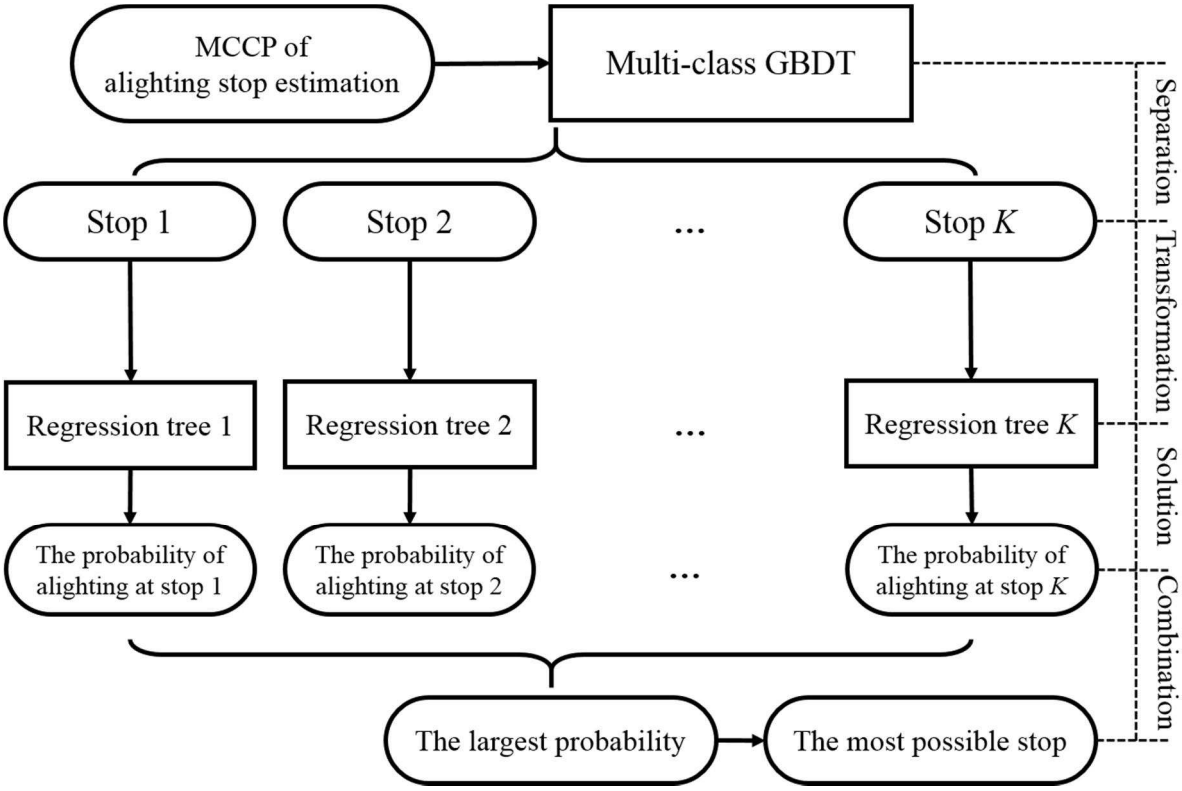

Figure 4 The processes of a multi-class GBDT.

4.3.2 Gradient boosting decision tree algorithm

The GBDT model combines a decision tree algorithm, a gradient updating algorithm, and a boosting algorithm, in an iterative process (outlined in Figure 5) to improve the training results.

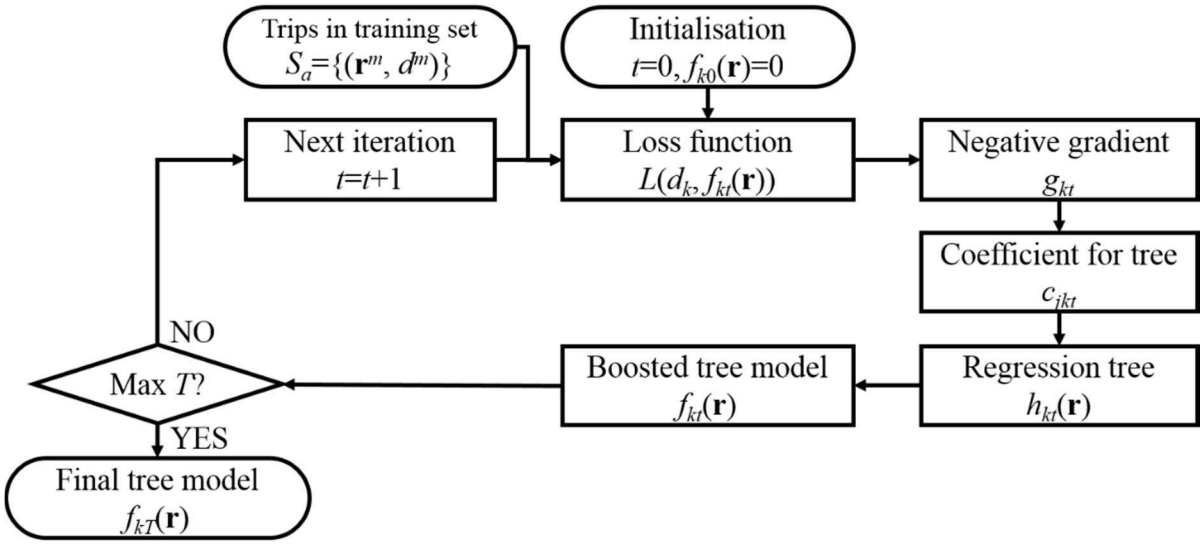

Figure 5 The detailed algorithm of GBDT model for a single stop k.

The classification and regression tree algorithm in GBDT is the most widely used decision tree model. Each internal node on the tree represents a test on a feature of the trip, while the branch represents the test output (represented in probability terms). The terminal nodes of the tree represent the alighting probability at the bus stop along the branch. Let $h_{kt}(\mathbf{r})$ denotes the estimation result for a trip $\mathbf{r}$ from a simple regression tree for stop k at the t^{th} iteration. The probability of alighting at the stop k at the t^{th} iteration is measured as the additive form:

$$h_{kt}\left(\mathbf{r}\right)=\sum_{j=1}^{J_{kt}} c_{jkt}\, I\left(\mathbf{r}\in R_{jkt}\right) \tag{5}$$

where R_{jkt} is the disjoint region j that collectively covers all the trips for stop k at iteration t, and J_{kt} is the number of regions for stop k at iteration t. These regions are represented by the terminal nodes of the tree. c_{jkt} is a coefficient corresponding to region j for stop k at t iteration t, which defines the boundaries of the regions. The indicator function $I(\cdot)$ has the value 1 if the argument is true, and zero otherwise.

The idea of boosting is to identify ways to improve the simple trees. Let $f_t(\mathbf{r})$ denotes the estimation result of the boosted tree model after iteration t. Hence, the boosted tree model, or $f_T(\mathbf{r})$, can be obtained from:

$$f_{kT}\left(\mathbf{r}\right)=\sum_{t=1}^{T} h_{kt}\left(\mathbf{r}\right) \tag{6}$$

To increase the accuracy of the estimates, a loss function is being minimised step by step in the iterative GBDT process. A gradient algorithm is used to calculate the direction where the loss function decreases the most, and the gradient of numerical decent. The negative

direction of the gradient refers to the direction where the loss function decreases the most. In GBDT, the loss function employs the log-likelihood loss function (Friedman, 2002):

$$L\left(d_k, f_k\left(\mathbf{r}\right)\right) = -\sum_{k=1}^{K} d_k \log_{10} p_k\left(\mathbf{r}\right) \tag{7}$$

where d denotes the real alighting stop; $f_k(\mathbf{r})$, calculated from Equation (6), is the probability of the trip estimated to alight at stop k; and d_k is the probability of the trip belonging to alighting stop k, where d_k equals to 1 if k is the real alighting stop, otherwise, 0.

