J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions

Khatun E Zannat[1,2] *and Charisma F. Choudhury*[1*]

Abstract | The rapid advancement of information and communication technology has brought a revolution in the domain of public transport (PT) planning alongside other areas of transport planning and operations. Of particular significance are the passively generated big data sources (e.g., smart cards, detailed vehicle location data, mobile phone traces, social media) which have started replacing the traditional surveys conducted onboard, at the stops/stations and/or at the household level for gathering insights about the behavior of the PT users. This paper presents a systematic review of the contemporary research papers related to the use of novel data sources in PT planning with particular focus on (1) assessing the usability and potential strengths and weaknesses of different emerging big data sources, (2) identifying the challenges and highlighting research gaps. Reviewed articles were categorized based on qualitative pattern matching (similarities/dissimilarities) and multiple sources of evidence analysis under three categories—use of big data in (1) travel pattern analysis, (2) PT modelling, and (3) PT performance assessment. The review revealed research gaps ranging from methodological and applied research on fusing different forms of big data as well as big data and traditional survey data; further work to validate the models and assumptions; lack of progress on developing more dynamic planning models. Findings of this study could inform transport planners and researchers about the opportunities/challenges big data bring for PT planning. Harnessing the full potential of the big data sources for PT planning can be extremely useful for cities in the developing world, where the PT landscape is changing more rapidly, but traditional forms of data are expensive to collect.

1 Introduction

Rapid urbanization and associated increase in population are resulting in a higher growth of motorized traffic flow in urban areas. As a consequence, cities are experiencing different problems such as air pollution, road accidents, and congestions. In response to these problems, public transportation (PT) could help to reduce air pollution, road congestion and travel time, and dependency on non-renewable energy, which benefit both riders and non-riders[71]. Understanding travel pattern and travel behavior of PT users, forecasting their demand, performance improvement of PT service, etc., are central to PT planning. Traditionally, public/transport research domain was reliant on manual travel surveys conducted onboard, at the stops/stations or at household level[14, 17]. These methods are useful in describing socio-demographic characteristics (age, gender, income, occupation, etc.) of the respondents

[1] *Choice Modelling Centre, Institute for Transport Studies, University of Leeds, Leeds, UK.*
[2] *Department of Urban and Regional Planning, Chittagong University of Engineering and Technology (CUET), Chittagong 4349, Bangladesh.*
**c.f.choudhury@leeds.ac.uk*

Springer IISc Press

along with their detailed travel information (trip purpose, mode choice etc.). However, these methods are labor intensive, leading to higher costs and consequently smaller sample size and lower update frequencies. This is particularly problematic in the context of the developing countries in the Global South where the resources are more limited. Further, the data can have reporting errors and are prone to statistical biases[17, 27]. Most importantly, data collected through manual surveys are typically non-panel in nature and unable to capture the short-term/long-term variations in PT usage. Due to the coarse spatial characteristics and non-dynamic nature, it is typically not possible to combine such data with land use, weather, and dynamic network conditions. This makes it difficult to use it for PT planning and operational decisions.

On the other hand, the rapid advancement of information and communication technology (ICT) has brought revolution in the domain of transport research over the last decade. These range from PT-specific sources like smart cards used for automatic fare collection and global positioning system (GPS) traces for automatic vehicle location (AVL) to more generic data like digital footprints of mobile phone users, geocoded social media records, etc. The ubiquity of these data sources has led to passive generation of unprecedented amount of data, which are precisely geo-referenced and spatiotemporal in nature and suitable to explain human mobility patterns at much lower costs[2, 57, 58]. Other heterogeneous data sources include loop detectors (collects traffic data), probe vehicles (measures traffic condition), Bluetooth (enumerates travel times or average speeds and the associated variability), video cameras, remote sensing, street imagery such as Google Street View (GSV), and Bing StreetSide[27, 55]. Given that most of these data sources are large in volume, they fall into the general umbrella of big data.

In comparison with the traditional data sources, these novel sources of data show many unique attributes and advantages[17, 26, 69]. Firstly, big data sources contain updated and near or real-time spatial and temporal information that is quite impossible to collect through traditional travel survey (e.g., face to face interview, telephone interview, travel diary, and web form survey). Secondly, they contain a large amount of individual level data with greater details and higher accuracy at lower cost. Some of these data can be potentially linked with supplementary data (e.g., land use, bus time tables, etc.) as well as with each other (different types of data of the same person), though data linking has the risk of breaching privacy issues. Thirdly, these can be used to reconstruct large-scale trajectory[1] data for a larger sample size and longer observation period. The availability of such large dataset unveils possibilities for more dynamic research in the field of transportation planning. Despite these opportunities, there are also challenges associated with collection, processing and analyzing of big data, which need to be addressed while employing these data in transport planning. Further, mining of big data for its meaningful application requires different methods and techniques for data processing and transport modeling, which raise new technical challenges to ensure computational efficiency, data processing, integration, evaluation, validation and user privacy. Major challenges include: (1) presence of data gap (e.g., discontinuity in the location data, errors and missing information); (2) some details (e.g., trip purposes, accompanying travelers) are not explicitly recorded; (3) in case of some forms of big data, termed as extrinsic mobility data, the location information is generated by non-transport activity (e.g., during phone call, text message etc.); hence, cannot be converted directly to mobility data for transport studies which often require significant processing; (4) absence of personal or socio-demographic information of the user, which are key inputs to some of the traditional models (econometric models for example).

In the last two decades, many articles had reviewed the application of different big data sources in transportation planning. Among the current review-based studies, Anda et al.[5] reviewed the current application of big data sources to understand travel behavior and develop travel demand models for transport planning. They prescribed using big data for 'activity-based model' and 'agent-based simulations' to understand individual travel behavior. They further described recent advancements in data-mining methods, applied for trip identification, activity inference, and mode inference at the individual level. Yue et al.[74] reviewed current studies related to travel behavior, which used trajectory-based data and explored scattered technologies, tools, and data sources (traditional travel survey data, GPS log data, smart card data, mobile phone data, and other non-conventional sources such as social media and banknote data). They have also documented major challenges of using trajectory data in behavioral studies such as data

Mode choice: Mode choice is the process by which the trip between traffic analysis zones in the study area are allocated using available modes

Loop detector: Traffic loop detectors are sensors used for measuring traffic flows and speeds, typically at intersections, links, parking lots, etc.

Agent-based transport model: Agent-based transport models treat each decision maker as an agent (with characteristics, goals and behavioral rules) and simulate their actions on the environment they inhabit and traverse.

Activity-based model: Activity based models are based on the principle that travel is ultimately derived from the necessity to participate in activities. Activity-based models predict which activities are conducted when, where, for how long, for and with whom, and the travel choices they will make to complete them.

[1] Series of location points at discrete time intervals.

privacy, data accessibility for different stakeholders, socio-demographic biasness in available data, and absence of new data modeling techniques to reduce computational cost. Additionally, they highlighted new opportunities of complex network science, computational social science and bottom-up approach in travel behavior studies. Further, Wang et al.[69] provided a review of travel behavior research based on mobile phone data, obtained from different sources (cellular network and smartphone sensor-based data). They highlighted that the majority of these studies are focused on the re-identification of human travel patterns (such as travel frequency and distance), but failed to exhibit the inherent mechanism responsible for the observed pattern. They have also elaborated the major potentials and challenges of using mobile phone data in travel behavioral studies. Chen et al.[17] reviewed the current methodologies of using mobile phone data for travel behavior analysis in three sub-areas: modeling travel behavior, behavioral factor, and human mobility pattern. They raised emerging necessities for cross-discipline research with the aim of conversations and collaboration among different disciplines. Huang et al.[34] reviewed the advantages and disadvantages of different methods to detect transport mode based on mobile phone network data. They emphasized that most of the studies were focused on easy-to-detect modes due to the lack of ground truth data. Also, they highlighted that most of the studies did not validate their results, or simply validated their proposed methods with aggregated data. However, the above-mentioned reviews are generic, whereas the research needs, available big data sources, analyses, and analyses methodologies are distinctly different for PT.

