International Journal of Transportation Science and Technology 12 (2023) 315–328

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

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

Origin-destination inference in public transportation systems: A comprehensive review

Mohammed Mohammed *, Jimi Oke

Department of Civil and Environmental Engineering, University of Massachusetts Amherst, MA 01003, USA

ARTICLE INFO

Article history:
Received 17 September 2021
Received in revised form 28 January 2022
Accepted 4 March 2022
Available online 23 March 2022

Keywords:
Transit
Origin-destination inference
Trip chaining
Automated data collection systems

ABSTRACT

Origin-destination (OD) modeling facilitates effective demand-responsive public transportation planning in order to meet emergent needs. Given recent advances in transit information and personal communications technology, transit OD estimation methods have evolved from relying on limited survey sources to automated big data sources. Innovative modeling approaches have also been developed over several decades to estimate trip ODs, not only for single routes, but also for full networks, including transfers. In this paper, we synthesize a review of the state of the art in research and practice, along with descriptions of key data types and methodological approaches, indicating how they interact. We also discuss current research gaps and opportunities for further innovation. This review provides a comprehensive resource that should facilitate the application of these methods to various transit systems, thus enabling planners and policymakers to gain insights from new and improved model estimates in various transit systems.

1. Introduction

Mass transit is a critical component of sustainable mobility, as it is energy-efficient and has a lower emissions footprint than other modes of transportation. With increased usage, mass transit can potentially reduce dependence on private vehicles and thus mitigate their environmental impacts (Deakin, 2001; Beaudoin et al., 2015). Transit adoption can be encouraged by effective planning and improved integration with active mobility and on-demand modes, which could further reduce travel and waiting times, thus making the service more attractive to passengers (Peterson, 2007; Miller et al., 2016). Origin-destination (OD) models are essential for such planning, as they provide spatio-temporal trip estimates which enable planners to understand which modes people use, how long they travel, and where they go. This leads to effective management and resource allocation and, ultimately, more efficient transit systems. We can also expect an improved overall quality of life, reduced greenhouse gas emissions, and lower energy consumption, among other economic benefits (Henke et al., 2020). Together, these outcomes can contribute to widespread adoption of mass transit, thus leading to sustainable mobility (Miller et al., 2016).

Since the late 1970s, various approaches and modeling frameworks have been developed to estimate OD flows in transportation networks (Wilson, 1970; Low, 1972; Robillard, 1975; Willumsen, 1978; Ben-Akiva et al., 1985; Ben-Akiva, 1987). In all cases, ground-truth data were obtained from surveys or other manually collected samples, which are laborious and

* Corresponding author at: Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Rd, Amherst, MA 01003, USA.
E-mail address: mamohammed@umass.edu (M. Mohammed).

https://doi.org/10.1016/j.ijtst.2022.03.002

expensive to assemble. Recently, the advent of automated data collection systems (ADCS), along with modern mobile communications, has generated big datasets at various resolutions that now make it possible to estimate public transportation flows with greater accuracy. Instead of relying on small survey samples with low spatial or temporal resolution, modern OD estimation models can utilize big data sources to help planners understand mobility patterns. Beyond route-level flows, innovative approaches have been developed to obtain complete network estimates, which include transfer information between routes.

Given the importance of OD modeling for public transportation networks, there remains a dearth of comprehensive sources that describe the linkage between available data sources and state-of-the-art methods. In this paper, we contribute to filling this gap by providing a synthesis of the data and methodological approaches to OD estimation for transit systems. We survey recent technologies and developments, as well as highlight opportunities for future research and applications. We expect that this review will be valuable to researchers and practitioners, both as a reference and as a guide to further investigations of case studies or methods of interest.

The rest of the paper is organized as follows. Section 2 provides information about the data sources required for transit OD modeling. In Section 3, we discuss the major methodological approaches. Section 4 provides an overview of notable recent developments in various aspects of transit OD estimation. We follow this with a discussion of current knowledge gaps and future research opportunities in Section 5, and then conclude in Section 6. We summarize the acronyms used throughout this paper in Table 1.

2. Data sources

We consider two types of origin-destination matrices: route-level OD matrices and network-level OD matrices. Route-level OD matrices contain the passenger flows from their origin stops to their destination stops along the same route. These are useful for various planning decisions, such as extending routes or building new ones (Ji et al., 2015). Network-level OD matrices contain the passenger flows from their origin stops to their destination stops, including transfer stops across multiple routes. Passengers may transfer one or more times between these stops across routes. These matrices have other benefits such as improving service planning and transit policies as a result of understanding the true origins and destinations of the network users.

Origin-destination matrices are the primary input for mass transit planning (Wong et al., 2005). In the past, however, OD matrices were only calculated when a passenger survey was being conducted. Furthermore, surveys were infrequent, due to their cost, and not always relevant to OD estimation needs (Ben-Akiva and Morikawa, 1989). Thus, up-to-date OD matrices were not usually readily available (Cui, 2006).

In recent years, there have been rapid advancements in information and communication technologies, which have given rise to advanced data collection methods. Data sources, such as automatic fare collection (AFC) systems used specifically for

Table 1
List of acronyms used in this paper.

Acronym	Meaning
ABM	Activity Based Model
ADCS	Automated Data Collection Systems
AFC	Automatic Fare Collection
APC	Automatic Passenger Counter
AVL	Automatic Vehicle Location
BE	Bayesian Estimation
BODS	Bus Origin and Destination Survey
CDR	Call Data Records
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FSDP	Fast Search by Density Peak
FSM	Four Step Model
GHG	Greenhouse Gases
GIS	Geographic Information System
GLS	Generalized Least Squares
GM	Gravity Model
GPS	Global Positioning System
GSM	Global System for Mobile communications
GTFS	General Transit Feed Specification
IPF	Iterative Proportional Fitting
MBTA	Massachusetts Bay Transportation Authority
ME	Maximum Entropy
MLE	Maximum Likelihood Estimation
OD	Origin–Destination
OPTICS	Ordering Points To Identify the Clustering Structure
PVTA	Pioneer Valley Transit Authority
TFL	Transport for London

mass transit systems, and other sources such as mobile phone data (Zannat and Choudhury, 2019) have become almost ubiquitous. Such systems can continuously collect, store, and disseminate high resolution big data, which can be exploited to analyze mobility patterns at much lower costs and in greater detail than was previously possible (Cui, 2006; Nassir et al., 2011). Moreover, these systems have the advantage of real-time or near real-time data collection. Data sources with these innovative characteristics have opened up possibilities for dynamic research and improved transit planning (Zannat and Choudhury, 2019). We describe them in detail below.

2.1. Automated data collection systems

In recent years, the abundance of raw and detailed data, which are collected by automated data collection systems (ADCS), attracted the attention of researchers and inspired them to create various origin-destination estimation approaches that use what is called the trip chaining method. This is because ADC systems proved to be very rich and detailed data sources that are both realistic and up to date. This subsequently led to the development of new methods of inferring the full-network origin-destination matrix referred to by some researchers as *Origin-Destination-Transfer* (ODX) models (Vanderwaart, 2016; Sánchez-Martínez, 2017). By reducing costs of data collection, utilizing larger sample sizes, and automating the process of these new methods, up-to-date OD matrices can be expected to become available on a much more frequent basis (Cui, 2006).