Following the method of Friedman et al. (2000), we use the symmetric multiple logistic transform:

$$p_k\left(\mathbf{r}\right) = \frac{\exp\left(f_k\left(\mathbf{r}\right)\right)}{\sum_{l=1}^{K} \exp\left(f_l\left(\mathbf{r}\right)\right)} \tag{8}$$

Then the decent direction of trip m at stop k and iteration t can be calculated as:

$$g_{kt}^{m} = -\left[\frac{\partial L\left(d_l^m, f_{kt}\left(\mathbf{r}^m\right)\right)}{\partial f_{kt}\left(\mathbf{r}^m\right)}\right]_{f_{l,t}(\mathbf{r}) = f_{l,t-1}(\mathbf{r})} = d_k^m - p_{k,t-1}\left(\mathbf{r}^m\right), \qquad l = 1, 2, \cdots, K \tag{9}$$

Equation (9) states that the error is the difference between the real probability of the alighting stop k that trip m maps and the corresponding estimated probability at iteration t-1.

Next, a new tree can be generated by following Equation (5), where the coefficient can be optimised as:

$$c_{jkt} = \arg\min_{\underset{c_{jk}}{\smile}} \sum_{m=1}^{M} \sum_{k=1}^{K} L\left(d_k^m, f_{k,t-1}\left(\mathbf{r}^m\right) + \sum_{j=1}^{J} c_{jk} I\left(\mathbf{r}^m \in R_{jt}\right)\right) \tag{10}$$

Following Friedman et al. (2000), Equation (10) is approximated as:

$$c_{jkt} = \frac{K-1}{K} \frac{\sum_{\mathbf{r}^m \in R_{jkt}} g_{kt}^{m}}{\sum_{\mathbf{r}^m \in R_{jkt}} \left|g_{kt}^{m}\right|\left(1 - \left|g_{kt}^{m}\right|\right)} \tag{11}$$

With Equation (11), a new regression tree for each stop can be generated by Equation (5), and the boosted tree model can be updated by using Equation (6):

$$f_{kt}\left(\mathbf{r}\right) = f_{k,t-1}\left(\mathbf{r}\right) + \sum_{j=1}^{J_{kt}} c_{jkt} I\left(\mathbf{r} \in R_{jkt}\right) \tag{12}$$

The iterative process continues until an empirical stopping criterion is met by comparing the merrors of training and verification dataset from Equation (4). In our case, a pre-specified number of iterations is reached.

Unlike many other machine learning methods, GBDT is able to evaluate the relative importance of the independent features of the trip. Since the depth of the tree is constrained by the hyper-parameter, the simple tree that is used in each iteration only includes a randomly chosen set of features (as opposed to all features). Hence, the frequency of the features used across all trees can be used to measure the relative importance.

4.4 Model evaluation

One measure of the performance of the model is the estimation error merror from Equation (4). However, it is a far too simple measurement to evaluate the machine learning models. Generally, a confusion matrix of measures, composed of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), is used to evaluate the performance of the binary classification model (Stehman, 1997). For our multi-class classification model, we introduce a confusion matrix for the estimated results of each alighting stop and then calculate evaluation indexes (Powers, 2011; Zhou, 2016). Table 5 presents the confusion matrix for a single alighting stop.

Table 5 The confusion matrix for the estimated results of a single alighting stop k

Real alighting stop / **Estimated alighting stop**	**k (positive)**	**Other stops except for k (negative)**
k (positive)	TP_k	FP_k
Other stops except for k (negative)	FN_k	TN_k

Our evaluation indexes then include precision (macro P), recall (macro R) and F1 score (macro F1), as defined in equations (13 – 15). Precision and recall reflect the quality of the model in terms of the reliability for the results and the applicability for the sample. The F1 score, the harmonic mean of the precision and recall, measures and provides an overall performance of the model. The higher F1 scores indicate more superior models.

$$\mathrm{macroP} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k} \tag{13}$$

$$\mathrm{macroR} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FN_k} \tag{14}$$

$$\mathrm{macroF1} = 2 \cdot \frac{\mathrm{macroR} \cdot \mathrm{macroP}}{\mathrm{macroR} + \mathrm{macroP}} \tag{15}$$

5 Feature selection and experiment designs

In this section, we introduce the features selected that characterise the trips, and the machine learning experiments designed to evaluate the relative performances of the algorithms and data features.

5.1 Feature selection

Each trip in this study contains 18 observed features, denoted as r_1 to r_{18} in Equation (2), and the one target label, d; these are listed in Table 6. The observed features contain three groups of data: (i) the basic bus trip information as recorded by the smart card and the boarding information as inferred from the GPS records of the bus services; (ii) the smart-card user's recent travel history, also extracted from the historical smart card data; and (iii) the ambient weather data for the trips taken in (i). The temporary features and boarding stops are the necessary information from the smart card data and often used in the previous studies for the bus ridership estimation. To investigate the regularity of travels and describe the travelling preference, we introduce features about the passengers' recent travel history. As noted in the Introduction, weather can impact on the travel destinations (Sabir, 2011), we introduce independent weather variables in the estimation of alighting stops.

Table 6 The selected model features and the target label.

Feature groups	Features	Types	Investigated range
Basic smart card information	Month	Discrete	4 - 9 [for April to September]
	Day	Discrete	1, 2, …, 31 [day]
	Hour	Discrete	0, 1, 2, …, 23 [hrs]
	Days of week	Categorised	Mon., Tues., Wed., Thurs., Fri., Sat., Sun.
	Holiday	Binary	0: working day; 1: holiday
	Boarding stop ID	Nominal	060101, 060102, etc.
	Boarding line	Nominal	6, 7, 63, 123, 147, 150, 168
Travel history	Number of trips on the previous day	Discrete	0, 1, 2, …
	Number of trips in the same hour on the previous day	Discrete	0, 1, 2, …
	Number of trips on all the previous 7 days	Discrete	0, 1, 2, …
	Number of trips on the same day of the last week	Discrete	0, 1, 2, …
	Number of trips in the same hour on the same day of last week	Discrete	0, 1, 2, …

Feature groups	Features	Types	Investigated range
Weather conditions	Temperature	Continuous	-6 - 40 [°C]
	Precipitation	Continuous	0 - 58mm
	Humidity	Continuous	0 - 100 [%]
	Visibility	Continuous	0 - 10 [km]
	Wind speed	Continuous	0 - 10 [mph]
	Weather events	Categorised	Clear, rainy, misty, cloudy, overcast, unknown.
Model label	Alighting stop IDs	Nominal	060101, 060102, etc.

5.2 Experimental design

As GBDT is a relatively new machine learning algorithm, we adopt two other classic algorithms, multinomial logistic regression (MLR) and neural network (NN), to compare their relative performances. MLR and NN are the two most popular algorithms used in machine learning approaches. Both have been used in a variety of transport applications, e.g. traffic forecasting, travel mode choice modelling and trip distribution modelling (Karlaftis and Vlahogianni, 2011). The hyper-parameters of the GBDT and NN algorithms are set as inputs for the machine learning model. Table 7 displays the different initial settings of the hyper-parameters for the algorithms during the training process.

Table 7 The initial setting of the hyper-parameters during the training process.