In the context of PT, Pelletier et al.[57] focused on the application of smart card payment data in PT, showing that in addition to fare collection this data can be used for many purposes such as strategic-level studies (long-term planning), tactile-level studies (service adjustment, transfer journey studies, etc.), and operational level studies (calculation of PT performance indicators, payment management, etc.). This study however focuses only on smart card data and is slightly dated. In particular, over the last decade, there has been emergence of many other types of big data which can be used for PT planning. Therefore, further study is needed to investigate what the new big data sources have to offer and what novel methods can be deployed to best utilize them for informed decision-making to guide the improved performance of the PT modes.

The main purpose of this systematic review-based study is to explore the recent research in PT planning using big data and assess the usability and potentiality of the novel data sources (in comparison to conventional data such as household travel diary surveys and population census). Therefore, existing relevant literature is reviewed according to different aspects of PT planning such as analyses of travel pattern and understanding travel behavior. Our review differs from existing review-based articles in the following ways. Firstly, our focus is on the research, targeting to validate the application of big data source in PT planning. Secondly, the aim is to give the reader an overview of the application of big data in the domain of PT planning. The article is organized in the following order. Section 2 summarizes the methodological approach followed in this study. Section 3 includes available novel data sources used directly/indirectly in PT research. Section 4 critically reviews the selected paper. Section 5 explores the conclusions and future research path.

It may be noted that most of the articles related to big data use in PT studies were from North America and Asia. About 11% reviewed articles were from North American and 18% from the Asian context. The USA, Canada, UK, China, South Korea, and Australia were among the leading countries practicing big data extensively for PT planning. From the Global South, very few articles were from Chile and Brazil. Only one article was found in the African context.

2 Methodology of Review

This study adopted a three-stage systematic literature review approach proposed by Bask and Rajahonka[11]. The stage 1, entitled as "Planning stage", includes objectives and review protocol for a review, defining sources and procedures for article/paper searches. At this stage, we identified our research question based on our research aim and objective: What is the current state-of-the-art application of big data in PT planning? We further selected the inclusion and exclusion criteria for the final review. To ensure a comprehensive search, our database included five data sources: Scopus, Science Direct, Wiley Online Library, Taylor & Francis, and Google scholar. The following keywords were used to search articles: "Big Data" or "Smart Card" or "Mobile Phone" or "Social Media" or "Passive Data" AND "Public Transport" or "Transportation" AND "Planning". The same key words were used in the five data sources to avoid any biasness in the search process. A total of 272 articles were found after this

stage. The inclusion criteria were determined to fulfill the research aim.

The stage 2 is the "Screening", which includes descriptive and structure analysis. The titles and abstracts of 272 articles were screened for preliminary selection. The screening was primarily based on the following criteria: accepted/published academic journal articles, full text available, and published in English. We found a total of 102 articles at this stage. When the abstract of an article indicated a research aim similar to our study, the full text was scanned to include it in the review pool. Besides, we excluded articles appearing redundantly in different search engines, editorials, book chapters, conference proceedings, and articles not in English. Finally, 47 articles were selected for critical review. The selected 47 articles were further categorized based on qualitative pattern matching (similarities/dissimilarities) and multiple sources of evidence analysis (validation and acceptance of hypothesis)[73]. Thus, the aims and hypotheses considered in the selected article were tabulated and the articles were categorized under three themes according to the respective research purpose. Other supporting articles were also included during the review of the 47 articles to elaborate and support the overall findings of the reviewed paper.

In Stage 3 "Reporting and Dissemination", we organized our findings to write the review paper. In the writing stage, we followed the category developed in the "Review Stage" to summarize our findings. Eventually, we explained the significance of the reviewed paper and their contributions in the PT research domain along with the research gap and opportunities of future research.

Multimodal public transport system: Multimodal public transport system can be defined as a system through which a movement is accomplished using two or more different modes

3 Types of Big Data in PT Planning

Among the different definitions of big data, we refer to the simplest one: "any data that cannot fit into an Excel spreadsheet"[12]. These encompass automatically and routinely generated diverse dynamic information coming from different sources (e.g., sensors, devices, third parties, web applications, and social media) at various speeds and frequencies[12, 17, 59]. These data sources are thus also aligned with the definitions proposed by Laney[41] and [31], where big data is defined with 3 Vs (volume, velocity, variety) and 5 V + C (3 V + variability, veracity, complexity), respectively.

From the last two decades, a plethora of studies applied multiple passively collected data sources for transport planning research. However, the discussion regarding the application of big data for PT planning is fragmented and distributed in different outlets of PT planning domain. Systematic and comprehensive reviews on the application of big data in PT planning are unavailable. Hence, in this study we only focused on disparate application of the novel data sources for PT planning. The systematic review of literature revealed three key categories of data for PT planning.

(a) *Smart card data* The key purpose of smart card data is to ensure a smooth automatic fare collection in PT. Currently, many large-/medium-sized cities in the world such as London (Oyster card), New York (SmartLink), Boston (Charlie card), Beijing (Yikatong), and Hong Kong (Octopus card), have their own smart card system. These cards are based on radio-frequency identification (RFID) technology and passengers are required to tap the cards during entry and/or exit. Depending on the type of smart card, it automatically and continuously collects different trip records while using PT. For example, an entry-only smart card records information (boarding time, location of stop, transport mode, and/or stop number) when passengers enter/board a transit station[10, 25]. Hence, no information in destination stop is recorded. On the other hand, multimodal smart cards such as London Oyster card also collects information related to both entry and exit point and transport mode[28]. Therefore, information collected from the smart card can be used for PT planning other than merely the fare collection[57].

(b) *Mobile phone data* At present, most individuals carry mobile phone almost everywhere, which results in mobile phone data—the largest human mobility data source[8]. There are broadly two sources of mobile phone data—cellular network-based data and smartphone sensor-based data[69]. Cellular network-based data are collected by telecommunication companies. Two types of network-based data have been used in the contemporary PT studies: call data record (CDR) and global system for mobile communication (GSM) data[1, 43]. CDR data comprises a set of phone activity (phone call, text message or Internet access) records

along with the time and location information of cell towers channeling the call. The GSM data are generated from an interaction between a device and the mobile network as long as it is turned on[69]. For a single mobile phone, CDR or GSM data are dispersed and provide very little information. However, aggregation of thousands of mobile phone data overcomes the above-mentioned limitation[1]. Among the two types, GSM data has a higher frequency compared to CDR data, but is typically more difficult to get access to. CDR data on the other hand is routinely saved by the mobile phone companies for billing purposes and involves no additional effort in data provision.

On the other hand, smartphone sensor-based data can be collected by dedicated applications[69]. For example, check in information from social media (Twitter, Facebook, etc.) or popular sports tracking apps provide higher spatial resolution data compared to network-based data. However, this form of data is associated with serious sampling bias and poor temporal granularity; therefore, very few studies used this data form for PT research[58]. Nevertheless, advancement of Internet and smartphone technology provides a unique opportunity to transport planners/modelers to estimate demand fluctuations under special events (e.g., Olympic Games, Formula 1).

