Public Transport
https://doi.org/10.1007/s12469-021-00282-0

CASE STUDY AND APPLICATIONS

The value of additional data for public transport origin–destination matrix estimation

Abderrahman Ait-Ali[1] · **Jonas Eliasson**[2]

Accepted: 18 August 2021

Abstract

Passenger origin–destination data is an important input for public transport planning. In recent years, new data sources have become increasingly common through the use of the automatic collection of entry counts, exit counts and link flows. However, collecting such data can be sometimes costly. The value of additional data collection hence has to be weighed against its costs. We study the value of additional data for estimating time-dependent origin–destination matrices, using a case study from the London Piccadilly underground line. Our focus is on how the precision of the estimated matrix increases when additional data on link flow, destination count and/or average travel distance is added, starting from origin counts only. We concentrate on the precision of the most policy-relevant estimation outputs, namely, link flows and station exit flows. Our results suggest that link flows are harder to estimate than exit flows, and only using entry and exit data is far from enough to estimate link flows with any precision. Information about the average trip distance adds greatly to the estimation precision. The marginal value of additional destination counts decreases only slowly, so a relatively large number of exit station measurement points seem warranted. Link flow data for a subset of links hardly add to the precision, especially if other data have already been added.

Keywords Dynamic origin–destination · OD estimation · Entropy maximization · Lagrangian relaxation · Smart card · Public transport

JEL Classification C60 · C89 · R41

✉ Abderrahman Ait-Ali
abderrahman.ait.ali@vti.se

[1] The Swedish National Road and Transport Research Institute (VTI), P.O. Box 55685, 114 28 Stockholm, Sweden

[2] Department of Science and Technology, Linköping University, Luntgatan 2, 602 47 Norrköping, Sweden

Published online: 14 September 2021

1 Introduction

Passenger origin–destination data is an important input for public transport (PT) planning. PT demand is summarized in time-dependent origin–destination (OD) matrices, which state the number of trips between pairs of stations, i.e., the number of passengers from an origin to a destination station per time interval, such as 15-min intervals. The knowledge of such matrices may improve the efficiency of PT supply (Pelletier et al. 2011), e.g., cost-effective timetable designs (Sun et al. 2014), or for studying passenger costs from timetable changes to solve track capacity conflicts (Ait-Ali et al. 2020).

In recent years, many PT systems have adopted new technological solutions such as automated fare collection (AFC), automated vehicle location (AVL), and vehicle weighing systems that measure passenger link flows. These solutions generate useful data, e.g., smart card data or automatic vehicle weights, which can be used for OD matrix estimation. However, acquiring such data can sometimes be costly, since it often requires installation and maintenance of measurement equipment on stations, tracks, and vehicles. Having measurements on all stations and links can be prohibitively costly, so a PT agency needs to weigh these costs against the benefits of a more precisely estimated OD matrix.

In this study, we investigate how much the precision of an estimated dynamic OD matrix for a single train line increases when additional data becomes available. We use a case study from the London Piccadilly underground line. Starting with origin counts only, we incrementally add data about exit counts, link flows and average trip distance, and measure how the precision of the estimated matrix increases with additional data.

We concentrate on the precision of the most policy-relevant variables, namely time-dependent link flows and station arrival rates, since these determine policy decisions such as service frequency (Ait-Ali et al. 2020) and capacity of stations and trains. They are also the key variables when analyzing passenger costs and benefits when adjusting timetables to solve capacity conflicts with other trains, as explained by Ait-Ali et al. (2020).

Our results suggest that entry and exit data alone is far from enough to estimate link flows with any precision. Information about the average trip distance adds greatly to estimation precision. Moreover, extrapolating from a limited number of destination counts or link flow measurements to the rest of the network results in lower added value, especially if prior data such as average travel distance is already included. Measuring a relatively large number of link flows and exit stations thus seems warranted.

Section 2 briefly summarizes the large literature on OD estimation. Section 3 describes the methodology and the case study. Results are presented in Sect. 4, and Sect. 5 concludes the paper.