Hyper-parameters	GBDT	NN
Learning rate or step-size	0.0005,0.001, 0.005, 0.01, 0.05, 0.1	0.0005,0.001, 0.005, 0.01, 0.05, 0.1
Maximum depth of each tree	3,5,8,10,12,15	-
Fraction of data for training next tree	0.2, 0.4, 0.6, 0.8	-

Note: The next two groups of experiments with GBDT follow this initial setting.

We conduct six experiments, with increasing total number of trips in the training and verification set, while keeping the same dataset as the testing set. Table 8 lists the details of the training and verification datasets used for the six experiments. All the six experiments employ the same testing dataset, the trips made on 30th September 2016. The other data are combined and used as training and verification data (depending on the sample sizes). In each of the six samples, 30% of the combined training and verification data is chosen randomly as the verification data and the rest 70% as the training data. From Samples 1 to 6, the number of days and data in the combined training and verification data increases. These experiments are

designed to illustrate the relationship between the size of the training set and the accuracy of the results.

Table 8 Data sample for the experiments.

Experiments	The training and verification data		The testing data	
	Days	Number of records	Day	Number of records
Sample 1	01-29/09/2016 (23 days)	807136	30/09/2016 (1 day)	22602
Sample 2	01/08 - 29/09/2016 (54 days)	1576132		
Sample 3	01/07 - 29/09/2016 (85 days)	2324989		
Sample 4	08/06 - 29/09/2016 (108 days)	2954072		
Sample 5	08/05 - 29/09/2016 (123 days)	3388117		
Sample 6	08/04 - 29/09/2016 (143 days)	3983788		

Note: Sample 4 – 6 exclude days when the data was missing.

As introduced in Section 1, in this study, we are interested in the weather conditions and travel history of passengers' behaviour. To test the hypothesis, we set up several GBDT models with four different combinations of feature groups (dubbed as FG 1 to 4). The feature groups and the setting of hyper-parameters are displayed in Table 9. All four experiments use the dataset of Sample 6 experiment. FG 1 uses only the basic smart card information, similar to many previous methods. FG 2 adds only the travel history to FG 1, and FG 3 adds only the weather variables to FG1, while FG 4 includes the full set of features proposed in Table 6 (i.e. adds both travel history and weather conditions to FG 1). The experiments are designed to help us understand the effect of different groups of features.

Table 9 Experimental designs with different feature groups (FG).

Experiments	Basic smart card information	Travel history	Weather conditions
FG 1	✓		
FG 2	✓	✓	
FG 3	✓		✓
FG 4	✓	✓	✓

6 Model results

6.1 Model comparison

Sample 6 and feature group FG 4 are applied to the GBDT model introduced in this paper, and to the MLR and NN models. For the GBDT and NN model, the final set of the hyper-parameters are displayed in Table 10.

Table 10 Values of the hyper-parameters in GBDT and NN.

Hyper-parameters	GBDT	NN
Learning rate or step-size	0.005	0.001
Iteration	150	30
Maximum depth of each tree	8	-
The fraction of data for training next tree	0.4	-
Number of nodes in layers	-	(342,333,333,306)
Activation function for hidden layers	-	Sigmoid
Dropout rate	-	0.3
Activation function for output layer	-	SoftMax

The relative estimation power of the three models, as measured by their precision (macro P), recall (macro R) and F1 score (macro F1), are illustrated in Figure 6. The F1 score of a random classification is used as the baseline for comparison, as is indicated as the 'baseline' in Figure 6. It can be seen that the F1 scores of all these three algorithms are higher than that of the baseline, suggesting all three models are theoretically acceptable, while GBDT has the best performance according to macro P and macro F1. Looking at the precision and recall of the model estimations, we can see that the values of macro P are always higher than macro R for their respective machine learning algorithm. This suggests that the estimation accuracy in all three models is better than their recall power. GBDT has the highest prevision accuracy, while NN has slightly higher recall power than GBDT. Overall, GBDT performs the best, and its prediction power is higher in accuracy than in its comprehensiveness.

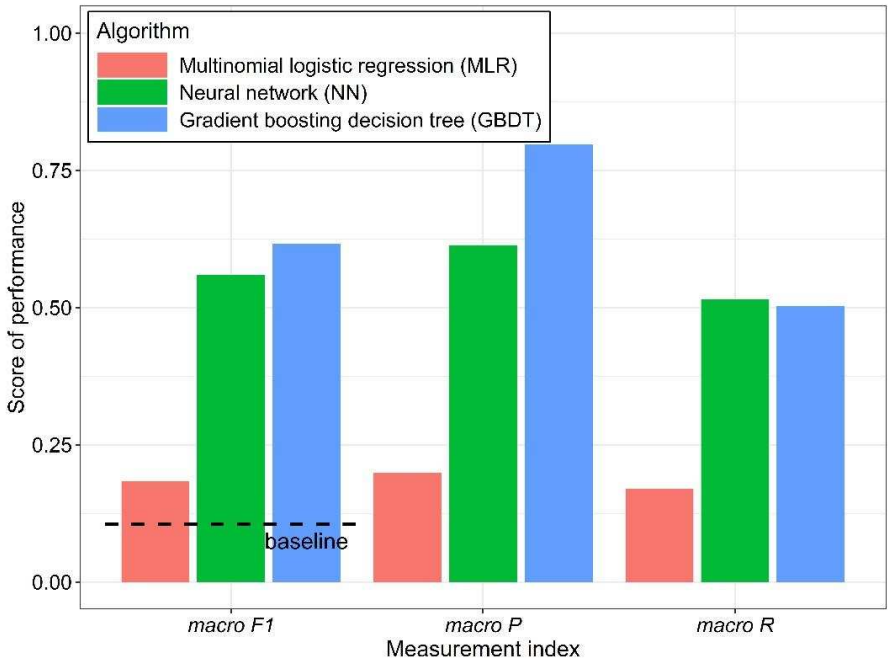

Figure 6 Comparison of the performance of the three training algorithms.

6.2 Effect of the training data size

The different training datasets, as defined in Table 8, are applied to the GBDT model with feature group FG 4. The values of hyper-parameters are displayed in Table 11.

Table 11 Values of the hyper-parameters in Sample 1 to Sample 6.

Setting	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Learning rate	0.001	0.001	0.001	0.005	0.001	0.005
Iteration	70	70	100	120	110	150
Maximum depth of each tree	6	8	8	5	7	8
Fraction of data for training next tree	0.5	0.7	0.6	0.7	0.5	0.4

The prediction measures are shown in Figure 7. We can see that, in general, increasing the training data size improves the prediction power. The most significant improvement happens between experiment Sample 1 and Sample 2, while the improvements gradually become smaller as the total sample sizes get larger. The level of precision is universally higher than the recall. It indicates that this model does better in the precision estimation than in the extensive estimation.