(c) *GPS data and automatic vehicle location (AVL)* The GPS technology installed in the vehicle enables collecting time, location, and service status of transport modes[66]. AVL database is prepared by collecting various geo-coded information about the mode (such as latitude, longitude, time, date) using GPS (on vehicle) at a constant or varying time interval. AVL data is widely used in conjunction with smart card data in different outlets of PT planning research[30, 48, 49, 54, 80], and extends the application of geographic information system (GIS) to perform spatial and temporal analysis in an urban landscape. Even though GPS data has many other applications (in private transport) such as car route mapping, measuring taxi service, freight tracking, and commercial fleet management, in this study we only considered GPS data that is collected in the PT modes and used to develop the AVL database.

Other heterogeneous data sources such as loop detectors, Bluetooth, and video camera can be considered as big data by definition and used in the transport research domain; however, these were not included in this study, as their application was non-specific in the reviewed articles on PT research domain.

4 Current State-of-the-Art of PT Studies Using Big Data

4.1 Theme 1: Use of Big Data in Travel Pattern Analysis

In the broad spectrum of urban planning and transport studies, it is important to understand and model how individuals move in time and space, called travel behavior analysis, which is important to understand the travel demand[14, 18]. In the context of PT, travel pattern and related statistical analyses are important while adopting plans to improve current and promote new PT services. In this section, we explore whether "big data" can substitute the conventional data collected through field survey for travel behavior analysis in PT planning.

More than one-third (36%) of the articles reviewed in this study were focused on travel behavior analyses of PT users using big data. Major research found under this theme is about—identification of aggregate/individual, single day/multiday travel behavior, inference of trip purpose, socio-demographic status of transit users, analysis of activity pattern, and spatial and temporal variability in transit use (Table 1). As seen in the Table 1, all authors mainly used smart card data for travel behavior analysis of transit users. Along with smart card data, AVL data were also incorporated in travel behavior analysis.

4.1.1 Aggregate vs Individual Travel Behavior

Day to day travel behavior analysis is difficult using conventional data, due to the cost and complexities associated with the data collection method. Among different novel data sources, smart card data has the capability to enhance existing decision-making tools, minimizing the complexities associated with conventional data collection systems (e.g., of onboard surveys)[42]. Several studies explored the potentiality of this data for aggregated and individual travel behavior analysis[37]. Zhang et al.[75] proposed a method to identify group travel behavior (GTB) with PT smart card data based on proxemics theory. Although aggregated behavior analysis draws the trip pattern of the general user, it fails to capture the individuality in travel behavior[36]. Hence, Kieu et al.[36] proposed a new method entitled Weighted

Table 1: Review of studies on big data for travel pattern analysis in PT planning. Source: prepared by the author, 2019.

	Article	Focus	Data used	Key findings
Aggregate vs. individual travel behavior	Zhao et al.[77]	Proposing a methodology to predict daily individual mobility and testing the method using smart card records from more than 10,000 users in London, U.K. over 2 years	Smart card data	Promising results obtained, which shows the proposed method can predict daily travel behavior and accuracy varies by the attributes considered
	Zhang et al.[75]	Developing a method to identify group travel behavior (GTB) of the PT user using 1-week records of subway fare card	Smart card data	Proposed method and smart card data have the potentiality to describe the characteristics and the spatio-temporal pattern of GTB
	Briand et al.[14]	Proposing a generative model to regroup PT passengers based on their temporal habits using smart card information collected for 5 years	Smart card data	Proposed unsupervised clustering method (Gaussian mixture model) makes it possible to model continuous representation of temporal travel pattern using smart card data
	Kieu et al.[36]	Developing an algorithm to understand individual passenger travel behavior using smart card data of 4 months period	Smart card data	Proposed Weighted Stop Density-Based Scanning Algorithm with Noise (WS-DBSCAN) algorithm is able to detect the daily changes in spatial travel pattern using smart card data with lower computation time
	Chu and Chapleau[21]	Proposing a method to enhance transit trip characterization by adding a multiday dimension to a month of smart card transactions	Smart card data and AVL data	Proposed rule-based algorithm and classification enables multiday travel behavior analysis of individual and subgroup using smart card transactions
Inference of travel behavioral attributes	Alsger et al.[3]	Proposing a model calibrated and validated to infer individual trip purpose of a PT user using smart card information	Smart card data, HTS[a], land use database, SEQSTM[b], GTFS, O–D survey data	Promising results obtained, shows a strong capability of the proposed model to predict trip purpose at a high level of accuracy
	Amaya et al.[4]	Estimating the residence zone of card users to enable socioeconomic variable for travel pattern analysis	Smart card data and OD survey data	The proposed method applicable for a segregate society allows to infer residence of the cardholders who are frequent PT users showing over 70% correct estimation
	Wang et al.[68]	Modeling location choice of metro commuters for after-work activities using smart card data	Smart card data	The proposed model performs well in explaining the station choices for after-work activities
	Goulet-Langlois et al.[29]	Proposing a methodology to identify clusters of PT users with similar activity sequence using smart card data	Smart card data and Socio-demographic information	Smart card data can be used to identify the connections existed between the demographic attributes of users and activity patterns identified exclusively from fare transactions

Table 1: continued

Article	Focus	Data used	Key findings
Han and Sohn[32]	Proposing a method to infer the sequences of activity using smart card information	Smart card data and land use information	Proposed continuous hidden Markov model (CHMM) is able to predict activity patterns which are consistent with the actual activity pattern
Long et al.[46]	Understanding extreme PT riders using smart card data	Smart card data and household travel survey data	Smart card data along with household survey data can be used to identify the spatio-temporal patterns of extreme transit behaviors
Arana et al.[7]	Determining the influence of meteorological conditions on transit ridership using smart card data	Smart card and AVL data	Smart card data can be used to determine the influence of external factor on PT ridership
Kusakabe and Asakura[40]	Developing a data fusion methodology to estimate behavioral attributes of trips using smart card data to observe continuous long-term changes in the attributes of trip	Smart card data and person trip survey data	Smart card data can successfully estimate (86.2%) the trip purpose using the proposed naïve Bayes probabilistic model which has low calculation load
Lee and Hickman[42]	Inferring trip purpose and activity information of transit users using smart card transaction data	Smart card data, land use, GTFS[c]	Inferences can be made through the trip purpose assignment process using a typical weekday smart card data integrated with land use data
Tao et al.[63]	Examining spatio-temporal dynamics or BRT trips in comparison with non-BRT trips using smart card data	Smart card data	Smart card data has the potentiality to reflect spatial heterogeneity in BRT/non-BRT trips
Ma et al.[49]	Proposing data-mining procedures that models the travel patterns of transit riders using smart card data	Smart card data	Using data mining techniques, smart card data can be used to understand individual travel pattern and travel regularity
Morency et al.[51]	Measuring the spatial and temporal variability of PT network use using smart card	Smart card data	Proposed data-mining techniques successfully provides continuous profile of spatial and temporal variability of transit use

[a] Household travel surveys (HTS)
[b] South East Queensland Strategic Transport Model (SEQSTM)
[c] Google's general transit feed specification (GTFS) is an open format updated by hundreds of transit agencies in the USA and used by Google to incorporate transit information

Stop Density-Based Scanning Algorithm with Noise (WS-DBSCAN) using smart card information to detect the spatial variability of individual travel pattern. Smart card data was further used for single day to multiday travel behavior analysis. Zhao et al.[77] proposed a regularized logistic regression model to predict daily individual travel pattern using historic sequence of individual trip records collected from smart card and reached 20–30% accuracy in predicting time, origin, and destination combinedly. In addition, studies proposed data-mining methods and probabilistic models to analyze multiday travel behavior and regularities of individual and subgroup[14, 21].