The three major ADCS are: automatic vehicle location (AVL), automatic passenger counter (APC), and automatic fare collection (AFC) systems. When all these sources are combined, they provide more accurate and reliable information as they cover different aspects of the transit network service. There are, however, some challenges involved. It is usually not intended for these systems to be used together to provide integrated data, as their purpose is not to generate OD matrices. Thus, to exploit them for OD estimation, these sources require further processing (Zhao, 2004).

2.1.1. Fare collection

Around the world, transit systems have become increasingly reliant on automated fare collection (AFC) systems (Huang et al., 2020). AFC systems vary in type from smart card to smartphone-based systems. A variety of smart fare card systems are used by many cities, such as London's Oyster card, New York's Smart Link, Boston's Charlie card, Beijing's Yika tong, and Hong Kong's Octopus card (Zannat and Choudhury, 2019). In these systems, passengers indicate their boarding or alighting activity via smart card tap-in or smartphone activation, among others, depending on the type of network and the transit agencies' policies. Individual passenger information is then collected by the AFC system and stored by the transit operator (Cui, 2006). AFC systems can be further categorized as:

- *Open*: Only passenger boardings are recorded; common in areas where fare prices are fixed.
- *Closed*: Both boardings and alightings are indicated; typical where fares are distance-based.

Wu et al. (2021) provide a thorough review of the use of smart card data for OD estimation.

2.1.2. Vehicle location

The location of vehicles in a network is important for accurate OD estimates. However, the AFC systems in many bus networks do not collect spatial information along with payment data. Automatic vehicle location (AVL) systems address this problem by utilizing GPS data to supplement AFC systems. GPS devices facilitate the collection of payment locations, in addition to the time recorded by the AFC (Cui, 2006; Zannat and Choudhury, 2019). However, some issues might arise when combining the two data sources. For instance, GPS inaccuracies might lead to incorrect boarding or alighting stop inference. If AVL data are not available, or if the data are highly inaccurate, vehicle locations can be inferred from alternate sources, such as transit schedules (Cui, 2006).

2.1.3. Passenger counts

Automatic passenger counter (APC) systems provide spatial and temporal boarding and alighting counts (Vanderwaart, 2016; Cui, 2006). The data are typically collected from sensors installed near vehicle doors, which record passengers entering and exiting. APC data can also be used to infer the number of passengers onboard a vehicle at a given time.

2.2. Mobile data

As cellular devices have become a very important aspect of most individuals' lives, large amounts of data are collected by the telecommunication companies from these devices. Call data records (CDR), which can provide the location of the nearest cell towers that serve a given user, and global system for mobile communication (GSM) data, which are always being collected as long as the device is operational, can be leveraged for mass transit planning. While the information they provide may sometimes be limited to time and location, the size of these datasets allows for the application of data mining algorithms to infer mobility patterns (Zannat and Choudhury, 2019).

3. Key methods

Several methods have emerged for OD estimation using big data from ADCS, as a result of its increasing availability. These approaches can be categorized as traffic modeling-based and statistical inference-based (Wong et al., 2005). Various methods are used for computing the trip origins, such as spatial clustering based on GPS, time-stamped boardings, bus schedules, etc. Others are used to estimate trip destinations via trip chaining approaches, such as assumption-based heuristic methods, neural networks, etc. These result in seed OD matrices, from which route-level and network-level solutions can be obtained. Transfer inference is required for network-level OD estimates. We describe the key model frameworks for transit OD estimation in the following subsections. Fig. 1 indicates the inputs (data sources), models, outputs, and purposes of OD estimates in a transit setting. We summarize the notation used throughout this section in Table 2. Notable research efforts in transit OD estimation are summarized in Table 3.

3.1. General full-network OD matrix estimation approach

A complete estimation of the trip distribution in a given public transportation network requires inference on trip origins, destinations, and transfers. The inference of trip origins and destinations is typically performed simultaneously. Transfer inference then identifies the destination stops that serve as transfers to another vehicle for a given passenger. These methods provide a highly spatio-temporally resolved analysis of network usage than is possible with previous data collection methods. This offers a complete and reliable picture of riders' activity and route choices that does not rely on assumptions. It can also provide planners with specific information about where to target their available resources and make adjustments to achieve maximum efficiency of the transit network (Vanderwaart, 2016). A general approach for network-level OD estimation involves the following steps (Cui, 2006):

- Obtain a sample of ridership counts (from APC, AFC, etc.), a seed matrix (from AFC, mobile network data, etc.), and transfer flows (from AFC, etc.).
- Using the boarding and alighting totals for each route and the seed matrix, estimate the route-level OD matrices with methods such as the iterative proportional fitting (IPF).
- Infer route transfer probabilities from AFC data and use these to estimate the network-level OD matrix.

3.2. Iterative proportional fitting

Iterative proportional fitting (IPF), which is sometimes referred to as bi-proportional fitting, is usually used to estimate a route-level OD matrix from a seed (base) matrix and the total boarding and alighting counts (marginal totals) (Cui, 2006). A key advantage of IPF over other matrix estimation methods is that it can be computed easily without sacrificing accuracy (Ben-Akiva et al., 1985). It is also considered the state-of-the-art method for estimating OD flows from boarding/ alighting data (Ji et al., 2014).

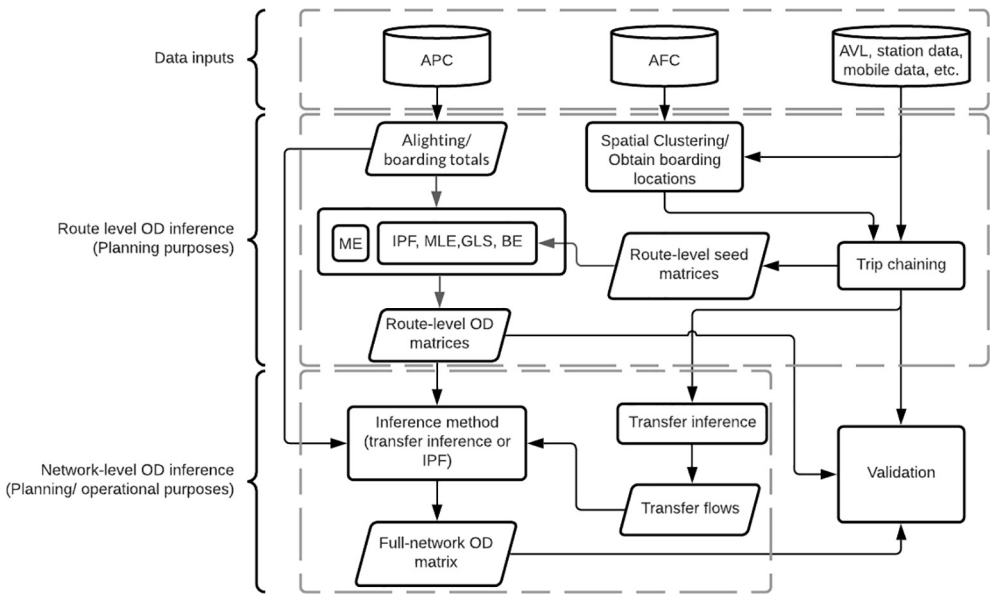

Fig. 1. Flowchart of full-network OD estimation methods, required data sources, and purpose of output.