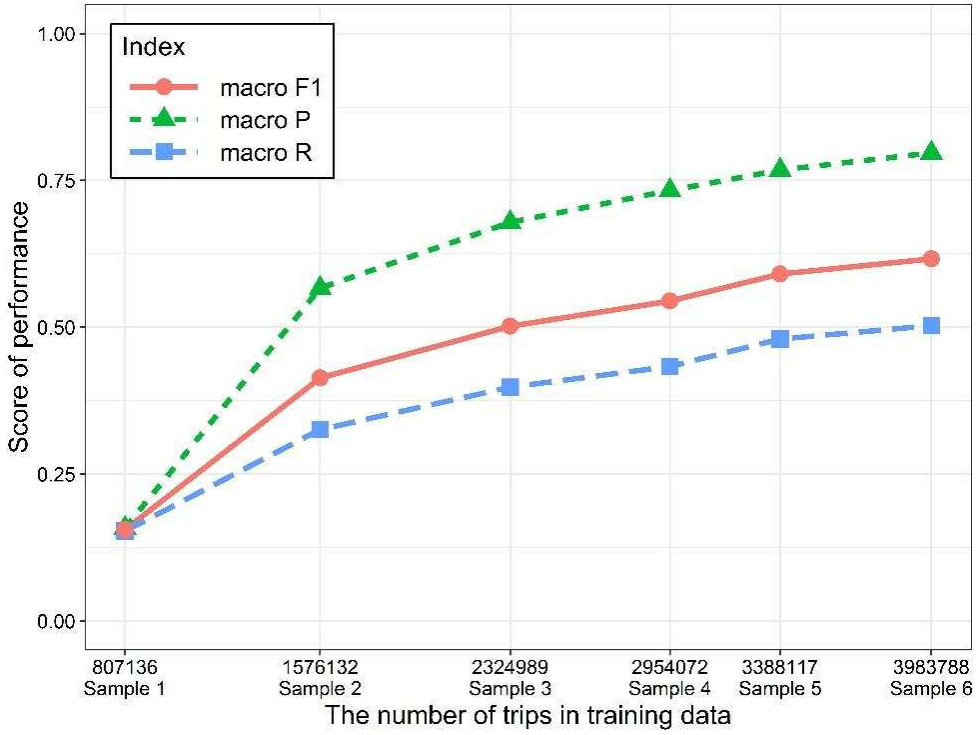

Figure 7 The performance measurements of the model in different size of the training set.

6.3 Impact of weather condition and travel history

We illustrate the impacts of including the weather variables and historical trips on our models by comparing the results of the four groups defined in Table 9. The final value of the hyper-parameters of each group is presented in Table 12.

Table 12 Values of the hyper-parameters for the experiments with different feature groups.

Experiments	**FG 1**	**FG 2**	**FG 3**	**FG 4**
	Basic smart card information	FG 1 + travel history	FG1 + weather conditions	FG1 + travel history + weather conditions
Learning rate	0.05	0.05	0.005	0.005
Iteration	80	80	120	150
Maximum depth of each tree	3	5	5	8
Fraction of data for training next tree	0.6	0.5	0.6	0.4

In Figure 8, from FG 1 to 4, as reflected in the F1 scores, we see improvements in the performances of the models. Between FG 2 and 3, the improvement from adding historical trips is less compared to the improvement resulted from adding the weather conditions. We speculate that this is because the information about historical trips captures the regularity of the behaviour, but the travel behaviour of the passengers are affected more by the changing weather conditions. FG 4 has the best F1 scores indicating that including both travel history and weather leads to the best performance.

Looking at the precision and recall sides of each model, the main increase from FG 1 to 2 is in the recall index (macro R). The similar situation occurs from FG 3 to 4. When we remove the historical trips out of the model (from FG 2 to 3), the recall ability of the model reduces slightly. However, there is a significant increase in precision from FG 2 to 3. The results suggest that the two groups of features can improve the model in different ways. Principally, the features about the historical trips improve the comprehensiveness of the model and the weather variable makes the estimation more accurate.

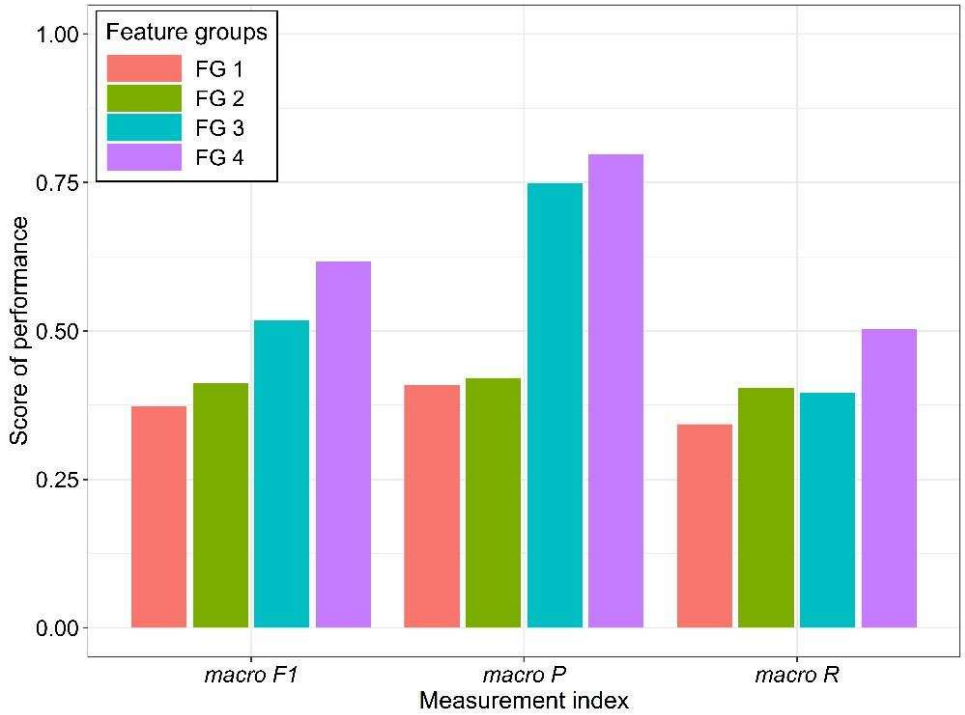

Figure 8 Evaluation of the impacts of different feature groups.

6.4 Relative importance of feature variables

We capture the relative importance of the features in models FG 1 to 4 in Figure 9. It can be seen in Figure 9(a) that, for FG 1(when only considering smart card data), the boarding stops is the most significant feature. The month and day are not significant features. The feature of holiday is the least significant feature, for which the day in a week contains the information on holiday to some extent. For FG 2 (as seen in Figure 9(b)), almost all the features about historical trips score low on impact. In Figure 9(c) for FG 3, the boarding stop scores the highest followed by the weather events with the temperature being the most important of the group. However, the importance of other quantitative weather features, i.e. wind speed, humidity, visibility and precipitation, is not as significant. This may imply that where passengers decide to alight is not influenced by their qualitative cognition (e.g. whether it is rainy or not) than by the quantitative information (e.g. how much the precipitation is). This is also reflected in practice how bus companies adjust their service frequency under different weather event, i.e. they provide more frequent bus services on a rainy day regardless of the level of precipitation. Even if the bus company cannot respond to individual weather variables, e.g. humidity and visibility, our study suggests that the simple register of a 'weather event' would improve the origin-destination demand estimation and better (re)scheduling of their bus services. In Figure 9(d) for FG 4, we can see that the weather condition is a much more important group of features than the travel history in the model. This reinforces the findings drawn in Section 6.3. Thus

based on the importance analysis and performance measurement, FG 4 (the model including all the features) is demonstrated as the best model.