2 Literature review

The research literature about OD estimation is rich and has a long history that can be traced back to the early twentieth century, e.g., with gravity models (Reilly 1931), entropy maximization (Cesario 1973), and Furness methods (Morphet 1975).

Various origin–destination problems appear in many fields of transportation research (Doblas and Benitez 2005). Most studies treat the time-independent (or static) problem (Wang et al. 2012), but there has been an increasing interest in the (harder) time-dependent (or dynamic) version (Cho et al. 2009; Zúñiga et al. 2021) which is the focus of this paper. This is partly due to increased data availability through AFC data, which is valuable for more precise estimation of (dynamic) OD matrices (Gordillo 2006). Better OD estimates can be used to improve PT services in various ways, for instance by inferring the purpose of the trips (Alsger et al. 2018), pricing and allocating railway capacity (Ait-Ali et al. 2020), or by estimating in-vehicle crowding costs (Hörcher et al. 2017; Yap et al. 2018).

The OD estimation problems also differ in terms of the considered zones and the studied type of transport traffic. Some studies looked at the flow of road vehicles (Wang et al. 2012) whereas fewer considered passenger flow in PT systems (Alsger et al. 2018; Zúñiga et al. 2021) such as buses (Wang et al. 2011), freight (Shen and Aydin 2014) or passenger rail (Gordillo 2006). Similar to the study by Wong and Tong (1998), this paper focuses on the passenger flow in a commuter rail system.

The formulation of the problem also depends on whether prior (target) matrices exist. Many authors assume the existence of such a matrix (Wang and Zhang 2016). However, this is not the case in our study and many others (Cho et al. 2009).

Generally speaking, the OD estimation problem consists of finding the most probable matrix that is consistent with observations or minimizing the deviation from observations. The definition of "most probable" (and "deviation") leads to different formulations of the objective function and functional constraints, and thus to a number of OD estimation models. For instance, deviation functions can be modeled in various ways, e.g., using discrete choice models (Ben-Akiva and Lerman 1985), generalized least square (Cascetta and Nguyen 1988), Kalman filters (Cho et al. 2009), mean least square with entropy (Xie et al. 2011), gravity models (Shen and Aydin 2014). Other modeling approaches also exist, such as genetic algorithms with entropy (Fu 2012), principal component analysis (Djukic et al. 2012), Bayesian inference (Carvalho 2014), trip chaining (Alsger et al. 2016; Hora et al. 2017), Markov chain models (Abareshi et al. 2019) or artificial neural networks (Zúñiga et al. 2021). Due to the continuous development of new approaches, several authors summarize and compare many of the different OD estimation models, e.g., Cascetta and Nguyen (1988), Abrahamsson (1998), Peterson (2007), Bera and Rao (2011), Deng and Cheng (2013), and more recently Alsger (2017) and Li et al. (2018).

In the absence of a target matrix, this paper adopts the entropy maximization (EM) principle which is also equivalent to several models such as gravity (Wilson 1967), minimum information (Van Zuylen and Willumsen 1980) and discrete choice models (Mishra et al. 2013). The EM principle originates from the statistical theory of probability. In the context of OD estimation, the EM principle relies on the idea that there are many possible trip distributions (or system states) and that the most probable OD estimate (or state) is the one that maximizes the total entropy (or randomness). Variants of such a formulation have been adopted in many OD estimation studies. Fisk (1988) used a similar (time-independent) formulation and considered that the choice of the path depends on the total travel time (or congestion). Similarly, Brenninger-Göthe et al. (1989) used it in a multi-objective program for OD

estimation using traffic counts. The same formulation was also more recently used by Xie et al. (2011) and Fu (2012).

Different types of data have been used to estimate OD matrices, e.g., cell phone network (Wang et al. 2013), tolling (Wang and Zhang 2016), GPS data and travel surveys (Ge and Fukuda 2016). In PT systems, the increasing adoption of AFC and AVL, and thus the availability of the corresponding smart card data, has led to the emergence of new applications based on such data (Nassir et al. 2011; Alsger et al. 2015), including historical and/or real-time data (Zúñiga et al. 2021).