Since the proposed methodologies aided in predicting individual and aggregated trip patterns of PT users using smart card and AVL data, the outcome of the research would be useful to propose short-term and long-term policies and strategies to support predictable users. For example, the proposed trip prediction models can be used to develop traveler information system which will inform the user about service shortage and delays in certain areas[77]. As smart card and AVL data are collected routinely and passively, such behavioral prediction can be done within a short/long time interval which will provide the opportunities to evaluate the impact of policy changes.

4.1.2 Inference of Travel Behavioral Attributes

It is considered that additional attributes such as socio-demographic and activity information improves the understanding of travel behavior[21]. Smart card or other passive data sources are criticized due to the absence of socio-demographic information, which is collected in survey-based methods. Hence, efforts have been made to integrate socio-demographic information with big data to portray a comprehensive picture of travel behavior. Several studies proposed data-mining and probabilistic models to use smart card data to infer trip purposes[3, 23, 40, 42]. Goulet-Langlois et al.[29] followed agglomerative hierarchical clustering techniques to understand passenger heterogeneity from longitudinal representation of each user's multiweek activity using smart card data. Further, Amaya et al.[4] attempted to identify residence zone of smart card users to include this socioeconomic variables in travel pattern analysis.

Moreover, various studies assessed spatio-temporal variability of PT use using clustering techniques and probabilistic model integrating smart card multiday data record[38, 47, 49, 51, 63]. Wang et al.[68] proposed a location choice model of metro commuters for predicting after-work activity location using smart card data. The method proposed by Ma et al.[47] achieved 94.1% overall accuracy in identifying commuter trips. Further, Long et al.[46] classified different extreme PT riders using both traditional household data and smart card information. Besides, big data was used to determine the influence of external factors such as weather conditions on transit ridership using regression analysis[7]. Therefore, by synthesizing these novel data sources, more meaningful insights about travel behavior can be inferred.

Though manual survey contains socio-demographic attributes and activity information, this information is more often associated with sampling bias. Therefore, conventional models developed using manual survey-based data are often misleading to interpret travel pattern. On the other hand, the proposed models (followed classification, clustering, and prediction techniques) developed using big data are more dynamic in nature, capable of incorporating multiple data sources from fine to coarse resolution to predict socio-demographic attributes, and reduces the requirement of conventional survey for want of socio-demographic information.

4.2 Theme 2: Use of Big Data in PT Modeling

Since the 1960s, one of the most prominent and widely acknowledged transport modeling approaches has been the four-stage modeling[44]. It is widely used as a systematic framework for both public and private transport modeling, which follows the sequence of (a) trip generation, (b) trip distribution, (c) modal split, and (d) trip assignment. This modeling approach requires a large amount of spatio-temporal individual trip information and demographic data, the collection of which is a manpower-, time-, and investment-intensive process[5]. After the emergence of big data sources, efforts have been made to use big data as an alternative to conventional survey data in transport modeling. But the question remains as to whether it is possible to replace the costly conventional survey data using big data in PT modeling. About 14 (30%) articles reviewed in this study answered this question.

Among the different sources of big data, the use of smart card information in PT modeling is observed in all 14 relevant articles reviewed in this study (Table 2). After the emergence of "smart card" for PT fare payment, it is considered as a useful data source for the planner and researcher, since it produces a large amount of

Trip distribution: Trip distribution is the process by which all the trips generated in a study area are allocated among the zones

Trip generation: Trip generation is the process of estimating the number of trips that will begin or end in each zone within a study area.

Location choice model: Location choice model predicts the individual household's choice of residence to aggregated zones with an aim to quantify the impact of different residential location characteristics and their interaction with household characteristics

Table 2: Review of studies on big data for PT modeling. Source: prepared by the author, 2019.

	Article	Focus	Data used	Key findings
O–D estimation	Kumar et al.[39]	Inferring the trajectory of PT user using trip-chaining method, deriving information from smart card data	Smart card data and GTFS data	Proposed trip-chaining algorithm tries to relax the assumptions on the parameters such as GPS inaccuracy buffer zone for boarding stop inference for origin and destination inference
	Tamblay et al.[62]	Inferring the zonal O–D matrix for observed trip between two PT stops (i.e., bus stops or metro stations)	Smart card data, socioeconomic, land use, and network information	Smart card data can be used with other data to infer zonal O–D matrix and the proposed model captures the expected effect of land use on trip generation and trip distribution
	Gordon et al.[28]	Proposing a method to infer origins and destination matrices for bus passenger	Smart card data, AVL and survey data	Smart card data is a useful source to infer boarding and alighting times and locations for individual bus passengers, transfers between passenger trips of various public modes, and origin–destination matrices of linked intermodal transit journeys
	Munizaga and Palma[54]	Proposing a method suitable for large multimodal PT system to predict origin–destination (OD) matrix obtained from entry-only smartcard data	Smart card data, AVL data and geo-coded PT network	Linked journey O–D matrix can be constructed for a multimodal PT system using interchange inference algorithm using smart card data
	Wang et al.[67]	Inferring O–D for bus passenger using data from automated data collection systems (ADCS)	Smart card data, AVL and manual survey data	It is feasible and easy to apply trip-chaining to infer O–D for both weekend and weekday and alighting stop of bus passenger, using smart card data
	Barry et al.[9]	Proposing a method to determine O–D trip tables by using entry-only smart card data for all PT user (subway, local and express buses, ferry, and tramway)	Smart card data	O–D matrix can be developed using smart card data for multimodal transport network even when AVL data is not available
	Chu and Chapleau[20]	Derive more complete information from smart card for planning purposes	Smart card data	Smart card can be used to infer passenger journeys, analyze transfer activity and synthesize vehicle load profile for the better analysis of linked trips, trip chains, and activity space using run time estimation
	Trépanier et al.[65]	Estimating trip destination using smart card fare collection data	Smart card data	Smart card data can be used for destination inference of bus passenger when alighting information is not available

Table 2: continued

	Article	Focus	Data used	Key findings
	Zhao et al.[76]	Proposing a method for predicting rail passenger O–D table from smart card data	Smart card data and AVL data	Smart card data can be used to infer O–D matrix (rail to rail and rail to bus) for rail passenger from an origin-only smart card data to replace expensive passenger O–D surveys.
	Barry et al.[10]	Proposing a method to determine station-to-station O–D trip tables by using entry-only smart card data for subway user (unimodal trip)	Smart card data, Travel diary survey data for validation purpose	Proposed destination inference model using metro card information can be used to create O–D matrix for different temporal scales
	Farzin[25]	Creating O–D matrix using multiple data source along with smart card data	Smart card data and AVL data	Multiple data can be infused with smart card data to develop and validate zonally aggregated O–D matrix
Route choice modeling	Cheon et al.[19]	Developing a method for analyzing the route choice of travelers in multimodal transit networks by considering multiple attributes	Smart card data	The developed model considers coexistence of various mode in a single network including multiple attributes to effectively reduce the unreasonable paths with high accuracy
	Nassir et al.[56]	Proposing a path choice model using smart card	Smart card data	Proposed recursive link-based choice model allows model calibration with incomplete path choice observations with transit smart card data in higher-frequency bus and rail services
	Jánošíková et al.[35]	Estimating route choice model for urban PT using smart card data	Smart card data, street map and time table	In-vehicle travel time, transfer walking time and to get from alighting stop to trip destination, the need to change, and the time headway of the first transportation line, can be determined by the combination of smart card data with other data sources

boarding/alighting information depending on the type of card[57].