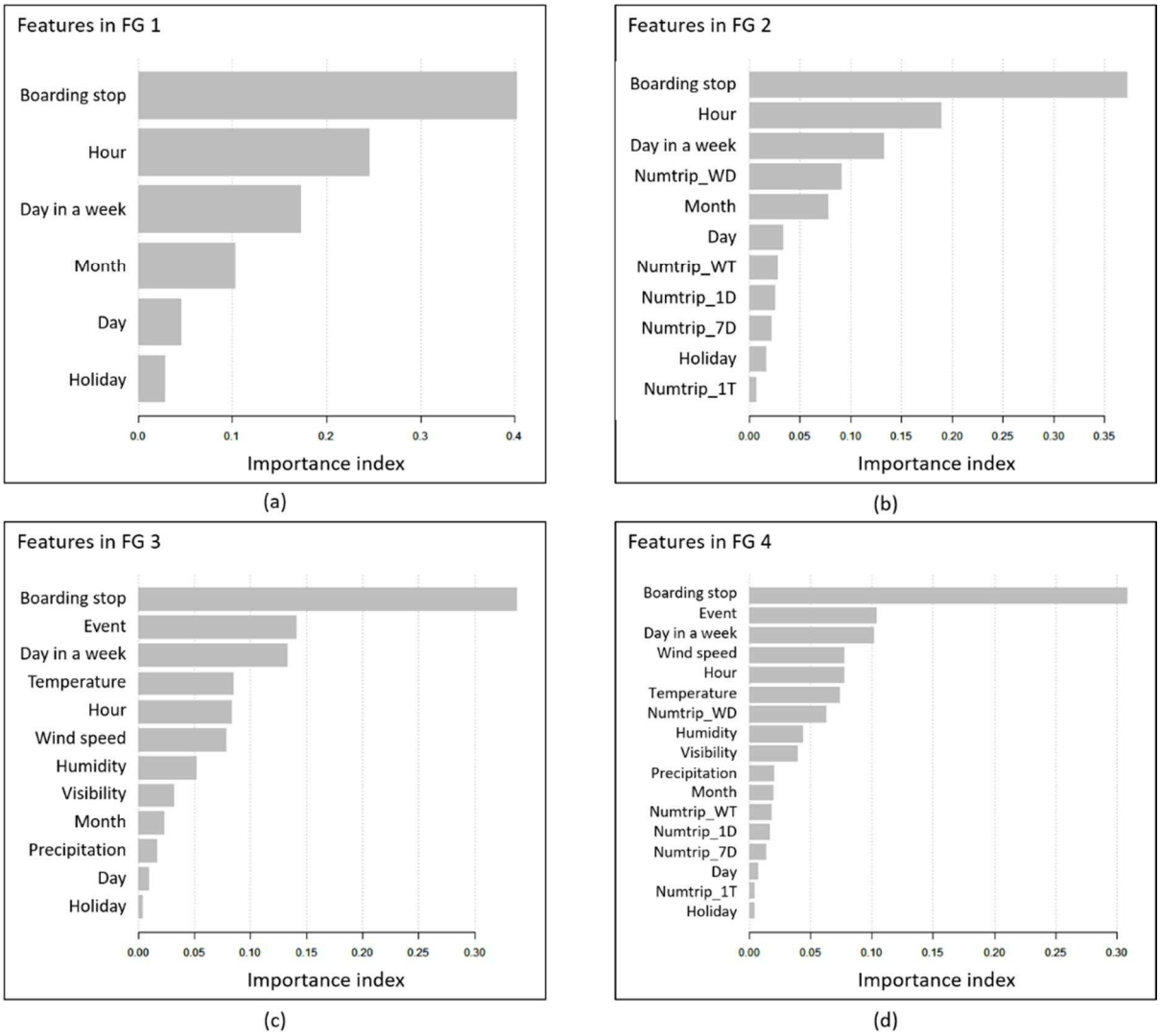

Figure 9 Ranking of the feature importance in different feature group experiments.

6.5 Ridership estimation

For bus planning, the overall demand (and distributions) of bus ridership is the most critical factor to consider. In this section, we apply our two trained models with and without weather features (FG 4 versus FG 2) to the test dataset, to predict the following aggregated bus ridership:

- The number of alighting passengers at each station,
- The load-profile and max load on each line.

The models with feature groups FG2 and FG4 are used in the prediction, and the predicted alighting stops are compared with the 'true' alighting stops as inferred from the trip-chaining model (Wang et al., 2011). We utilise the GEH Statistic (DfT, 1996) to compare the

difference between the estimated and true alighting numbers at bus stops, which is formulated as:

$$GEH = \pm\sqrt{\frac{2\left(N_e - N_r\right)^2}{N_e + N_r}} \tag{16}$$

where N_e and N_r represent the number of correctly estimated and true alighting stops, respectively. Additionally, we add the signs to represent when $N_e > N_r$ (positive) and when $N_e < N_r$ (negative). Figure 10 presents the GEH statistics for seven bus lines (in both directions). In general, an absolute value of GEH less than 5 is considered acceptable (DfT, 1996).

Overall, 98% of the alighting stops have a GEH value less than 5, suggesting that the estimation accuracy is high. There are six stops for FG 2 and seven stops for FG 4 with GEH value between 5 and 10, while four stops in both FG 2 and FG 4 have GEH values greater than 10. We find that most of those stops (with GEH higher than 5) are the last stop of the bus route. It is possible that the accumulation of the errors at intermediate stops leads to those large errors at the last stops. Besides, in the middle of Line 150 outbound services, there are two stops with high GEHs. We take Stop 16 in the model with FG 4 as an example: the true number of alighting is 7, while the weather-included model (FG 4) estimates that the number of alighting at this stop is 36. So, a small base number might cause a higher GEH value. Another possibility is that the trips alighting at this stop only make up 1.8% of the total training data. This load imbalance might also cause inaccuracy in the estimation. It is worth noting that the GEHs of Line 63, 123, 147 and 168 are near 0, suggesting that both of the models with FG 2 and FG 4 produce accurate matches in the alighting numbers at bus stops along these four bus lines. Furthermore, comparing the two models, 60% of stops have lower GEH in FG 4 than in FG 2, suggesting that including weather conditions help our machine learning model estimate more accurate alighting stops.

Figure 11 shows the ground truth and estimated ridership with FG 4 and FG 2, the two trained machine learning models with and without weather variables. As seen in the figure, the estimated load-profile has similar profiles as the ground truth and correctly matches the max-load stop in the ground truth with little differences in the absolute value of the maximum load. This is especially the case for Line 63, 123, 147 and 168, which have fewer passengers and which get an almost perfect matching. Although both models reflect the ground truth reasonably well, the model containing the weather variables (FG 4) is closer to the ground truth. Again, this comparison confirms that including weather variables makes the ridership estimation more accurate.

Additionally, the significant errors occur at the downstream stops of each bus services, which can be attributed to accumulation effects. We take a closer look at the outbound services of Line 6 and 147 to gain better insights about the accumulation errors. Line 6 has the largest error, while there are few errors in the latter one. Having a look at these two services in Figure 10, most stops in Line 6 has negative errors, and the only two positive errors are in the middle and at the end, respectively. With accumulating the negative error, the increasing number of passengers are counted on board. So, the estimated ridership increases by these errors. The difference becomes larger and larger until the final positive error corrects the previous accumulated negative errors. However, the situation in Line 147 is different. The negative and positive errors occur alternately so that the following error can correct the previous errors in time, and the errors are not accumulated.