Such studies focus on different aspects to improve PT planning and its efficiency. For instance, some research is about the estimation (Mosallanejad et al. 2019) or the validation (Alsger et al. 2016) of OD matrices in different PT systems, or more particularly about the problem of trip destinations (Trépanier et al. 2007). Moreover, different case studies exist from various PT networks around the world. These include entry-only and/or entry-exit systems from New York (Barry et al. 2002), Santiago (Munizaga and Palma 2012), China (Chen and Liu 2016) and London (Wang et al. 2011). London is also the case study in this paper. Readers interested in a summary of the different OD estimation studies using smart card PT data are referred to the review paper by Li et al. (2018).

Wang et al. (2012), one of the only studies on the value of data, looked at the additional value of well-located sensors for improving road traffic OD estimates. However, although rich, the research literature on PT data does not include, to our knowledge, studies that look at the value of knowing such additional data for dynamic OD estimation. Hence, the purpose of this paper is to fill this gap in the literature by studying the value of smart cards and additional PT data.

3 Methodology and data

In this section, we describe and formulate the main problem, i.e., the OD estimation, the solution method (details in the appendix) and the case study.

Let n_{ij}^t be the number of passengers starting from station i in time interval t, going to station j. The (dynamic) OD matrix estimation consists of finding a time-dependent origin–destination matrix $\left\{ n_{ij}^t \right\}$ that is consistent with observations. This is done by estimating the entropy-maximizing matrix (the "most probable" matrix) that is consistent with observations of origin counts O_i^t, destination counts D_j^t, link flows F_l^t and the average trip distance $\overline{d}$.

In the following, we assume that origin counts are always available since many PT systems collect such data at entry gates. Destination counts, however, are not always collected, since equipping exit gates with data collection equipment is costly. Link flow measurements require specialized equipment, such as automated vehicle weighing. The average trip distance is usually estimated using travel surveys.

From the network and timetable, the travel time matrix τ_{ij} can be calculated. Given these, the estimated number of arriving passengers at station j in time interval t can be

calculated as $\Sigma_i n_{ij}^{t-\tau_{ij}}$. The estimated flow on link l at time t can be calculated as $\sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{ij}}$, where $\{i < l\}$ denote all stations i preceding link l and $\{j > l\}$ denote all stations succeeding it. Given distances between stations d_{ij}, the estimated average trip distance is calculated as $\frac{\Sigma_{ijt} n_{ij}^t d_{ij}}{\Sigma_{ijt} n_{ij}^t}$.

The core question of the paper is how the precision of the estimated matrix improves when more and more data become available. Let L be the subset of links where link flows are known, and Δ is the subset of stations where exit counts are known. Similar to the EM model by Xie et al. (2010) and Fu (2012), the studied dynamic OD estimation problem is formulated in Eq. (1).

$$\begin{cases} \min_{n_{ij}^t \geq 0} \sum_{ijt} \left(n_{ij}^t \log\left(n_{ij}^t \right) - n_{ij}^t \right) \\ \sum_j n_{ij}^t = O_i^t; \quad \forall i, \quad \forall t \, (1.1) \\ \sum_i n_{ij}^{t-\tau_{ij}} = D_j^t; \quad \forall t, \quad \forall j \in \Delta \, (1.2) \\ \sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{ij}} = F_l^t; \quad \forall t, \quad \forall l \in L \, (1.3) \\ \frac{\Sigma_{ijt} n_{ij}^t d_{ij}}{\Sigma_{ijt} n_{ij}^t} = \bar{d} \, (1.4) \\ n_{ii}^t = 0; \quad \forall i \end{cases} . \tag{1}$$

The central question can now be stated as: By how much is the precision of the estimated OD matrix n_{ij}^t improved when additional data becomes available, i.e., when the sets Δ and L become larger?