Table 2
List of mathematical symbols used in this paper.

Symbol	Definition
a_i, b_j	Lagrangian multipliers
$\lvert C_k \rvert$	number of points in the *kth* cluster, C_k
D_j	total trips to destination station j
i	origin station index
j	destination station index
K	number of clusters
L	likelihood function
O_i	total trips from origin station i
Q_i, Q_j	row and column marginal totals
$\boldsymbol{X}$	seed matrix
x_{ij}	number of trips from i to j
x_{ij}^0	number of trips from i to j in the seed matrix
$\hat{x}_{ij}$	estimated number of trips from i to j
$\bar{x}_{kn}$	mean for feature n in the *kth* cluster
x_n	n-th point in cluster C_k

The IPF algorithm is described as follows:

1. Compute marginal totals Q_i and Q_j for boardings and alightings, respectively, from available data sources.
2. Obtain a seed matrix X from smart card or other AFC data:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{3.1}$$

3. Using the marginal totals and the seed matrix, for each iteration k:
   (a) Calculate the scaled updates by row:

$$x_{ij}^{k+1} = \frac{x_{ij}^k}{\sum_j x_{ij}^k} Q_i \tag{3.2}$$

   (b) Then calculate the scaled updates by column:

$$x_{ij}^{k+2} = \frac{x_{ij}^{k+1}}{\sum_i x_{ij}^{k+1}} Q_j \tag{3.3}$$

   where x_{ij}^k is the number of trips between origin stop i and destination stop j (OD pair) in the kth iteration, and Q_i and Q_j are the row and column marginal totals, respectively.
4. Repeat above steps until $\sum_j \hat{x}_{ij}$ is sufficiently close to Q_i and $\sum_i \hat{x}_{ij}$ is sufficiently close to Q_j, where $\hat{x}_{ij}$ is the final estimated number of trips from i to j.

A limitation of IPF is the non-structural zeros problem (Ben-Akiva et al., 1985). It arises when OD pairs with low flow in the seed matrix are estimated as zero flow cells in the solution. Another issue is that the quality of the seed matrix can significantly affect the IPF solution accuracy (Liu et al., 2021). When data are unavailable to generate a seed matrix, a null seed matrix is typically used. As a result, however, trip estimates may be inflated. To address this issue, Ben-Akiva et al. (1985) proposed three approaches: resampling, zero-cell augmentation and Bayesian smoothing. More recently, Ji et al. (2014) proposed an iteratively resampled approach, coined "IPF method with an iteratively improved base" (IPF-IB).

3.3. Maximum entropy

In the 1970s, researchers took advantage of data generated from traffic monitoring systems to develop indirect mathematical models. They employed the maximum entropy (ME) principle (equivalent to minimum information) to obtain the most likely trip matrix estimation based on traffic counts (Willumsen, 1978; Van Zuylen and Willumsen, 1980; Willumsen, 1981). The underlying hypotheses of these approaches are: first, there is a large number of possible trip distributions, and, second, that the optimal OD matrix estimate is the one that maximizes the total entropy (randomness) of the

Table 3
Relevant origin–destination estimation research efforts, highlighting key methods, data types and study areas.

Researchers	Method	Data	Mode	Study Area/City	Dynamic	Purpose	Level
Alsger et al. (2015)	Trip chaining	Smart card	Multimodal	Queensland, Australia	N	Planning	Route
Assemi et al. (2020)	Trip chaining (neural net)	Smart card	Multimodal	Queensland, Australia	N	Planning	Route
Ait-Ali and Eliasson (2019)	Maximum Entropy	Smart card station entry	Train	Stockholm, Sweden	Y	Planning	Network
Cheng et al. (2019)	Trip chaining, topic model	Smart card	Train	Guangzhou, China	N	Planning	Network
Cui (2006)	Trip chaining, IPF, MLE	AFC, APC, AVL	Bus	Chicago, USA	N	Planning	Network
Fan and Chen (2018)	Trip chaining	Smart card	Bus	Guangzhou, China	N	Planning	Route
Farzin (2008)	Trip chaining	AVL, AFC	Bus	São Paulo, Brazil	N	Planning	Route
Ge and Fukuda (2016)	Maximum Entropy	Mobility traces, survey	Train	Tokyo, Japan	N	Planning	Network
Gordon (2012)	Trip chaining, IPF	AFC, AVL	Bus/Train	London, UK	N	Planning	Network
Hazelton (2010)	Bayesian Method	Stop level counts	Bus	San Francisco, USA	N	Planning	Route
He et al. (2015)	Trip chaining	Smart card	Bus	Brisbane, Australia	N	Planning	Route
Hora et al. (2017)	Trip chaining	AFC	Bus/Train	Porto, Portugal	N	Planning	Network
Huang et al. (2020)	DBSCAN, Trip chaining	AFC, AVL	Bus	Suzhou, China	N	Planning	Route
Kang et al. (2020)	Trip chaining, expansion factors	Smart card, GPS, GIS	Bus	Tehran, Iran	Y	Planning	Network
Kumar (2019)	Trip chaining	AFC, AVL, GTFS	Bus/Train	Minneapolis-Saint Paul, USA	N	Planning	Route
Lam et al. (2003)	Generalized least squares	Observed counts, hist. OD matrices	Bus/Train	Kowloon, China	N	Planning	Route
Li et al. (2011)	Trip chaining	APC	Bus	Jinan, China	N	Planning	Route
Liu et al. (2021)	IPF, Integer programming	AVL, APC	Bus	Ann Arbor-Ypsilanti, USA	N	Planning	Network
Liu et al. (2020)	Trip chaining	Smart card, GPS	Bus	Shenzhen, China	N	Planning	Network
Luo et al. (2017)	Transfer inference, K-means	Closed system AFC	Bus/Train	Haaglanden, Netherlands	N	Planning	Network
Munizaga and Palma (2012)	Trip chaining, expansion factors	Smart card, GPS	Bus/Train	Santiago, Chile	N	Planning	Network
Nassir et al. (2011)	Trip chaining	AFC, AVL, APC, GTFS	Bus	Minneapolis–Saint Paul, USA	N	Planning	Network
Navick and Furth (1994)	IPF, MLE	Surveys, stop level counts	Bus	Boston, USA	N	Planning	Route
Nunes et al. (2016)	Trip chaining	AFC		Porto, Portugal	N	Planning	Route
Sánchez-Martínez (2017)	Trip chaining via dynamic programming	AFC, AVL	Bus/Train	Boston, USA	N	Operational	Network
Trépanier et al. (2007)	Trip chaining	Smart Card AFC	Bus	Gatineau, Canada	N	Planning	Route
Wang (2010)	Trip chaining, expansion factors	AFC, AVL, APC, GIS	Bus	London, UK	N	Planning	Route
Widyawan (2017)	Trip chaining	AFC	Bus	Jakarta, Indonesia	N	Planning	Route
Wong et al. (2005)	Maximum Entropy	Traffic counts	Multimodal	Hong Kong	N	Planning	Network
Wu et al. (2021)	Transfer inference, MLE	AFC in a closed system	Multimodal	Seoul, South Korea	N	Planning	Network
Zeng et al. (2014)	Trip chaining	AVL, AFC, GTFS	Bus/Train	New York City, USA	Y	Operational	Network
Zhao et al. (2007)	Trip chaining, IPF	AFC, APC, AVL	Train	Chicago, USA	N	Planning	Route