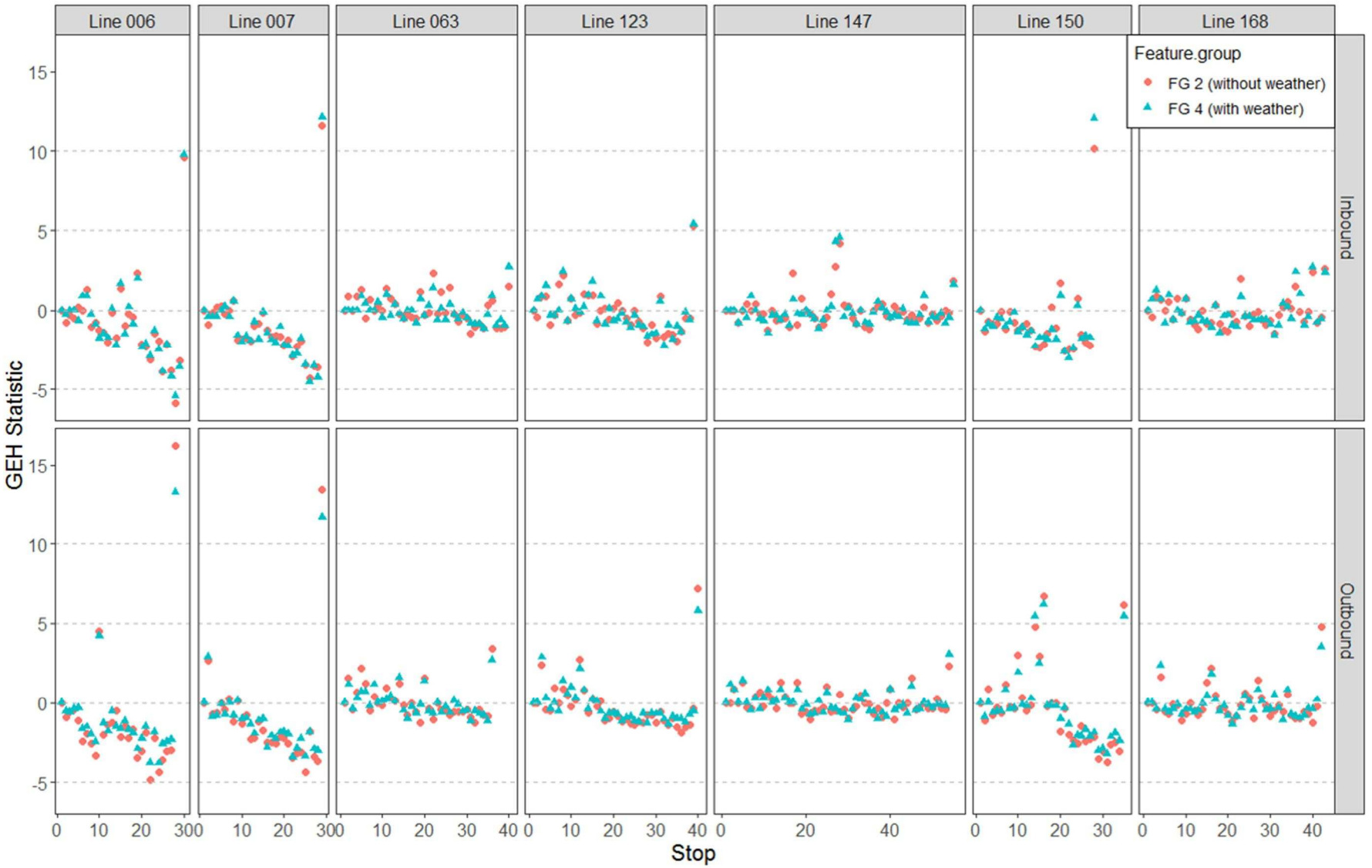

Figure 10 The GEH statistic of the alighting number at each station.

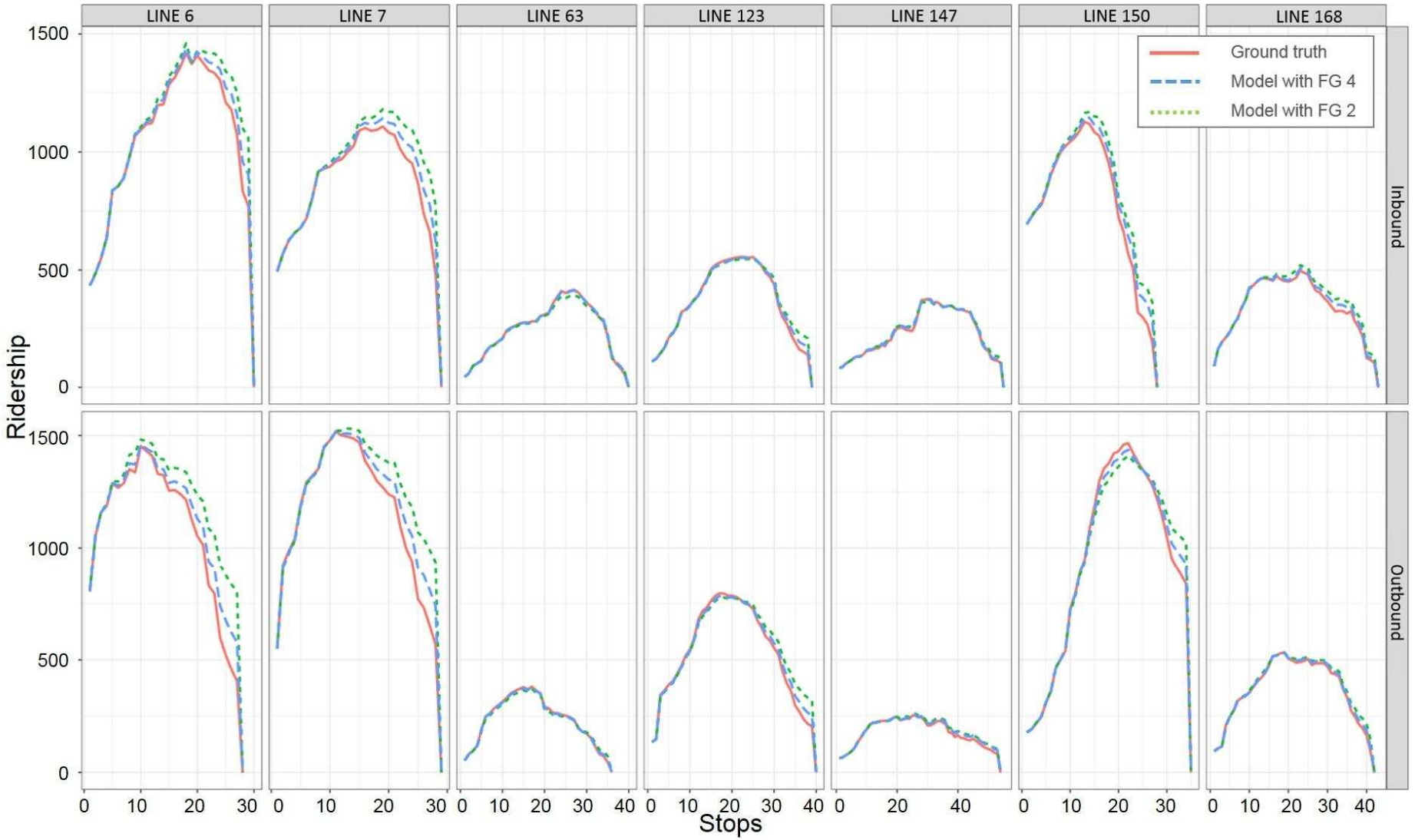

Figure 11 The load-profile of each service in ground truth, weather-included model and weather-excluded model.

7 Discussion and Conclusion

Developing smart public transport system is a vital task in building sustainable cities. Understanding passengers' origin-destination and travel pattern are of great importance to improve the level of services and attractiveness of buses.

This study proposes a machine learning model with advanced gradient-boosting decision tree (GBDT) solution algorithm to estimate the alighting stops from the smart card logs of an open automatic fare collection system. We explicitly incorporate features that represent weather conditions and information of the individual's travel history in the model, so the estimation is not only based on the characteristics of the trip itself but also referring to the impacts of the ambient environment and the passengers' habitual travel behaviour.