We must thus define what kind of "precision" we are interested in. In applied policy-making, e.g., timetable design and investments in links or stations, the exact cells of the OD matrix are less important. What matters most are station flows and link flows, since this determines the crowding levels in vehicles and stations. This is used for decisions about link and station capacity upgrades, station staff planning, and timetable design (timetable optimization depends mainly on passenger departure and arrival rates per line segment, and on crowding levels on different links). Hence, we will concentrate on how close to reality the estimated OD matrix is in terms of link flows and arrival rates per station (origin rates are assumed to be known). We thus measure the relative root mean square error (or deviation) for link flows ($\text{RMSE}_{\text{link}}$) and arrival rates at destination stations ($\text{RMSE}_{\text{dest}}$), and study how these vary with more available information such as when the sets of available link flows and destination counts, L and Δ, become larger.

Let $\widehat{D}_j^t = \sum_i n_{ij}^{t-\tau_{ij}}$ be the estimated number of passengers arriving at station j and time interval t, and $\widehat{F}_l^t = \sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{ij}}$ the estimated link flow on link l and time interval t. The relative errors are then defined as in Eqs. (2) and (3).

$$\mathrm{RMSE}_{\mathrm{dest}} = \frac{\sqrt{\Sigma_{jt}\left(\widehat{D}_j^t - D_j^t\right)^2}}{\sqrt{\Sigma_{jt}\left(D_j^t\right)^2}} \tag{2}$$

$$\mathrm{RMSE}_{\mathrm{link}} = \frac{\sqrt{\Sigma_{lt}\left(\widehat{F}_l^t - F_l^t\right)^2}}{\sqrt{\Sigma_{lt}\left(F_l^t\right)^2}} \tag{3}$$

For those stations and links where data is available ($j \in \Delta$ and $l \in L$) the errors will of course be zero (assuming that the optimization problem is feasible). The errors hence measure the deviations for the unobserved stations and links—in other words, how well the available station and link data can be extrapolated to unobserved stations and links.

Table 2 lists the different combinations of destination counts, link flows and average travel distance that we will explore.

3.1 Solution method

The EM estimation model is a convex (nonlinear) minimization program with a nonlinear objective function (the total entropy) and linear constraints. Finding a solution for time-dependent real-world instances (e.g., large networks and/or longer periods) is generally hard. Thus, instead of using state-of-the-art solvers, we derive the iterative solution methods using Lagrangian relaxation.

We first relax the constraints and associate corresponding Lagrangian multipliers as presented in Table 1. This leads to the formulation of a Lagrange function (or relaxed dual objective function). More details can be found in the Appendix.

Using first-order optimality conditions on the Lagrange function, we can formulate the (primal) solution, i.e., OD estimate as a function of the (dual) Lagrangian multipliers. Depending on the studied data, we find different solution formulations of the dynamic OD estimate n_{ij}^t. Table 2 presents the formulations for the different studied variants. A more detailed derivation of these solution formulations is described in the appendix.

To estimate the multipliers, we use the problem constraints. In some trivial cases, it is possible to find a closed-form expression such as in the basic O model

Table 1 Lagrangian multipliers and the corresponding relaxed constraints

Constraint(s)	Description	Lagrangian multiplier(s)
(1.1)	Origin counts	λ_{it}; $\forall i, \forall t$
(1.2)	Destination (or exit) counts	μ_{jt}; $\forall t, \forall j \in \Delta$
(1.3)	Link flow counts	φ_{lt}; $\forall t, \forall l \in L$
(1.4)	Average travel distance	θ