system (Ait-Ali and Eliasson, 2019). Following Wilson (2011), the method has gained importance as a modeling and planning tool. The ME method is viable for estimating flows because of its generalizability and ability to use any information from the observed flows (Willumsen, 1981).

The objective function in ME is given as (Xie et al., 2010; Ait-Ali and Eliasson, 2019):

$$\max_{x_{ij}} \sum_{ij} \left(x_{ij} \log(x_{ij}) - x_{ij}\right) \tag{3.4}$$

where i is the origin station index, j is the destination station index, and x_{ij} is the number of trips from i to j. Objective (3.4) can be subject to several constraints, e.g.:

$$\sum_{j} (x_{ij}) = Q_i \tag{3.5}$$

This constrained model can be solved by formulating the Lagrangian to incorporate the constraints into the objective function (augmented Lagrangian method) to find the OD estimates ($\hat{x}_{ij}$). In this basic solution, the OD estimates can be calculated by:

$$x_{ij} = \frac{Q_i}{|D|} \tag{3.6}$$

where $|D|$ is the number of possible destinations from the origin station. This is a basic solution since it assumes all destinations have the same level of attraction (Ait-Ali and Eliasson, 2019).

3.4. Maximum likelihood estimation

In the maximum likelihood estimation (MLE) approach, we assume that OD pairs in the seed matrix are independent (a simplifying but not necessarily true assumption) and identically distributed. Each OD pair can thus be considered a random selection from a hypothetical joint distribution, with the resulting likelihood of all ODs obtained as the joint probability of the OD pair observations. The estimated OD matrix is then obtained by maximizing the likelihood (Ben-Akiva et al., 1985; Lu, 2008).

Assuming a Poisson distribution for the OD pairs, the probability of observing the number of passengers in any OD pair is given by:

$$P(\hat{x}_{ij} = x_{ij}^0) = \left(\frac{\hat{x}_{ij}^{x_{ij}^0}}{x_{ij}^{0!}} \exp(-\hat{x}_{ij}) \right) \tag{3.7}$$

where $\hat{x}_{ij}$ is the estimated number of trips from i to j and x_{ij}^0 is the number of trips from i to j in the seed matrix. The joint likelihood function, assuming independence, can then be formulated as:

$$L = \prod_i \prod_j \left(\frac{\hat{x}_{ij}^{x_{ij}^0}}{x_{ij}^{0!}} \exp(-\hat{x}_{ij}) \right) \tag{3.8}$$

The likelihood function can be subjected to constraints for feasibility, e.g.:

$$\sum_j \hat{x}_{ij} = O_i \tag{3.9}$$

$$\sum_i \hat{x}_{ij} = D_j \tag{3.10}$$

where O_i is the total trips from origin station i and D_j is the total trips to destination station j.

Maximizing the likelihood function under the constraints (3.9) and (3.10) can be done via the augmented Lagrangian method as follows:

$$\ln L^0 = \sum_i \sum_j [-\hat{x}_{ij} + x_{ij}^0 \ln(\hat{x}_{ij}) - \ln(x_{ij}^{0!})] - a_i(\sum_j \hat{x}_{ij} - O_i) - b_j(\sum_i \hat{x}_{ij} - D_j) \tag{3.11}$$

The resulting maximum likelihood estimator can then be written as:

$$\hat{x}_{ij} = \frac{x_{ij}^0}{1 + a_i + b_j} \tag{3.12}$$

where $\hat{x}_{ij}$ is the estimated trips between i and j, a_i and b_j are the Lagrangian multipliers. As can be seen from the constrained likelihood function solution in Eq. (3.12), the estimated OD flows are calculated as the seed OD trips divided by the summation of these balancing factors.

Aerde et al. (2003) proposed a numerical solution to the maximum likelihood synthetic OD problem that can overcome some limitations of the traditional formulations, such as the computational difficulty, by using only Stirling's approximation. Their model is highly efficient and can be implemented on routes without a seed matrix.

3.5. Constrained generalized least squares

The estimation of OD flows $\hat{x}_{ij}$ can also be formulated as a constrained generalized least squares (CGLS) problem. CGLS was formally described by Theil (1971) and then demonstrated as viable for transit OD estimation by Hendrickson and McNeil (1984); McNeil and Hendrickson (1985); Ben-Akiva et al. (1985). The approach can be justified by assuming that an expansion factor f applied to the seed matrix elements x_{ij}^0 provides an unbiased estimate of the true flows x_{ij}. The expansion factor f is obtained as the ratio of the total number of trips (the sum of all the boardings, for instance) and the total trips from the seed matrix $\boldsymbol{X}$. The CGLS problem can then be written as:

$$\min (f\boldsymbol{x}^0 - \boldsymbol{x})^T V^{-1} (f\boldsymbol{x}^0 - \boldsymbol{x}), \tag{3.13}$$

where V is the $K \times K$ variance-covariance matrix of $\boldsymbol{x}^0$, and constrained as follows:

$$R\boldsymbol{x} = r \tag{3.14}$$

where R is the $C \times K$ constraint-incidence matrix and r is the $C \times 1$ vector of linearly independent row and column constraints (corresponding to total boardings and alightings). In this formulation, $\boldsymbol{x}$ is expressed as a vector of K origin-destination flows. Thus, in a system with $|O|$ possible origins and $|D|$ possible destinations, $C = |O| + |D|$ and $K = |O| \cdot |D|$. The estimator can then be written as:

$$\hat{\boldsymbol{x}} = f\boldsymbol{x}^0 + VR^T (RVR^T)^{-1} (r - R\boldsymbol{x}^0) \tag{3.15}$$

Bell (1991) later presented an algorithm to solve the GLS problem subject to inequality constraints. In a notable extension, Xie et al. (2011) proposed a combined ME-LS method that improved accuracy by incorporating elastic, instead of fixed, demand.