To illustrate the performance of our proposed method, we conduct three comparative studies: (i) GBDT method vs two commonly used machine learning models; (ii) the size of the training dataset; and (iii) the inclusion (or not) of weather conditions and travel history in the estimation model. The results show that the machine learning method can accurately estimate the alighting stops from smart card data and that GBDT performs better than NN and MLR overall, and in particular from the view of precision. Intuitively, increasing the size of training dataset improves the estimation accuracy. However, we discover that there is less improvement after a certain point (in our case, 3-month data in training dataset). The results also confirm that we can obtain a more accurate estimation when considering more features in the model, although the effect of the features varies. Weather conditions improve the accuracy (in precision), and historical trips improve the comprehensiveness (in recall). Additionally, the high ranking of the feature, weather events, and its significant contribution in increasing the precision of the GBDT model highlights that the effect of this variable is worthy of detailed testing when analysing and predicting passengers' decision of alighting stops.

Whilst the model trained in this paper with smart card data from Changsha is only applicable to this specific study network, the proposed GDBT framework is generic and can be applied to other smart card systems (open or closed): firstly using the smart card data to obtain a trained model, and then apply the trained model to predict bus ridership in the near future where travel conditions (such as weather conditions) can be readily predicted. Even the application in different cities can customise their model by easily adding or deleting the features in the model.

This study can also be used to predict the alighting stops in short-term as opposed to long-term trend prediction. The short-term prediction emphasises on the detailed value and

minor changes (dynamics) and leverages the availability of accurate short-term weather forecasts. The target application is to make minor adjustments in the schedule of high-frequency bus services.

Overall, this paper makes new advances in these main aspects. Firstly, we employ a machine learning model with that GBDT algorithm in bus data mining, a novel technique in processing the massive smart card data. Secondly, our method is general and applicable to individual bus trips made by regular and irregular passengers as recorded in smart card data and fills the gap of the trip-chaining model, which requires the identification of an unbroken trip chain for every smart card user. Third, we incorporate the impacts of weather conditions and travel history in the estimation of detailed origin-destination and ridership. Our model estimates the alighting stops for each smart card log, making it possible to readily compute the origin-destination-based load-profile for each bus line, important baseline information for planning more attractive bus services for the public.

We conclude this paper by critically examining the limitations of the current study. By its very nature, the true alighting stops of the open smart card system are not known, and we did not have access to an alternate source of ground truth data. Rather, we only have access to smartcard data from one bus company and use the naïve trip-chaining method to generate/obtain an estimated 'ground-truth'. Lack of validation from the real data is a limitation of this study. Validation against true alighting stops collected from surveys and/or inferred from other data sources (e.g. video recordings) will help verify the assumption and support the machine learning model estimation process. Additionally, this paper, in fact, applies the proposed machine learning model on the trip chains (X1) and part of transfer trips (X2) and we are not able to consider travellers who change their travel modes under different weather conditions. This could potentially lead to survivor bias for the model, as a result of the limited availability of data sources. As more multi-source data, such as bus video recordings and automatic passenger counted data, becomes readily available in the future, they open opportunities for data fusion and new models for estimating bus ridership.

Further, due to the limitation of data accessibility, our case study is a subset bus network in Changsha, which may have caused the underlying self-selection bias. As stated in Section 3.2, it is common for more than one bus companies to operate in a city and they do not share their data. In our study, we try to reduce the bias by selecting representative bus lines in the city (discussed in Section 3.1). However, this still leaves a certain level of self-selection bias, which we cannot completely avoid. It is worth further investigation to compare the deviation of the models trained by the data from the whole and subset network.

Our model has 306 stops in the study sub-network of Changsha. In larger networks, the number of bus stops can be very big. A large number of stops can cause difficulty in training a good model; this is a general challenge for multi-class machine learning problem. As far as we are aware that machine learning methods have only been applied to cases with limited stops (classes), for example, Jung and Sohn (2017) consider only five candidate alighting stops in their model. One possible advance might be to separate the trips by bus lines and to build machine learning models line by line. This reduces the number of candidate stops (classes) in each model. However, the increased number of models are likely to make it more difficult for bus companies to use them, and to coordinate the different changes in service schedule at a network level.

The current study is only the first step in applying machine learning techniques to estimate bus ridership form open smart card data. We believe that there will be more efforts and gain can be made from using machine learning techniques to gather passenger origin-destination demand and ridership information to support developing a sustainable public transport system.

Acknowledgements

We thank the Hunan Longxiang Bus Ltd Co for making the smart card data available to this study. We also acknowledged the support from the UK Department for Transport (Project "Future Streets") and the National Natural Science Foundation of China (71890972/71890970).

Appendix A

Table 13 Results of the one-way analysis of variance of trip types vs feature 'weather event'.

Total number of trips in categories	Trip types	Homogeneity of test	p-value
Trips in a chain	X1	0.737	0.023
Segments in transfer journey excluding last one	X2	0.338	0.004
Last segment in the transfer journey	X3	0.403	0.112
Other trips	X4	0.483	0.068
Transfer trips	X2+X3	0.559	0.020
Trips chains and transfer trips	X1+X2+X3	0.889	0.018
All the trips	X1+X2+X3+X4	0.588	0.032
Trips used in the paper	X1+X2	0.777	0.017

References

Aaheim, H.A., Hauge, K.E. (2005) Impacts of climate change on travel habits: a national assessment based on individual choices. CICERO report.

Alpaydin, E. (2014) Introduction to Machine Learning. MIT Press.

Arana, P., Cabezudo, S., Peñalba, M. (2014) Influence of weather conditions on transit ridership: a statistical study using data from Smartcards. Transportation Research Part A: Policy and Practice 59, pp. 1-12.

Bagchi, M., White, P.R. (2005) The potential of public transport smart card data. Transport Policy 12, pp. 464-474.

Barry, J., Freimer, R., Slavin, H. (2009) Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research Record: Journal of the Transportation Research Board 2112, pp. 53-61.

Barry, J., Newhouser, R., Rahbee, A., Sayeda, S. (2002) Origin and destination estimation in New York City with automated fare system data. Transportation Research Record: Journal of the Transportation Research Board 1817, pp. 183-187.

Berrebi, S.J., Watkins, K.E., Laval, J.A. (2015) A real-time bus dispatching policy to minimize passenger wait on a high frequency route. Transportation Research Part B: Methodological 81, pp. 377-389.

Böcker, L., Dijst, M., Prillwitz, J. (2013) Impact of everyday weather on individual daily travel behaviours in perspective: a literature review. Transport Reviews 33, pp. 71-91.

Bordagaray, M., dell'Olio, L., Ibeas, A., Cecín, P. (2013) Modelling user perception of bus transit quality considering user and service heterogeneity. Transportmetrica A: Transport Science 10, pp. 705-721.

Ceder, A. (2007) Public Transit Planning and Operation: Theory, Modeling and Practice. Butterworth-Heinemann, Oxford, UK.

Corman, F., Kecman, P. (2018) Stochastic prediction of train delays in real-time using Bayesian networks. Transportation Research Part C: Emerging Technologies 95, pp. 599-615.

DfT (1996) Traffic Appraisal in Urban Areas. Department for Transportation, p. 70-71.

Dou, H., Liu, H., Yang, X. (2007) OD matrix estimation method of public transportation flow based on passenger boarding and alighting. Computer and Communications 2, pp. 79-82.

Friedman, J., Hastie, T., Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics 28, pp. 337-407.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics, pp. 1189-1232.

Friedman, J.H. (2002) Stochastic gradient boosting. Computational Statistics & Data Analysis 38, pp. 367-378.

Gordon, J., Koutsopoulos, H., Wilson, N., Attanucci, J. (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transportation Research Record: Journal of the Transportation Research Board 2343, pp. 17-24.

Guo, Z., Wilson, N., Rahbee, A. (2007) Impact of weather on transit ridership in Chicago, Illinois. Transportation Research Record: Journal of the Transportation Research Board 2034, pp. 3-10.