4.2.1 Estimation of Origin and Destination (O–D)

At the beginning of the last decade, this form of data was employed to elicit basic information required for transport modeling such as origin and destination (O–D) to quantify transport demand between geographical regions in a city. Various algorithms were proposed to determine station-to-station O–D trip tables by using smart card data for unimodal PT trip (subway/bus)[10, 25, 65, 67]. Then, various studies proposed methods suitable for large multimodal PT system to estimate the O–D matrix from smart card data[9, 28, 54, 76]. As such, along with rail to rail trip, Zhao et al. [76] considered rail to bus trip to develop the O–D matrix. Further, Barry et al. [9] incorporated multiple modes (subway, local and express buses, ferry, and tramway) for O–D estimation. In these studies, other data sources such as AVL data and geocoded GPS data were infused with smart card information. Using the proposed algorithms by Barry et al. [9], O–D trip tables were created for short-term and long-term demand estimation. Besides disaggregate O–D estimation, a study conducted by Tamblay et al. [62] also attempted to infer zonal O–D matrix form smart card information to reflect the city-wise travel demand. Since PT journeys often comprise multiple transfers, smart card data were also used in the contemporary studies to understand linked trips to derive more complete information of PT trip. Methods were proposed to infer passenger journeys and analyze transfer pattern [20, 28, 54, 60]. Through the analysis of transfer pattern of linked trip, multiple modes used to complete the trip were also detected to complete the O–D estimation.

To validate the various methods for the use of smart card data and other passive data sources in O–D estimation, multiple validation techniques have been proposed in several studies. As such, travel diary, manual survey data, and historical observations were used to validate the methods proposed to estimate the O–D matrix [10, 28]. These methods can predict O–D with an accuracy level ranging from 66% to 90%. In addition, Munizaga et al. [53] proposed an endogenous validation method (analyze the data to verify assumption) to validate the assumptions considered for PT origin–destination (O–D) matrices using smart card data and survey data. On the other hand, Kumar et al.[39] proposed a new trip-chaining algorithm for O–D inference that tries to relax the assumptions on the parameters such as GPS inaccuracy (buffer zone for boarding stop inference).

4.2.2 Route Choice Modeling

Three articles dealt with route choice of travelers in PT networks by considering multiple attributes. The interchanges between the segments of a linked journey can be recognized using smart card data infused with other data sources[35]. Nassir et al.[56] proposed a recursive link-based path choice model using smart card data in higher-frequency bus and rail services and added a new measure of "attractiveness" to allow for randomness in the choice of attractive routes. Cheon et al.[19] proposed a trip assignment model considered the coexistence of various modes in a single network considering multiple attributes to effectively reduce the unreasonable paths. Besides, Kim et al.[37] introduced a new attribute entitled as 'stickiness' to understand individuals' habitual route choice through cross-sectional and longitudinal analysis using the whole trajectories of individual passengers, constructed from smart card data.

Therefore, the application of big data in PT modeling is observed in three (trip generation, distribution, and trip assignment) of the four stages of transport modeling. Future research is needed to delineate modal classification using big data source. Thus far, it was attempted to establish passively collected data as an alternative source of conventional survey data for PT modeling. By using different data sources (smart card, AVL, travel survey data, etc.), the proposed models exhibit promising result in developing aggregate and disaggregate O–D matrix. After precise identification or prediction of origin and destination of different trips, outcomes (spatial and temporal information of trip) from the models would be useful for further analysis to understand the reasons behind the trip-making behavior, seasonal and daily travel demand, and macro- to micro-level interaction among the factors governing travel pattern,. Also, the proposed methods overcome the challenges to capture the complex travel pattern in modern PT system (involving multiple transfers and modes) that may be quite impossible to infer precisely from traditional survey data (e.g., cordon line, screen line, or household survey data).

Traffic assignment/Route choice model: Traffic assignment determine the routes that will be used.

Trip chain: A trip chain is travel involving multiple purposes to single or multiple destinations and begins and ends at home or a similar origin.

Table 3: Review of studies on big data in PT performance improvement. Source: prepared by the author, 2019.

	References	Aim		Major observation
Measurement of performance assessment indicators	Lee et al.[43]	Measuring the accessibility to PT using mobile phone data	CDR data and census data	Proposed Huff model-based floating catchment area method measures reliable time-varying accessibility to PT using mobile phone data
	Tavassoli et al.[64]	Modeling passenger waiting time at transit stop using smart card data	Smart card data and GTFS data	Log-logistic AFT models is inferred to be the best fit for passenger waiting time using smart card data
	Tu et al.[66]	Analyzing the spatial variations of urban PT ridership using smart card data	Smart card data and GPS trajectories	The effects of demographic, land use and transportation factors on the ridership of PT can be explored using GWR analysis using smart card data
	Zhou et al.[79]	Developing a model to calculate bus arrival time using smart card data	Smart card and travel survey data	Bus arrival time is calculated using the distribution of the card swiping time distribution, occupancy and the seating capacity information
	Zhu et al.[80]	Proposing a model to infer left behind passenger using smart card data	Smart card and AVL data	The estimated probabilities of passengers being left behind using smart card data is similar to manual survey results and provide crowding information
	Hong et al.[33]	Estimating both the physical and schedule-based connections of metro passengers by examining the Smart Card data	Smart card data	Promising results obtained and the model estimated precisely the passengers boarding, transferring, and alighting of trains based on the entry and exit times and stations of a passenger
	Min et al.[50]	Proposing method to recover the arrival times of trains from the gate times of metro passengers from smart card data	Smart card data	The proposed method is applicable, when logs are missing for an entire line and the procedure recovered the arrival time of higher accuracy
	Aguiléra et al.[1]	Measuring quality of service and passenger flows using mobile phone data	GSM, smart card data and field observation data	Train occupancy levels, travel times, and origin–destination flows is estimated at a very fine-grain level using GSM data and compared with the field observations (train trajectory) and smart card data

Table 3: continued

	References	Aim		Major observation
Evaluation of performance	Liu et al.[44]	Proposing method to replicate the multi-modal PT system	Smart card data	The disaggregated replication provides trip information with precision of a few minutes and the outputs are precise temporal and spatial travel demand analysis, transfer pattern analysis, traffic condition investigation and bus utilization analysis
	Liu et al.[45]	Assessing the impact of fare policy change on ridership using smart card information	Smart card data	Impact of fare policy on PT ridership can be assessed through the comparison of number of card users, their journeys, and travel costs before and after the policy reform
	Zhou et al.[78]	Explicating the potentiality of big data in quantifying and visualizing the relationship between transit fare, space and justice	Smart card data	Proposed and implemented methods such as "trajectory rebuilding", "fare matching", "segment tagging", "desired line/stop visualization", "commuter identification" and "scenario analysis" using smart card data
	Moyano et al.[52]	Evaluating the importance of access and egress times to/from HSR stations in an urban context	GTFS, speed profile and Twitter data	Travel time measures are analyzed temporally and spatially for access/egress to/from stations considering both taxis and PT
	Yap et al.[72]	Improve the prediction accuracy of the impact of planned, temporary disturbances on PT usage	Smart card data	Proposed rule-based three-step search procedure results in higher accuracy in predicting PT usage during disturbances
	Pereira et al.[58]	Predicting PT arrivals under special events using Internet	Internet	Proposed a methodology extracts events information from the Internet and matches such information with bus and subway tap-in/tap-out data
	Williams et al.[70]	Collecting a comprehensive data set on a semi-formal transit system using cell phones	Cell phone data	The proposed method shows how to transform cell phone data into a GTFS format useful for planning, research, operations, and transit routing applications
	Ma and Wang[48]	Developing a data-driven platform for online PT performance assessment	Smart card data, AVL data and geospatial data	The proposed framework demonstrates several transit performance indicators at different scales (e.g., network level, route level, and stop level) and the feasibility of establishing a web-based e-science system for transit performance measures

4.3 Theme 3: Use of Big Data in PT Performance Assessment

Transit performance assessment: Transit performance assessment is a framework for evaluating public transport service quality from the perspective of different stakeholder (user/operator/government authority).