Table 2 Solution models and formulations of different variants

Variant	Model	Formulation
O	Origin counts only, for all stations (basic model)	$e^{\lambda_{it}}$
O-d	Origin counts (for all stations) and average travel distance	$e^{\lambda_{it}+\theta d_{ij}}$
O-D	Origin counts (for all stations) and destination counts for a subset of stations Δ	$\begin{cases} e^{\lambda_{it}+\mu_{j,t+\tau_{ij}}}; j \in \Delta \\ e^{\lambda_{it}}; j \notin \Delta \end{cases}$
O-d-D	As O-D plus average travel distance	$\begin{cases} e^{\lambda_{it}+\theta d_{ij}+\mu_{j,t+\tau_{ij}}}; j \in \Delta \\ e^{\lambda_{it}+\theta d_{ij}}; j \notin \Delta \end{cases}$
O-F	Origin counts (for all stations) and link flows for a subset of links L	$\begin{cases} e^{\lambda_{it}+\varphi_{lt}}; l=(i,j) \in L \\ e^{\lambda_{it}}; l=(i,j) \notin L \end{cases}$
O-D-F	As O-F but with destination counts for all stations	$\begin{cases} e^{\lambda_{it}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} l=(i,j) \in L \\ e^{\lambda_{it}+\mu_{j,t+\tau_{ij}}}; l=(i,j) \notin L \end{cases}$
O-d-D-F	As O-D-F plus average travel distance	$\begin{cases} e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}}; l=(i,j) \in L \\ e^{\lambda_{it}+\theta d_{ij}+\mu_{j,t+\tau_{ij}}}; l=(i,j) \notin L \end{cases}$

where $\sum_j n_{ij}^t = O_i^t \Rightarrow n_{ij}^t = e^{\lambda_{it}} = \frac{O_i^t}{|S|-1}$, i.e., all destinations have a similar attractivity. In other (more interesting) cases, this is often difficult (sometimes impossible). Thus, we attempt to find numerical solutions by iteratively balancing the relaxed constraints corresponding to additional studied data. Figure 1 is an example of an iterative algorithm to find the numerical solution of the multipliers for the O-d model. More details about the iterative algorithms can be found in the Appendix.

The iterative solution algorithm stops when the constraints are satisfied, up to a certain tolerance ϵ. Note that the use of the (hard) constraint of origin counts (from smart cards) to derive an analytic expression of the dynamic OD estimate yields the formulation in (4).

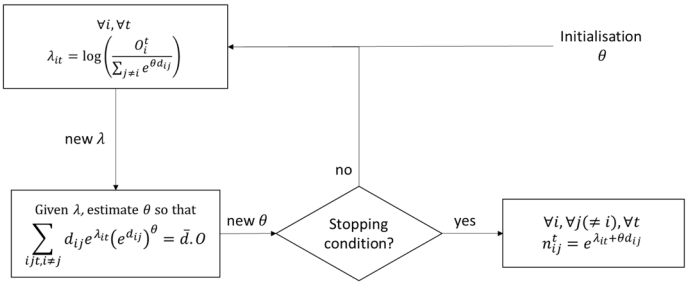

Fig. 1 Iterative algorithm for the O-d variant

$$n_{ij}^t = O_i^t p(j|i,t)$$
$$\text{where } p(j|i,t) = \frac{e^{u_{ij}^t}}{\sum_j e^{u_{ij}^t}} \text{ and } u_{ij}^t = K_j^t + \theta_1 k_{ijt}^{(1)} + \ldots + \theta_m k_{ijt}^{(m)} \quad (4)$$

The term $p(j|i,t)$ can be seen as the probability of choosing destination j when departing from origin i at time interval t. In this case, the exponent u_{ij}^t can be interpreted as the total utility for traveling to j from i during time interval t. Such utility may include parameters $k_{ijt}^{(1)}, \ldots, k_{ijt}^{(m)}$ corresponding to m types of additional data, if available. The coefficients (or multipliers) $\theta_1, \ldots, \theta_m$ are estimated to reflect utilities (if $\theta \geq 0$) or disutilities (if not). The constant K_j^t is specific to the destination station j and time interval t.

Such interpretation can also be found in discrete choice models (without the random error term) where the discrete choices are between the different destination stations j given an origin i. The (alternative-specific) constants K_j^t and parameters $\theta_1, \ldots, \theta_m$ are specific to the PT system where the OD estimation is performed. They need to be estimated to reflect the (dis-)utilities explaining the choice of the passengers. It is possible to estimate the values of these parameters using additional data, e.g., from smart cards, stated (or revealed) preference surveys, old OD or target matrices from the same PT system.