3.6. Bayesian estimation

Bayes' Theorem gives the posterior distribution of unknown parameters $\boldsymbol{\theta}$ given an observed measurement $\boldsymbol{Y}$ as proportional to the likelihood of the observation and the prior probability of the unknowns $p(\boldsymbol{\theta})$. This is typically expressed as

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3.16}$$

In the case of the transit OD estimation problem, the unknowns are the OD flows between stops and are taken as conditioned on a given observation, such as trip counts/totals, or boarding/alighting counts. The estimates can then be obtained as those giving the maximum a posteriori (MAP) density. We note that some of the methods earlier discussed are either special cases of Bayesian estimation (BE) or share a correspondence. For instance, MLE corresponds to the MAP estimate where a uniform prior is assumed. Furthermore, it has been shown that the method of maximum entropy is a generalization of Bayesian inference (Giffin and Caticha, 2007).

In BE approaches, closed-form estimates of the likelihood are rarely possible, however, given the size of the problem. Thus, sampling approaches, such as Metropolis-Hastings (Hastings, 1970) and Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2011) are typically used to obtain the estimates and their uncertainties. Bayesian methods have been demonstrated for route-level estimation (Ji et al., 2015), as well as for full-network estimation (Park et al., 2008; Blume et al., 2021). Several challenges perhaps contribute toward the increased application of BE approaches to transit OD estimation. A well-documented issue is that of exhaustive OD enumeration required to evaluate the likelihood function. However, this can be addressed via sampling methods such as MCMC or via the expectation maximization algorithm (Dempster et al., 1977; Meng and Dyk, 1997; Ji et al., 2015). Another factor, depending on the formulation of the problem, is the requirement of a complete observation of OD flows in a given snapshot. These data are not always available, however. Addressing this challenge requires innovations in model formulation and sampling approaches. Blume et al. (2021) provide a recent example of this.

3.7. Trip chaining

Trip chaining is the process of inferring the alighting stop location of each trip that passengers take. This is necessary in the absence of alighting data for individual passengers from the AFC data, as is usually the case. It is therefore an essential step in obtaining a route-level seed matrix. In addition, trip chaining is used to infer transfers which helps in assigning trip legs in order to complete a passenger's entire trip sequence (journey), and thus integral to building a full-network origin-destination model.

Trip chaining is commonly implemented as a rule-based heuristic. The algorithm takes individual boarding locations obtained from AFC and AVL or bus schedule data and creates buffer zones around each boarding location to infer the alighting location that preceded it. Other rules/assumptions specific to the transit system area are also applied (Cui, 2006). Using the AFC data, transfers can either be inferred based on probabilistic models or by using deterministic criteria within the trip chaining algorithm. This ultimately results in transfer probabilities that can be used to assign transfer flows across routes in order to obtain the full-network result.

The potential use of automatic data collection systems (ADCS) to estimate trip chains was investigated by Wang (2010), who analyzed bus passenger travel behavior in the Transport for London (TFL) network. He predicted boardings and alightings and examined transfer patterns using the trip chaining principles employed in Chicago (Cui, 2006). He was also the first to validate his results with manually administered surveys (the Bus Origin and Destination Survey data (BODS)). This work showed the ease of applying the trip chaining method to ADCS data and inferring a network-level OD matrix. An enhanced method to infer complete passenger journeys in a mass transit system was subsequently developed by Gordon (2012). This effort provided a solution to the issue of non-representation of passengers who did not use farecards via a modified IPF method. He also demonstrated that this process could be applied across the entire network and on a daily basis, thus eliminating the need for small and infrequent samples.

The typical criteria for trip chaining and transfer inference include the following general assumptions (Alsger et al., 2016):

- *Transfer time threshold*: This value ranges from 30 min to 90 min.
- *Walking distance threshold*: Similarly, walking distances can range from 400 m to 1100 m, depending on the area and the network.
- *Last destination*: There are two assumptions for inferring the final alighting location of the day. One approach is to assume that the final alighting location is the same as the first boarding location of that day. The alternative assumption is that the final alighting location is the closest stop to the first boarding location on the final trip's route.

3.7.1. Validation of trip chaining assumptions and results

The historical trip chaining assumptions that have guided modeling efforts in recent decades have not been validated in many cases, or at least not completely. Revisiting and re-validating these assumptions can be challenging, however, especially in the absence of high resolution data. Nonetheless, when the major trends that emerged in the past several years such as ride-hailing and micro-mobility services are considered, the need for validation is unavoidable (Shaheen et al., 2020; Palm et al., 2021).

Big data sources could provide avenues for integrating relevant variables into novel datasets that could enhance validation efforts (Zannat and Choudhury, 2019). Recently, some researchers attempted to assess some of the trip chaining assumptions using AFC smart card data in Queensland, Australia (Alsger et al., 2015; Alsger et al., 2016). Findings indicated that: the assumed allowable transfer time does not have a significant impact on the OD matrix estimates; most passengers (90%) are willing to walk less than 10 min for transfers; and most passengers return to their first origin of the day. This effort also yielded an algorithm to generate OD matrices based on individual user transactions.

3.8. Spatial clustering

A problem commonly faced when trying to estimate an OD matrix in a system that has no boarding stop information is origin inference. Spatial data from the AFC system need to be assigned to the correct stop location. This often requires some spatial clustering algorithm to obtain the most accurate assignment. Clustering techniques such as K-means clustering, density-based spatial clustering (DBSCAN), hierarchical clustering, ordering points to identify the clustering structure (OPTICS), and fast search by density peak (FSDP) clustering (Schoier et al., 2017), can be applied to match AFC data to bus stop locations (Huang et al., 2020). We discuss the K-means and DBSCAN methods in detail in the following subsections.

3.8.1. K-means clustering

The K-means algorithm (MacQueen, 1967) classifies points or observations into K clusters by minimizing the sum of squared (Euclidean) distances between the points in each cluster. The method requires the prior input of the number of cluster centers (K).

The objective function of the K-means algorithm can be formulated as:

$$\min \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{n=1}^{|C_k|} (x_{kn} - \bar{x}_{kn})^2 \tag{3.17}$$

where K is the number of clusters, $|C_k|$ is the number of points in cluster C_k, x_{kn} is the *nth* point in cluster C_k, and $\bar{x}_{kn}$ is the mean for feature n in the cluster C_k. In the stops inference application, the points to be clustered are the locations reported by the AFC system and K would be the number of stations in the network. The resulting clusters should correspond to the location of the stations (Hartigan and Wong, 1979). The location of each centroid might not exactly match the locations reported by AFC. In that case, each cluster can be assigned to the nearest actual stop location.

3.8.2. Density-based spatial clustering

Density-based spatial clustering (DBSCAN) (Ester et al., 1996) generates clusters by finding regions where observations are tightly packed. DBSCAN differs from K-means clustering in several aspects. It does not require the prespecification of the number of clusters as it is designed to discover clusters on its own based on point density. It also discovers and removes noise from cluster inclusion. This is achieved by specifying the minimum number of points to be considered a cluster (*minPoints*) and a distance threshold (ε). DBSCAN takes the value of ε and creates a radius around each point in the data. It then decides, based on the minimum number of points parameter, which points have *minPoints* in their radius (which translates into a density threshold). The groups that satisfy this density threshold are then considered clusters. Any points that do not fall inside these dense areas, as defined by the algorithm, are considered noise points (Sander, 1998).