In recent years, big data has been applied in PT performance assessment. By reviewing 16 selected articles under this theme, we attempted to answer the question on whether big data can substitute conventional data in assessing performance of PT service. Unlike the previous two themes, where the hegemony of smart card was observed, here, the use of other passive data sources (e.g., mobile phone, social media) along with big data was evident.

Performance assessment has become an integral part in transport planning and management, which helps to ensure accountability, transparency, and service quality[16]. The resultant information from performance assessment is necessary for the decision makers to evaluate investment alternatives in the transport sector[61]. Therefore, augmenting the multipurpose uses of performance assessment expedites the development of performance measures, which are the integral part of the performance assessment[13]. Review of studies on big data in PT performance improvement is summarized in Table 3.

Performance Measures: Performance Measures are the indicators used to assess the performance. For example: for the performance assessment of public transport service ridership, cost efficiency, on-time performance, customer service, and financial performance are potential measures to consider.

4.3.1 Measurement of Performance Assessment Indicators

To use big data in performance assessment of PT services, attempts have been made to measure different performance indicators of PT. The majority of the existing contributions focused on developing methodologies for PT performance assessment. In the reviewed articles, big data was used to estimate regular performance measures such as quality of PT service using GSM data[1], physical and schedule-based connections of metro user using quadruple[33], bus arrival time using smart card data[79], left behind passenger using smart card data and AVL data[80], accessibility to PT service using mobile phone data[43], passenger waiting time using smart card data[64], and spatial variations of urban PT ridership using GPS trajectories and smart card data[66]. Further, Min et al.[50] proposed a method to recover the arrival times of trains from the exit times of metro passengers.

Accessibility: Accessibility can be defined by the "ability to access" and benefits from a system or entity. In transport planning, accessibility refers to a measure of the ease of reaching (and interacting with) destinations or activities distributed in space, e.g. around a city or country.

4.3.2 Evaluation of Performance

The use of big data is also observed in evaluating the performance of PT service such as the importance of access and egress times to/from high speed railway (HSR) stations[52], impact of fare policy change on PT ridership[45], and the relationship between transit fare, space and justice[78]. Also, an attempt has been made to create online data-driven platform for performance measurement in Beijing, China[48]. Liu et al.[44] proposed a method to replicate the multimodal PT system using smart card data and the resulting replication covers about 96% of trips made in PT in Singapore. Also, using cell phone data, a comprehensive dataset was built for para-transit service for performance improvement in Nairobi, Kenya[70]. Use of big data is also observed to monitor special events/circumstances. Pereira et al.[58] developed method to predict PT arrivals on the time of special events using Twitter data. In addition, to ensure prediction accuracy of the impact of planned, temporary disturbances (such as temporary track closures) of PT usage, Yap et al.[72] proposed a method using smart card data.

To initiate PT improvement and management programs, the prerequisite is to measure the performance of existing PT system to elicit problems, their root causes, and sectors requiring special attention[24]. Even though the performance assessment of PT service has been acknowledged as an effective planning, management, monitoring and evaluation tool, it is less prioritized and practiced in developing countries due to the scarcity of data required for performance assessment, even after the implementation of a project. Lack of tangible outcomes from performance assessment could create political negligence, compared to investment-intensive infrastructure development program. The application of big data for measuring performance indicators (e.g., service quality, accessibility to PT) and evaluation of performance (under special circumstances) has overcome the budget constraint associated with conventional data collection methods. An evolution of such application could enable the decision maker to use big data in public transport policy making[6].

5 Conclusion and Future Research Direction

This systematic review-based study critically analyzed the current advances in application of big data in PT planning. Following a three-stage review process, we categorized 47 review paper under three subsection—travel pattern analysis, PT modeling, and PT performance assessment. It is found that the high potential and usefulness of big data (particularly smart card data, mobile data, and AVL data) in PT planning is widely acknowledged. The general finding is that the emerging big data sources provide at least as good if not better models/tools for PT planning.

Further, it is claimed in the majority of the studies that due to the 'by-product' nature of such data, these tools are cheaper to develop compared to those involving collection of traditional data.

However, the majority of the reviewed studies have focused on investigating conventional PT planning topics (e.g., O–D estimation, route choice modeling) with passively collected data sources. There is a research gap in extending these to discover more novel applications of big data for PT planning. One particular promising direction in this regard is the potential to develop more dynamic planning models that better utilize the panel nature of the data in modeling the variability in behavior and the interaction of different influencing variables over time. For example, the trip-chaining data available from the smart cards can be utilized to optimize the overall transfer times; the increased/decreased boarding numbers on a certain bus stop can be utilized to make minor adjustments in the dwell time at the subsequent stops to prevent systematic 'bus bunching'; establishing relationship between weather and OD patterns can be used to improve the seasonal changes in the time table.

Existing algorithms and models in the contemporary studies to predict origin, destination, and route choice have many applications and substantial level of accuracy. But, the development of these models involved multiple steps and considered many assumptions and sampling approximation. While the validity of these models depends on the accuracy of these assumptions and sampling approximation, very few attempts have been made to validate these assumptions in a context different from the training/estimation context[53]. Therefore, future research is needed to propose new techniques to validate the underlying assumptions in PT modeling using big data and potentially provide insights about their forecasting performance.

Further, though there are possibilities of combining multiple big data sources or big and small data for PT planning (e.g., smart card data with mobile phone data), only few studies have focused on this aspect on a limited scale. There is potentiality to develop wide-scale models combining land use, weather, events, and other big data sources to understand travel behavior in different landscapes. While there exist various sources of big data, methods to integrate these different types of data to solve problems (e.g., congestion and service deficiency), which could improve the accuracy in predicting individual travel behavior, do not exist. Existing sources of big data are often criticized for not providing information such as socio-demographic characteristics of the passengers. Integrating data collected from small-scale surveys and correcting the potential biases with the big data sources (as proposed by Bwambale et al.[15] in the context of generic travel behavior) can provide more insights into travel behavior.

In addition, cross-cutting research is needed to explain the applicability of big data in the extended domain of PT, such as spatio-temporal relationship between origin and destination choice, transit users' transfer choice, and mode choice behavior[19]. Besides, disjointed O–D estimation, route choice, transfer choice, and mode choice research can be integrated to understand the complex trip chain and travel behavior[25]. Trip-chaining and linked trip analysis can be further extended for inferring trip purpose, analyzing spatial and temporal travel pattern, and analyzing route choice behavior of passenger[32, 39]. To improve the performance of PT service, research is needed to determine the relationship between passenger travel demand and performance indicator such as speed of vehicle, quality of service, accessibility to PT, and passenger waiting time[22, 43, 64].

In understanding travel behavior (individual/aggregated), clustering techniques such as density-based spatial clustering, K-means++ clustering, and Gaussian mixture model are generally applied. However, application of supervised classification process, interpretation, and validation with onboard data are needed to classify the heterogeneous transit user[42]. But, to understand activity-based travel behavior in PT use, none of the studies have either used big data in developing agent-based individual model or implemented model in an open-sourced agent-based micro-simulation tool[5]. Therefore, advanced mathematical model, machine learning toolkit, and application of spatial statistics integrated with spatial analysis are needed to understand the interaction of PT user with the surroundings. Since, big data has the potentiality to provide both short-/long-term records, future research can evaluate alternative scenarios of different PT policy environment (e.g., before and after policy implementation), which will enable to understand the potential impacts of a PT policy[4, 22].