3.2 Case study data

To explore the question formulated above, we use a case study based on the London Piccadilly underground line. Transport for London (or TfL) provides open access to a comprehensive multi-rail demand dataset as part of the NUMBAT project (TfL 2018). Based on the use of smart cards at entry/exit station gates during a typical 2018 autumn weekday, the dataset provides information about the number of passengers boarding and alighting at each station (per 15 min), and link flows (per 15 min) for a subset of the links (data for around 100 links are available, but we study 12 of the most crowded links). The data also contains an estimated OD matrix for longer time periods which is used in this case study to calculate the average travel distance.

The Piccadilly line (Fig. 2) is more than 70 km long and consists of 53 stations with two different western branches at Acton Town station. Note the one-way trajectory around the Heathrow airport from/to Hatton Cross through terminal 4 then 2 and 3.

In Fig. 3, stations are sorted according to their location on the studied line to make it easy to visualize the symmetry of the distance matrix. However, the matrix, as shown in the figure, is not completely symmetric, see around the airport due to the previously mentioned one-way trajectory.

To calculate the average travel times τ_{ij}, we use the travel distances between each pair of stations which is illustrated in Fig. 3. We assume that all trains are running according to the train timetable (headways) presented in Table 3, and that their average speed is 33 km/h (TfL 2018).

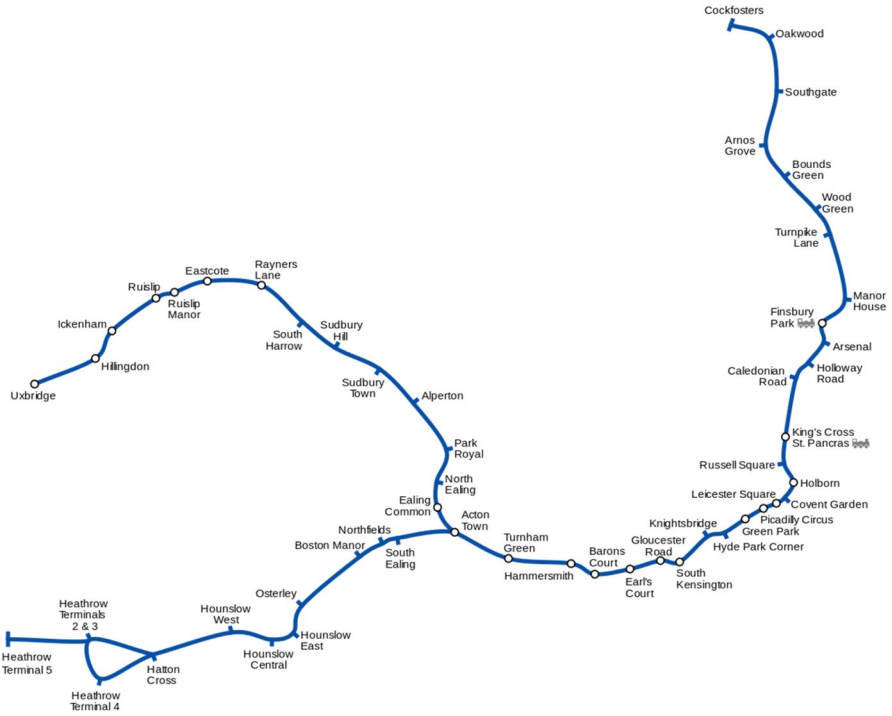

Fig. 2 Piccadilly line of the London commuter network

As presented in Table 3, we focus in this study on three different time periods, i.e., morning and afternoon (peak) as well as midday (off-peak). These periods are also illustrated in Fig. 4 which also shows the variation of both the origin (boarders) and the destination or exit (alighters) counts per 15-min time interval over the day. The studied time periods are separated by dashed vertical lines in the figure.