Due to the limitations of having to specify single values for each of the two parameters in DBSCAN, which may lead to some clusters going undiscovered, Ankerst et al. (1999) developed the alternative OPTICS algorithm. Rather than depending on ε, OPTICS instead finds high density samples based only on *minPoints* and expands from them. This helps it to identify clusters with different densities and sizes.

4. Recent developments

In this section, we discuss recent advances in transit OD estimation, notably in the areas of trip chaining and transfer inference, real-time estimation, and real-world applications.

4.1. Innovations in alighting and transfer inference algorithms

In many cases, the boarding stop location is not available due to limitations in fare collection systems. An algorithm to infer the boarding stop by identifying the boarding direction, without relying on GPS or survey data, was developed by Chen and Fan (2018) in Guangzhou, China. Other researchers applied machine learning approaches (neural networks) for more data-driven inference (Assemi et al., 2020) and optimization (Liu et al., 2021), both of which do not rely on heuristic assumptions to estimate trip chains. Hussain et al. (2021) provides a more critical analysis of the recent developments in the area of transfer inference and OD estimation using smart card data.

4.2. Real-time estimation

Static OD estimation focuses on finding the total trips between OD pairs for a certain period of time (Ait-Ali and Eliasson, 2019). Conversely, real-time (dynamic) OD estimation focuses on transit operations and management by adding a time dimension to its estimation (Cho et al., 2009). The real-time requirements of this type of OD estimation, such as the limited computation time available, make it very challenging (Peterson, 2007).

Some algorithms were introduced to solve the challenges of real-time estimation, such as the Kalman filter (Kalman, 1960) and LSQR—an iterative least squares approach via QR factorization (Paige and Saunders, 1982). These two algorithms were compared by Bierlaire and Crittin (2004) who adopted a least-squares formulation to reduce the computation time for large scale problems in the U.S. cities of Boston, Massachusetts and Irvine, California.

Other algorithms include: entropy maximization, which Ait-Ali and Eliasson (2019) used to solve the dynamic OD estimation problem when the available data are only station entry counts; the Local Ensemble Transformed Kalman Filter used to easily explain the nonlinearity between time-dependent OD flows and traffic data (Castiglione et al., 2021); and the ‘dynamic transfer link’ network generation method used for automating large-scale planning (Zeng et al., 2014). The latter proved suitable for high-resolution modeling on a daily basis and was used to infer passengers’ full journeys in New York City via online calibration of shortest-path models.

4.3. Real-world applications

Rather than improving the trip chaining algorithm, some researchers focused on the applications of network-level OD estimates. Vanderwaart et al. (2017) used an OD dataset from Boston provided by the MBTA from 2014 to identify service needs in key locations and suggest new planning techniques. This work demonstrated how OD modeling results can be used to obtain accessibility and equity outcomes of service changes. Planners can gain valuable information through these approaches and use them for better decision-making.

Sánchez-Martínez (2017) developed a dynamic programming model based on a generalized disutility minimization objective to estimate the path that a passenger will most likely take. This model, contrary to earlier ones, takes into account the effect of the number of transfers on path choice and the time spent waiting for or inside vehicles. It is currently being used to infer OD matrices in Boston, Massachusetts, with better results than previous efforts.

5. Opportunities for further research

There is a pressing need to develop full-network transit OD models that are readily deployable to various cities and regions worldwide. To facilitate these developments, further investigations that harness the latest statistical and machine learning innovations using big data on operational and passenger behavior are required. In an age where disruptive events are increasingly imminent, insights gained in these areas can improve planning efforts. We highlight these research avenues below.

5.1. Integration of new technologies

The rise of big data and advanced mobile technologies has yielded opportunities to exploit new sources of information (e.g. mobility traces, social networks) for potentially more efficient and accurate models. Ongoing developments in machine learning and network science are well poised to harness these developments to enhance the inferential capabilities of OD models. There is also an opportunity for innovation in survey methods via mobile applications and social media, in order to more readily obtain validation data. Furthermore, novel methods for survey data integration with ADCS can be developed for more robust and explainable OD estimates.

Fig. 2. World map indicating the cities that have served as study areas for origin–destination modeling research efforts between 1994 and 2021.

5.2. Incorporation of user behavior

The methods used for OD estimation in practice also often do not take in consideration user preferences for routes and transfers. Different demographics with differing behaviors can affect the accuracy of the model predictions (Assemi et al., 2020). Also, current methods assume the same usage of the transit system by all users and only scales the results obtained from the ADCS. However, passengers who do not use the AFC system could have different patterns from those who do. Models that capture heterogeneity across various groups can provide more detailed insights and analyses, and can be particularly useful for investigating equity concerns (Wu et al., 2021).

5.3. Inclusion of developing countries

Finally, the planning capabilities afforded by OD models could have significant impacts in developing or emerging economies, where transit usage is on the decline and car ownership on the rise (Stead et al., 2017). Yet, as seen in Fig. 2, there are significantly fewer studies in these countries. Future research efforts in emerging nations could harness big data available from cellphones to estimate OD models for improved planning. The potential for increasing transit ridership as a result could yield significant benefits for the environment.

5.4. Other research gaps

The following topics can also benefit from further research:

- The impact of AFC system type: The difference between tapping at boarding only, at boarding and alighting, or at alighting only can affect the assumed values and the methods that should be used for estimation.
- Use of fixed versus dynamic trip chaining assumptions: More research on using a variable value based on bus schedule, land use, station, etc., instead of a fixed value for the trip chaining assumptions is needed to increase the accuracy of inference. A non-residential area, for example, is less likely to be the last destination of the day.
- The effect of new transit services: Emerging transit services, such as ride-sharing and bike-sharing, could impact the transfer assumptions (Assemi et al., 2020), especially in networks where the accessibility to mass transit facilities is lacking. This can prevent the model from providing any meaningful full-network OD estimation. The effect of these needs to be investigated and incorporated into future models.
- Impact of disruption on transfers: A delayed transfer due to congestion, high demand, or other disruptive events, can lead to wrong inference in models where the transfer delay threshold is lower than that caused by the disruption. Disruptions could also affect the traveled distance, route choice, and type of vehicle used for the transfer which lowers the model's planning capabilities (Hussain et al., 2021).