Finally, apart from a few reviewed papers, big data has been used in transport research predominantly in context of developed countries, similar uses of such data in planning and operation of PT system is needed in developing countries, where the PT landscape is changing more rapidly.

It is expected that in future the accuracy and precision of the PT data will improve over time

Mode choice: Mode choice is the process by which the trip between traffic analysis zones in the study area are allocated using available modes

Traffic microsimulation: Traffic microsimulation is an agent-based analytical tool to simulate the decisions of each travler to deduce the network conditions. They are used as laboratories to evaluate the effects of proposed interventions before the implemented in the real world

leading to fewer missing data and gaps in the trajectory. The improved data quality holds the promise of leading to more robust PT planning models. A more promising direction is, however, the emergence of multimodal data (e.g., data from ride-sharing modes, shared bikes). These emerging data sources are promising for PT planning from two aspects. Firstly, such data can be leveraged to optimize the transport network as a whole as opposed to PT only. Secondly, they can be used to infer the latent PT demand and take PT planning measures to maximize the revenue.

This review could be a useful guide for fellow researchers who intend to work with "big data" and "PT planning", which will contribute in promoting PT. Despite the ever-increasing demand for car use, hopefully, academic research with big data will provide useful guideline on how to reduce car use considering the current situation of PT usage. Hence, exploring new methods and techniques is essential to employ big data in accurately explaining travel behavior and improving PT system.

Publisher's Note

Open Access

Received: 18 July 2019 Accepted: 19 September 2019
Published online: 9 October 2019

References

1. Aguiléra V, Allio S, Benezech V, Combes F, Milion C (2014) Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transp Res Part C Emerg Technol 43:198–211
2. Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C Emerg Technol 58:240–250
3. Alsger A, Tavassoli A, Mesbah M, Ferreira L, Hickman M (2018) Public transport trip purpose inference using smart card fare data. Transp Res Part C Emerg Technol 87:123–137
4. Amaya M, Cruzat R, Munizaga MA (2018) Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. J Transp Geogr 66:330–339
5. Anda C, Erath A, Fourie PJ (2017) Transport modelling in the age of big data. Int J Urban Sci 21:19–42
6. Aragona B, De Rosa R (2019) Big data in policy making. Math Popul Stud 26(2):107–113
7. Arana P, Cabezudo S, Peñalba M (2014) Influence of weather conditions on transit ridership: a statistical study using data from Smartcards. Transp Res Part A Policy Pract 59:1–12
8. Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Puchinger J (2019) Inferring dynamic origin–destination flows by transport mode using mobile phone data. Transp Res Part C Emerg Technol 101:254–275
9. Barry JJ, Freimer R, Slavin H (2009) Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transp Res Rec 2112:53–61
10. Barry JJ, Newhouser R, Rahbee A, Sayeda S (2002) Origin and destination estimation in New York City with automated fare system data. Transp Res Rec 1817:183–187
11. Bask A, Rajahonka M (2017) The role of environmental sustainability in the freight transport mode choice: a systematic literature review with focus on the EU. Int J Phys Distrib Logist Manag 47:560–602
12. Batty M (2013) Big data, smart cities and city planning. Dialog Hum Geogr 3:274–279
13. Benjamin J, Obeng K (1990) The effect of policy and background variables on total factor productivity for public transit. Transp Res Part B Methodol 24:1–14
14. Briand A-S, Côme E, Trépanier M, Oukhellou L (2017) Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transp Res Part C Emerg Technol 79:274–289
15. Bwambale A, Choudhury CF, Hess S, Iqbal MS (2019) Getting the best of both worlds—a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. In: 15th World conference on transport research India
16. Carter DN, Lomax TJ (1992) Development and application of performance measures for rural public transportation operators. Transp Res Rec 1338:28–36
17. Chen C, Ma J, Susilo Y, Liu Y, Wang M (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp Res Part C Emerg Technol 68:285–299
18. Chen SH, Wu CC, Li PY, Adhitana Paramitha P (2017) Evaluation of pedestrian transportation facilities in Taiwan using linear regression and support vector regression. Road Mater Pavement Des 18:170–179
19. Cheon SH, Lee C, Shin S (2019) Data-driven stochastic transit assignment modeling using an automatic fare

collection system. Transp Res Part C Emerg Technol 98:239–254
20. Chu KKA, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transp Res Rec 2063:63–72
21. Chu KKA, Chapleau R (2010) Augmenting transit trip characterization and travel behavior comprehension: multiday location-stamped smart card transactions. Transp Res Rec 2183:29–40
22. Cortés CE, Gibson J, Gschwender A, Munizaga M, Zúñiga M (2011) Commercial bus speed diagnosis based on GPS-monitored data. Transp Res Part C Emerg Technol 19:695–707
23. Devillaine F, Munizaga M, Trépanier M (2012) Detection of activities of public transport users by analyzing smart card data. Transp Res Rec 2276:48–55
24. Dhingra C (2011) Measuring public transport performance: lessons for developing countries. Sustain Urban Transp Tech Doc 9:1–43
25. Farzin JM (2008) Constructing an automated bus origin–destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. Transp Res Rec 2072:30–37
26. Gadziński J (2018) Perspectives of the use of smartphones in travel behaviour studies: findings from a literature review and a pilot study. Transp Res Part C Emerg Technol 88:74–86
27. Goel R, Garcia LM, Goodman A, Johnson R, Aldred R, Murugesan M, Brage S, Bhalla K, Woodcock J (2018) Estimating city-level travel patterns using street imagery: a case study of using Google Street View in Britain. PLoS One 13:e0196521
28. Gordon JB, Koutsopoulos HN, Wilson NH, Attanucci JP (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp Res Rec 2343:17–24
29. Goulet-Langlois G, Koutsopoulos HN, Zhao J (2016) Inferring patterns in the multi-week activity sequences of public transport users. Transp Res Part C Emerg Technol 64:1–16
30. Gschwender A, Munizaga M, Simonetti C (2016) Using smart card and GPS data for policy and planning: the case of transantiago. Res Transp Econ 59:242–249
31. GSR (2015) Big Data [Online]. GSR Technologies Inc. http://www.gsrti.com. Accessed 9 Mar 2016
32. Han G, Sohn K (2016) Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. Transp Res Part B Methodol 83:121–135
33. Hong S-P, Min Y-H, Park M-J, Kim KM, Oh SM (2016) Precise estimation of connections of metro passengers from Smart Card data. Transportation 43:749–769
34. Huang H, Cheng Y, Weibel R (2019) Transport mode detection based on mobile phone network data: a systematic review. Transp Res Part C Emerg Technol 101:297–312
35. Jánošíková Ľ, Slavík J, Koháni M (2014) Estimation of a route choice model for urban public transport using smart card data. Transp Plan Technol 37:638–648
36. Kieu L-M, Bhaskar A, Chung E (2015) A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. Transp Res Part C Emerg Technol 58:193–207
37. Kim J, Corcoran J, Papamanolis M (2017) Route choice stickiness of public transport passengers: measuring habitual bus ridership behaviour using smart card data. Transp Res Part C Emerg Technol 83:146–164
38. Kim M-K, Kim S, Sohn H-G (2018) Relationship between spatio-temporal travel patterns derived from smart-card data and local environmental characteristics of Seoul, Korea. Sustainability 10:787
39. Kumar P, Khani A, He Q (2018) A robust method for estimating transit passenger trajectories using automated data. Transp Res Part C Emerg Technol 95:731–747
40. Kusakabe T, Asakura Y (2014) Behavioural data mining of transit smart card data: a data fusion approach. Transp Res Part C Emerg Technol 46:179–191
41. Laney D (2001) 3D data management: controlling data volume, velocity and variety. META Group Res Note 6:1
42. Lee SG, Hickman M (2014) Trip purpose inference using automated fare collection data. Public Transport 6:1–20
43. Lee WK, Sohn SY, Heo J (2018) Utilizing mobile phone-based floating population data to measure the spatial accessibility to public transit. Appl Geogr 92:123–130
44. Liu X, Zhou Y, Rau A (2019) Smart card data-centric replication of the multi-modal public transport system in Singapore. J Transp Geogr 76:254–264
45. Liu Y, Wang S, Xie B (2019) Evaluating the effects of public transport fare policy change together with built and non-built environment features on ridership: the case in South East Queensland, Australia. Transp Policy 76:78–89
46. Long Y, Liu X, Zhou J, Chai Y (2016) Early birds, night owls, and tireless/recurring itinerants: an exploratory analysis of extreme transit behaviors in Beijing, China. Habitat Int 57:223–232
47. Ma X, Liu C, Wen H, Wang Y, Wu Y-J (2017) Understanding commuting patterns using transit smart card data. J Transp Geogr 58:135–145
48. Ma X, Wang Y (2014) Development of a data-driven platform for transit performance measures using smart card and GPS data. J Transp Eng 140:04014063
49. Ma X, Wu Y-J, Wang Y, Chen F, Liu J (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C Emerg Technol 36:1–12
50. Min Y-H, Ko S-J, Kim KM, Hong S-P (2016) Mining missing train logs from Smart Card data. Transp Res Part C Emerg Technol 63:170–181
51. Morency C, Trépanier M, Agard B (2007) Measuring transit use variability with smart-card data. Transp Policy 14:193–203