He, L., Trépanier, M. (2015) Estimating the destination of unlinked trips in transit smart card fare data. Transportation Research Record 2535, pp. 97-104.

Hofmann, M., O'Mahony, M. (2005) The impact of adverse weather conditions on urban bus performance measures. Proceedings of Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE, pp. 84-89.

Hollander, Y., Liu, R. (2008) Estimation of the distribution of travel times by repeated simulation. Transportation Research Part C: Emerging Technologies 16, pp. 212-231.

Hou, Y., He, M., Zhang, S. (2012) Origin-destination matrix estimation method based on bus smart card records. Journal of Transport Information and Safety 30, pp. 109-114.

Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C. (2015) Planning, operation, and control of bus transport systems: A literature review. Transportation Research Part B: Methodological 77, pp. 38-75.

Johnson, A. (2003) Bus transit and land use: illuminating the interaction. Journal of Public Transportation 6, pp. 21-39.

Jung, J., Sohn, K. (2017) Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. IET Intelligent Transport Systems 11, pp. 334-339.

Karlaftis, M.G., Vlahogianni, E.I. (2011) Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies 19, pp. 387-399.

Kashfi, S.A., Bunker, J.M., Yigitcanlar, T. (2015) Understanding the effects of complex seasonality on suburban daily transit ridership. Journal of Transport Geography 46, pp. 67-80.

Kwan, S.C., Hashim, J.H. (2016) A review on co-benefits of mass public transportation in climate change mitigation. Sustainable Cities and Society 22, pp. 11-18.

Li, T., Sun, D., Jing, P., Yang, K. (2018) Smart card data mining of public transport destination: a literature review. Information 9, p. 18.

Li, Y., Wang, X., Sun, S., Ma, x., Lu, G. (2017) Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. Transportation Research Part C: Emerging Technologies 77, pp. 306-328.

Liu, C., Susilo, Y.O., Karlström, A. (2015) Investigating the impacts of weather variability on individual's daily activity–travel patterns: A comparison between commuters and non-commuters in Sweden. Transportation Research Part A: Policy and Practice 82, pp. 47-64.

Liu, R., Sinha, S. (2007) Modelling urban bus service and passenger reliability. International Symposium on Transportation Network Reliability, Hague.

Liu, Y., Liu, Z., Jia, R. (2019) DeepPF: A deep learning based architecture for metro passenger flow prediction. Transportation Research Part C: Emerging Technologies 101, pp. 18-34.

Ma, F., Ren, F., Yuen, K.F., Guo, Y., Zhao, C., Guo, D. (2019) The spatial coupling effect between urban public transport and commercial complexes: A network centrality perspective. Sustainable Cities and Society 50, p. 101645.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D. (2014) Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies 44, pp. 70-79.

Munizaga, M.A., Palma, C. (2012) Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies 24, pp. 9-18.

Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M. (2011) Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. Transportation Research Record 2263, pp. 140-150.

Nunes, A.A.N., Dias, T.G.D., Falcão e Cunha, J. (2016) Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE transactions on intelligent transportation systems 17, pp. 133-142.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J., White, P. (2006) The demand for public transport: The effects of fares, quality of service, income and car ownership. Transport Policy 13, pp. 295-306.

Pei, M., Lin, P., Liu, R., Ma, Y. (2019) Flexible transit routing model considering passengers' willingness to pay. IET Intelligent Transport Systems 13, pp. 841-850.

Pelletier, M.-P., Trépanier, M., Morency, C. (2011) Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, pp. 557-568.

Powers, D.M. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2, pp. 37-63.

Sabir, M. (2011) Weather and travel behaviour. Vrije Universiteit Amsterdam.

Saneinejad, S., Roorda, M.J., Kennedy, C. (2012) Modelling the impact of weather conditions on active transportation travel behaviour. Transportation Research Part D: Transport and Environment 17, pp. 129-137.

Singhal, A., Kamga, C., Yazici, A. (2014) Impact of weather on urban transit ridership. Transportation Research Part A: Policy and Practice 69, pp. 379-391.

Sorratini, J., Liu, R., Sinha, S. (2008) Assessing bus transport reliability using micro-simulation. Transportation Planning and Technology 31, pp. 303-324.

Stehman, S.V. (1997) Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment 62, pp. 77-89.

Stover, V.W., McCormack, E.D. (2012) The impact of weather on bus ridership in Pierce County, Washington. Journal of Public Transportation 15, p. 6.

Toqué, F., Côme, E., El Mahrsi, M.K., Oukhellou, L. (2016) Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1071-1076.

Trépanier, M., Chapleau, R. (2006) Destination estimation from public transport smartcard data. IFAC Proceedings Volumes 39, pp. 393-398.

Trépanier, M., Tranchant, N., Chapleau, R. (2007) Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems 11, pp. 1-14.

United Nations (2015) Transforming our world: The 2030 agenda for sustainable development. General Assembley 70 session.

Wang, W., Attanucci, J.P., Wilson, N.H. (2011) Bus passenger origin-destination estimation and related analyses using automated data collection systems. Journal of Public Transportation 14, p. 7.

Wei, M., Liu, Y., Sigler, T., Liu, X., Corcoran, J. (2019) The influence of weather conditions on adult transit ridership in the sub-tropics. Transportation Research Part A: Policy and Practice 125, pp. 106-118.

Wei, Y., Chen, M.-C. (2012) Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. Transportation Research Part C: Emerging Technologies 21, pp. 148-162.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wu, W., Jiang, S., Liu, R., Jin, W., Ma, C. (2019a) Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: Gradient boosting decision tree model. Transportmetica A: Transport Science In press.

Wu, W., Liu, R., Jin, W. (2016) Designing robust schedule coordination scheme for transit networks with safety control margins. Transportation Research Part B: Methodological 93, pp. 495-519.

Wu, W., Liu, R., Jin, W. (2017) Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour. Transportation Research Part B: Methodological 104, pp. 175-197.

Wu, W., Liu, R., Jin, W., Ma, C. (2019b) Stochastic bus schedule coordination considering demand assignment and rerouting of passengers. Transportation Research Part B: Methodological 121, pp. 275-303.

Xie, B., An, Z., Zheng, Y., Li, Z. (2019a) Incorporating transportation safety into land use planning: Pre-assessment of land use conversion effects on severe crashes in urban China. Applied geography 103, pp. 1-11.

Xie, B., Jiao, J., An, Z., Zheng, Y., Li, Z. (2019b) Deciphering the stroke–built environment nexus in transitional cities: Conceptual framework, empirical evidence, and implications for proactive planning intervention. Cities 94, pp. 116-128.

Yin, J., Yu, D., Yin, Z., Liu, M., He, Q. (2016) Evaluating the impact and risk of pluvial flash flood on intra-urban road network: A case study in the city center of Shanghai, China. Journal of Hydrology 537, pp. 138-145.

Yu, B., Lam, W.H., Tam, M.L. (2011) Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies 19, pp. 1157-1170.

Zannat, K.E., Choudhury, C.F. (2019) Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions. Journal of the Indian Institute of Science.

Zhang, X., Zhang, Q., Sun, T., Zou, Y., Chen, H. (2018) Evaluation of urban public transport priority performance based on the improved TOPSIS method: A case study of Wuhan. Sustainable Cities and Society 43, pp. 357-365.

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, pp. 376-387.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., Cao, R. (2017) Impacts of weather on public transport ridership: Results from mining data from different sources. Transportation Research Part C: Emerging Technologies 75, pp. 17-29.

Zhou, Z. (2016) Machine Learning. Tsinghua University Press.