6. Conclusion

The theory and practice of OD inference have been refined over the past few decades in order to aid planning and decision-making efforts in various transit systems. In this paper, we have given a comprehensive review of the background, required data sources, and the state-of-the-art methods for estimating OD models. Given the importance of public

transportation for sustainable urban and regional mobility, OD approaches will continue to remain critical to serving planning and demand-response needs. Thus, we expect that the modeling approaches detailed here would be useful as starting points for future researchers and practitioners to expand on current methods and gain novel insights by applying them to new transit systems worldwide.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aerde, M.V., Rakha, H., Paramahamsan, H., 2003. Estimation of Origin-Destination Matrices: Relationship Between Practical and Theoretical Considerations. Transp. Res. Record 1831.1, pp. 122–130.
Ait-Ali, A., Eliasson, J., 2019. Dynamic Origin-Destination Estimation Using Smart Card Data: An Entropy Maximisation Approach. arXiv e-prints.
Alsger, A., Assemi, B., Mesbah, M., Ferreira, L., 2016. Validating and improving public transport origin-destination estimation algorithm using smart card fare data. Transp. Res. Part C 68, 490–506.
Alsger, A.A., Mesbah, M., Ferreira, L., Safi, H., 2015. Use of Smart Card Fare Data to Estimate Public Transport Origin-Destination Matrix. Transp. Res. Record 2535 (1), 88–96.
Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: Ordering Points To Identify the Clustering Structure. ACM Sigmod Record 28 (2), 12.
Assemi, B., Alsger, A., Moghaddam, M., Hickman, M., Mesbah, M., 2020. Improving alighting stop inference accuracy in the trip chaining method using neural networks. Public Transport 12.1, pp. 89–121.
Beaudoin, J., Farzin, Y.H., Lin Lawell, C.-Y.C., Lin, C., 2015. Public transit investment and sustainable transportation: A review of studies of transit's impact on traffic congestion and air quality. Res. Transp. Econ. Sustain. Transp. 52, 15–22.
Bell, M.G.H., 1991. The estimation of origin-destination matrices by constrained generalized least squares. Transp. Res. Part B 25 (1), 13–22.
Ben-Akiva, M., 1987. Methods to combine different data sources and estimate origin-destination matrices. Transportation and traffic theory, pp. 459–481.
Ben-Akiva, M.E., Morikawa, T., 1989. Data fusion methods and their applications to origin destination trip tables. Transport Policy, Management & Technology towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research. vol. 4.
Ben-Akiva, M.E., Macke, P.P., Hsu, P.S., 1985. Alternative Methods to Estimate Route-Level Trip Tables and Expand On-Board Surveys. Transp. Res. Record. p. 11.
Bierlaire, M., Crittin, F., 2004. An Efficient Algorithm for Real-Time Estimation and Prediction of Dynamic OD Tables. Oper. Res. 52 (1), 116–127.
Blume, S.O.P., Corman, F., Sansavini, G., 2021. Bayesian Origin-Destination Estimation in Networked Transit Systems using Nodal In- and Outflow Counts. arXiv preprint 2105.12798.
Castiglione, M., Cantelmo, G., Qurashi, M., Nigro, M., Antoniou, C., 2021. Assignment Matrix Free Algorithms for On-line Estimation of Dynamic Origin-Destination Matrices. Front. Future Transp. 2.
Chen, Z., Fan, W., 2018. Extracting bus transit boarding stop information using smart card transaction data. Journal of Modern Transportation 26.3, pp. 209–219.
Cheng, Z., Trepanier, M., Sun, L., 2019. Inferring trip destinations in transit smart card data using a probabilistic topic model. CIRRELT-2019-47.
Cho, H.-J., Jou, Y.-J., Lan, C.-L., 2009. Time Dependent Origin-destination Estimation from Traffic Count without Prior Information. Networks Spatial Econ. 9 (2), 145–170.
Cui, A., 2006. Bus passenger Origin-Destination Matrix estimation using Automated Data Collection systems (Thesis). Massachusetts Institute of Technology.
Deakin, E., 2001. Sustainable Development and Sustainable Transportation: Strategies for Economic Prosperity, Environmental Quality, and Equity. University of California Transportation Center.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. R. Stat. Soc.: Ser. B (Methodological) 39 (1), 1–22.
Ester, M., Kriegel, H.-P., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. kdd 96 (34), 226–231.
Fan, W., Chen, Z., 2018. Estimation of Origin-Destination Matrix and Identification of User Activities Using Public Transit Smart Card Data. Center for Advanced Multimodal Mobility Solutions and Education.
Farzin, J.M., 2008. Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. Transp. Res. Record 2072 (1), 30–37.
Ge, Q., Fukuda, D., 2016. Updating origin-destination matrices with aggregated data of GPS traces. Transp. Res. Part C 69, 291–312.
Giffin, A., Caticha, A., 2007. Updating Probabilities with Data and Moments. AIP Conference Proceedings 954, 74–84.
Gordon, J.B., 2012. Intermodal passenger flows on London's public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data (Thesis). Massachusetts Institute of Technology.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) 28 (1), 100–108.
Hastings, W.K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57 (1), 97–109.
Hazelton, M.L., 2010. Statistical Inference for Transit System Origin-Destination Matrices. Technometrics 52 (2), 221–230.
He, L., Nassir, N., Trepanier, M., Hickman, M., 2015. Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems. CIRRELT vol. 52.
Hendrickson, C., McNeil, S., 1984. Estimation of origin-destination matrices with constrained regression. Transp. Res. Record 976.
Henke, I., Cartenı, A., Molitierno, C., Errico, A., 2020. Decision-Making in the Transport Sector: A Sustainable Evaluation Method for Road Infrastructure. Sustainability 12 (3), 764.
Hora, J., Dias, T.G., Camanho, A., Sobral, T., 2017. Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system. Transp. Res. Proc. 27, 664–671.
Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L., Gong, C., 2020. A Method for Bus OD Matrix Estimation Using Multisource Data. J. Adv. Transp., 1–13
Hussain, E., Bhaskar, A., Chung, E., 2021. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. Transp. Res. Part C 125, 103044.