52. Moyano A, Moya-Gómez B, Gutiérrez J (2018) Access and egress times to high-speed rail stations: a spatiotemporal accessibility analysis. J Transp Geogr 73:84–93
53. Munizaga M, Devillaine F, Navarrete C, Silva D (2014) Validating travel behavior estimated from smartcard data. Transp Res Part C Emerg Technol 44:70–79
54. Munizaga MA, Palma C (2012) Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile. Transp Res Part C Emerg Technol 24:9–18
55. Nantes A, Ngoduy D, Bhaskar A, Miska M, Chung E (2016) Real-time traffic state estimation in urban corridors from heterogeneous data. Transp Res Part C Emerg Technol 66:99–118
56. Nassir N, Hickman M, Ma Z-L (2019) A strategy-based recursive path choice model for public transit smart card data. Transp Res Part B Methodol 126:528–548
57. Pelletier M-P, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C Emerg Technol 19:557–568
58. Pereira FC, Rodrigues F, Ben-Akiva M (2015) Using data from the web to predict public transport arrivals under special events scenarios. J Intell Transp Syst 19:273–288
59. Russom P (2011) Big data analytics. TDWI best practices report. Data Warehousing Institute, Renton
60. Seaborn C, Attanucci J, Wilson NH (2009) Analyzing multimodal public transport journeys in London with smart card fare payment data. Transp Res Rec 2121:55–62
61. Sheth C, Triantis K, Teodorović D (2007) Performance evaluation of bus routes: a provider and passenger perspective. Transp Res Part E Logist Transp Rev 43:453–478
62. Tamblay S, Galilea P, Iglesias P, Raveau S, Muñoz JC (2016) A zonal inference model based on observed smart-card transactions for Santiago de Chile. Transp Res Part A Policy Pract 84:44–54
63. Tao S, Corcoran J, Mateo-Babiano I, Rohde D (2014) Exploring Bus Rapid Transit passenger travel behaviour using big data. Appl Geogr 53:90–104
64. Tavassoli A, Mesbah M, Shobeirinejad A (2018) Modelling passenger waiting time using large-scale automatic fare collection data: an Australian case study. Transp Res Part F Traffic Psychol Behav 58:500–510
65. Trépanier M, Tranchant N, Chapleau R (2007) Individual trip destination estimation in a transit smart card automated fare collection system. J Intell Transp Syst 11:1–14
66. Tu W, Cao R, Yue Y, Zhou B, Li Q, Li Q (2018) Spatial variations in urban public ridership derived from GPS trajectories and smart card data. J Transp Geogr 69:45–57
67. Wang W, Attanucci JP, Wilson NH (2011) Bus passenger origin–destination estimation and related analyses using automated data collection systems. J Public Transp 14:7
68. Wang Y, De Almeida Correia GH, de Romph E, Timmermans H (2017) Using metro smart card data to model location choice of after-work activities: an application to Shanghai. J Transp Geogr 63:40–47
69. Wang Z, He SY, Leung Y (2018) Applying mobile phone data to travel behaviour research: a literature review. Travel Behav Soc 11:141–155
70. Williams S, White A, Waiganjo P, Orwa D, Klopp J (2015) The digital matatu project: using cell phones to create an open source data for Nairobi's semi-formal bus system. J Transp Geogr 49:39–51
71. Yamamoto T, Komori R (2010) Mode choice analysis with imprecise location information. Transportation 37:491–503
72. Yap MD, Nijënstein S, van Oort N (2018) Improving predictions of public transport usage during disturbances based on smart card data. Transp Policy 61:84–95
73. Yin RK (1994) Discovering the future of the case study. Method in evaluation research. Eval Pract 15:283–290
74. Yue Y, Lan T, Yeh AG, Li Q-Q (2014) Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. Travel Behav Soc 1:69–78
75. Zhang Y, Martens K, Long Y (2018) Revealing group travel behavior patterns with public transit smart card data. Travel Behav Soc 10:42–52
76. Zhao J, Rahbee A, Wilson NH (2007) Estimating a rail passenger trip origin–destination matrix using automatic data collection systems. Comput Aided Civ Infrastruct Eng 22:376–387
77. Zhao Z, Koutsopoulos HN, Zhao J (2018) Individual mobility prediction using transit smart card data. Transp Res Part C Emerg Technol 89:19–34
78. Zhou J, Zhang M, Zhu P (2019) The equity and spatial implications of transit fare. Transp Res Part A Policy Pract 121:309–324
79. Zhou Y, Yao L, Chen Y, Gong Y, Lai J (2017) Bus arrival time calculation model based on smart card data. Transp Res Part C Emerg Technol 74:81–96
80. Zhu Y, Koutsopoulos HN, Wilson NH (2017) Inferring left behind passengers in congested metro systems from automated data. Transp Res Proc 23:362–379

Khatun E Zannat is a PhD student at the Institute for Transport Studies (ITS), at the University of Leeds in the UK. She completed the Bachelor of Urban and Regional Planning degree from Bangladesh University of Engineering and Technology (BUET) followed by a Master's degree in International Cooperation in Urban Development, as an Erasmus Mundus scholar, from Technische Universität Darmstadt (TUD). Zannat has been working as a faculty member in the Department of Urban and Regional Planning of Chittagong University of Engineering and Technology (CUET) since 2013 (currently on study leave).

Charisma Choudhury is an Associate Professor at the Institute for Transport Studies and School of Civil Engineering at the University of Leeds (UoL) where she leads the Choice Modelling Research Group. She also serves as the Deputy-Director of the interdisciplinary Choice Modelling Centre, UoL. Charisma is an Honorary Guest Professor at Beijing Jiaotong University, China and a Fellow of the Alan Turing Institute - UK's national institute for data science and artificial intelligence.