In addition to the temporal variation (per time interval) of the number of boarders and alighters as shown in Fig. 4, we present the spatial variation (per station) in Fig. 5 over the day. The stations on the horizontal axis are sorted by the number of alighters (from highest). Figure 6 presents the link flows during the day for three of the largest links.

The average travel distance per passenger $\overline{d}$ is usually estimated from demand travel surveys. For our case study, we calculate it based on the available OD matrices (per time period). Table 4 shows the average travel distance in km per passenger for the different studied time periods of the day.

4 Results

In this section, we present results on how the precision of the estimated matrix varies when more data is included in the estimation. We focus on the precision of arrival rates at destination stations and link flows. Several scenarios with different

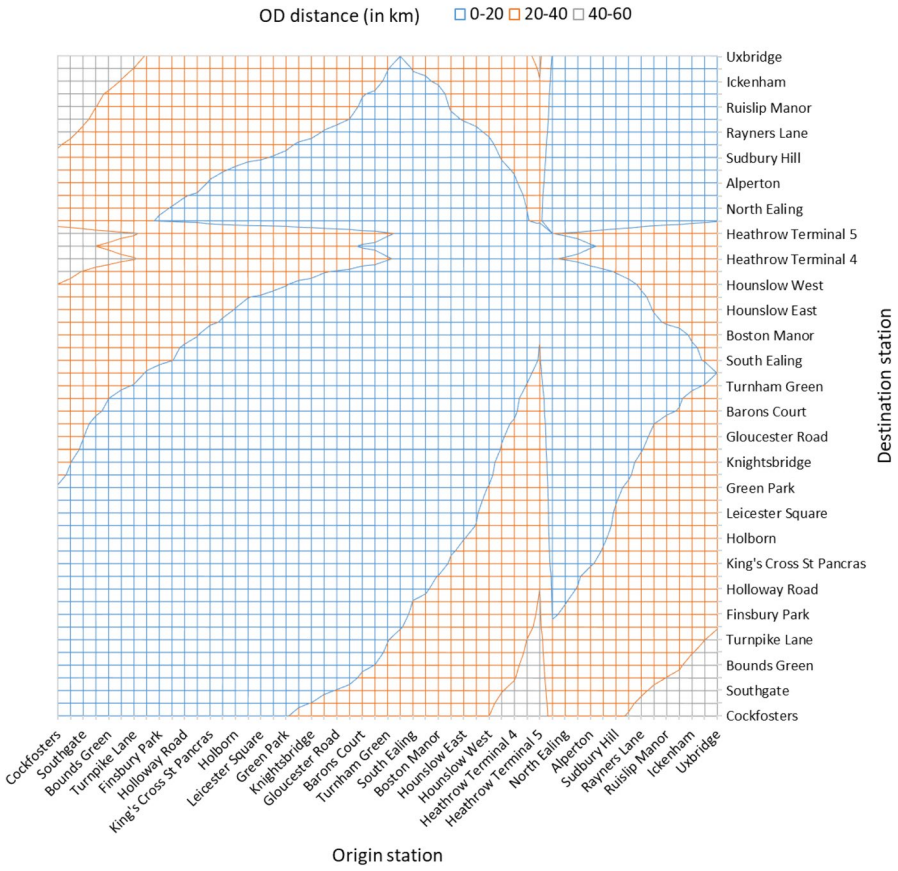

Fig. 3 Travel distances (in km) between the different pairs of stations

Table 3 Train headways in the studied time periods for both directions (TfL 2018)

	Peak i.e., morning (7.00–10.00) and afternoon (16.00–19.00)	Off-peak e.g., midday (10.00–16.00)
Main	5/2 min	5 min
Branches	5 min	10 min

types of additional data are tested. Table 5 presents an overview of the reported results, i.e., tested models and the corresponding presented estimation errors.

Two types of data are incrementally added, i.e., destination and link data. The value of such additional data is studied by testing different estimation models. The average travel distance is also studied in certain models.