Jafari Kang, M., Ataeian, S., Amiripour, S.M.M., 2020. A procedure for public transit OD matrix generation using smart card transaction data. Public Transport 13(1), 81–100.
Ji, Y., Mishalani, R.G., McCord, M.R., 2014. Estimating Transit Route OD Flow Matrices from APC Data on Multiple Bus Trips Using the IPF Method with an Iteratively Improved Base: Method and Empirical Evaluation. J. Transp. Eng. 140 (5), 04014008.
Ji, Y., Mishalani, R.G., McCord, M.R., 2015. Transit passenger origin-destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. Transp. Res. Part C 58, 178–192.
Ji, Y., You, Q., Jiang, S., Zhang, H.M., 2015. Statistical inference on transit route-level origin-destination flows using automatic passenger counter data. J. Adv. Transp. 49 (6), 724–737.
Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. J. Basic Eng. 82 (1), 35–45.
Kumar, P., 2019. Transit Origin Destination Estimation using Automated Data. M.S. United States- Minnesota: University of Minnesota.
Lam, W.H.K., Wu, Z.X., Chan, K.S., 2003. Estimation of Transit Origin-Destination Matrices from Passenger Counts Using a Frequency-Based Approach. J. Math. Model. Algorithms 2 (4), 329–348.
Li, D., Lin, Y., Zhao, X., Song, H., Zou, N., 2011. Estimating a Transit Passenger Trip Origin-DestinationMatrix Using Automatic Fare Collection System. In: Database Systems for Adanced Applications. Ed. by J. Xu, G. Yu, S. Zhou, R. Unland. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 502–513.
Liu, W., Tan, Q., Liu, L., 2020. Destination Estimation for Bus Passengers Based on Data Fusion. Math. Problems Eng.
Liu, X., Van Hentenryck, P., Zhao, X., 2021. Optimization Models for Estimating Transit Network Origin-Destination Flows with Big Transit Data. Journal of Big Data Analytics in Transportation 3.3, pp. 247–262.
Low, D.E., 1972. New approach to transportation systems modeling. Traffic Q. 26 (3).
Lu, D., 2008. Route Level Bus Transit Passenger Origin-Destination Flow Estimation Using Apc Data: Numerical And Empirical Investigations. The Ohio State University.
Luo, D., Cats, O., van Lint, H., 2017. Constructing Transit Origin-Destination Matrices with Spatial Clustering. Transp. Res. Record 2652 (1), 39–49.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297.
McNeil, S., Hendrickson, C., 1985. A note on alternative matrix entry estimation techniques. Transp. Res. Part B 19 (6), 509–519.
Meng, X.-L., Dyk, D.V., 1997. The EM Algorithm-an Old Folk-song Sung to a Fast New Tune. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 59 (3), 511–567.
Miller, P., De Barros, A., Kattan, L., Wirasinghe, S., 2016. Analyzing the sustainability performance of public transit. Transp. Res. Part D 44, 177–198.
Miller, P., de Barros, A.G., Kattan, L., Wirasinghe, S.C., 2016. Public transportation and sustainability: A review. KSCE J. Civil Eng. 20 (3), 1076–1083.
Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C 24, 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. Transp. Res. Record 2263 (1), 140–150.
Navick, D., Furth, P., 1994. Distance-Based Model for Estimating a Bus Route Origin-Destination Matrix. Transp. Res. Record, 16.
Nunes, A.A., Galvão Dias, T., Falcão e Cunha, J., 2016. Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. IEEE Trans. Intell. Transp. Syst. 17 (1), 133–142.
Paige, C.C., Saunders, M.A., 1982. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. ACM Trans. Math. Software 8 (1), 43–71.
Palm, M., Farber, S., Shalaby, A., Young, M., 2021. Equity Analysis and New Mobility Technologies: Toward Meaningful Interventions. J. Plann. Literature 36 (1), 31–45.
Park, E.S., Rilett, L.R., Spiegelman, C.H., 2008. A Markov Chain Monte Carlo-Based Origin Destination Matrix Estimator that is Robust to Imperfect Intelligent Transportation Systems Data. J. Intell. Transp. Syst. 12 (3), 139–155.
Peterson, A., 2007. Linköpings universitet, Institutionen för teknik och naturvetenskap, The origindestination matrix estimation problem: analysis and computations. Dept. of Science and Technology, Linköpings universitet, Norrköping.
Robert, C., Casella, G., 2011. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. Stat. Sci. 26 (1), 102–115.
Robillard, P., 1975. Estimating the O-D matrix from observed link volumes. Transp. Res. 9 (2), 123–128.
Sánchez-Martínez, G.E., 2017. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data: Dynamic Programming Approach. Transp. Res. Record 2652 (1), 1–7.
Sander, J.R., 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Min. Knowl. Discovery 2(2), 169–194.
Schoier, G., Gregorio, C., 2017. Clustering Algorithms for Spatial Big Data. In: Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C.M., Rocha, A.M.A., Taniar, D., Apduhan, B.O., Stankova, E., Cuzzocrea, A. (Eds.), Computational Science and Its Applications - ICCSA 2017. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 571–583.
Shaheen, S., Cohen, A., Chan, N., Bansal, A., 2020. Chapter 13 - Sharing strategies: carsharing, shared micromobility (bikesharing and scooter sharing), transportation network companies, microtransit, and other innovative mobility modes. Transportation, Land Use, and Environmental Planning. Ed. by E. Deakin. Elsevier, pp. 237–262.
Stead, D., Pojani, D., 2017. The Urban Transport Crisis in Emerging Economies: A Comparative Overview. In: Pojani, D., Stead, D. (Eds.), The Urban Transport Crisis in Emerging Economies. The Urban Book Series. Springer International Publishing, Cham, pp. 283–295.
Theil, H., 1971. Principles of econometrics. John Wiley & Sons, New York.
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), 1–14.
Vanderwaart, C., 2016. Planning transit networks with origin, destination, and interchange inference (Thesis). Massachusetts Institute of Technology.
Vanderwaart, C., Attanucci, J.P., Salvucci, F.P., 2017. Applications of Inferred Origins, Destinations, and Interchanges in Bus Service Planning. Transp. Res. Record 2652 (1), 70–77.
Van Zuylen, H., Willumsen, L., 1980. The Most Likely TripMatrix Estimated from Traffic Counts. Transp. Res. Part B 14, 281–293.
Wang, W., 2010. Bus passenger origin-destination estimation and travel behavior using automated data collection systems in London, UK (Thesis). Massachusetts Institute of Technology.
Widyawan, Prakasa, B., Putra, D.W., Kusumawardani, S.S., Widhiyanto, B.T.Y., Habibie, F., 2017. Big data analytic for estimation of origin-destination matrix in Bus Rapid Transit system. In: 2017 3rd International Conference on Science and Technology - Computer (ICST). 2017 3rd International Conference on Science and Technology - Computer(ICST). Yogyakarta, Indonesia: IEEE. pp. 165-170. .
Willumsen, L.G., 1978. Estimation of an O-D Matrix from Traffic Counts – A Review. Leeds, UK.
Willumsen, L.G., 1981. Simplified transport models based on traffic counts. Transportation 10 (3), 257–278.
Wilson, A.G., 1970. The Use of the Concept of Entropy in System Modelling. Oper. Res. Q. (1970–1977) 21(2), 247–265.
Wilson, A., 2011. Entropy in Urban and Regional Modelling. Routledge 1, 175.
Wong, K.I., Wong, S.C., Tong, C.O., Lam, W.H.K., Lo, H.K., Yang, H., Lo, H.P., 2005. Estimation of origin-destination matrices for a multimodal public transit network. J. Adv. Transp. 39 (2), 139–168.
Wu, L., Kang, J.E., Chung, Y., Nikolaev, A., 2021. Inferring origin-Destination demand and user preferences in a multi-modal travel environment using automated fare collection data. Omega 101, 102260.
Xie, C., Kockelman, K.M., Waller, S.T., 2010. Maximum Entropy Method for Subnetwork Origin-Destination Trip Matrix Estimation. Transp. Res. Record 2196 (1), 111–119.

Xie, C., Kockelman, K.M., Waller, S.T., 2011. A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation. Procedia - Social and Behavioral Sciences. The 19th International Symposium on Transportation and Traffic Theory 17. pp. 189–212.
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. J. Indian Inst. Sci. 99 (4), 601–619.
Zeng, Q., Reddy, A., Lu, A., Levine, B., 2014. Develop New York City Surface Transit Boarding and Alighting Ridership Daily Production Application Using Big Data. Draft for Trb 15, 1–25.
Zhao, J., 2004. The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples (Thesis). Massachusetts Institute of Technology.
Zhao, J., Rahbee, A., Wilson, N.H.M., 2007. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. Computer-Aided Civil Infrastruct. Eng. 22 (5), 376–387.