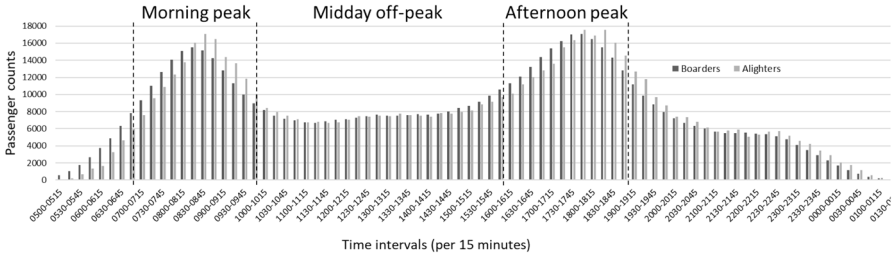

Fig. 4 Temporal variation of the total number of boarders and alighters

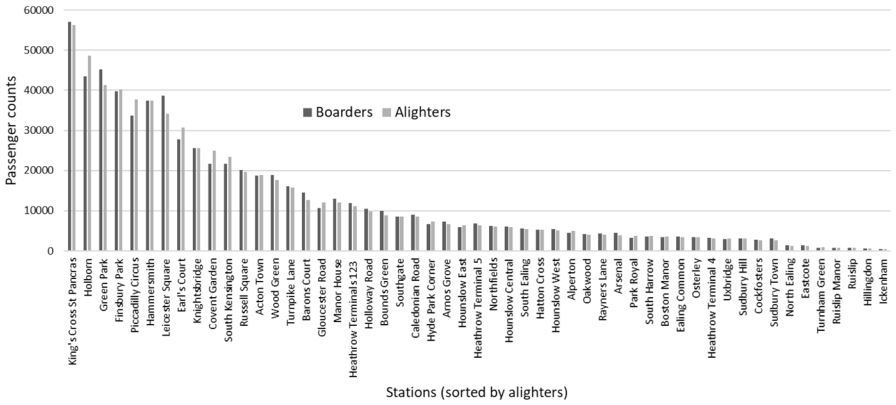

Fig. 5 Spatial variation of the total number of boarders and alighters

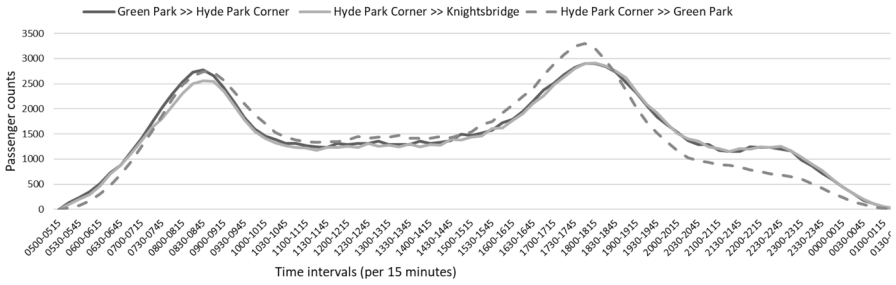

Fig. 6 Temporal variation of the passenger flow in three of the most crowded links

Table 4 Average travel distance (in km per pax) for the different periods

Period	Average travel distance (km per pax)
Morning (peak)	9.7
Midday (off-peak)	8.6
Afternoon (peak)	9.1

Table 5 Overview of the tested models and the presented estimation errors

Tested models		Estimation error
Destination data incrementally added	Link flow data incrementally added	
O-D, O-d-D	O-F	$RMSE_{dest}$
O-D, O-d-D	O-F, O-D-F, O-d-D-F	$RMSE_{link}$

Note that when incrementally including link flows in O-D-F and O-d-D-F, exit counts (destination data) from all stations are considered unlike other models (i.e., O-D and O-d-D variants) where these are also incrementally included.

4.1 Estimating arrival rates per station

We first focus on the destination estimation, i.e., the number of alighters in the system per 15-min interval and station. Figure 7 shows how the relative error ($RMSE_{dest}$) varies when data for more and more stations is added. Stations are sorted according to their total number of alighters, and for each step along the x-axis, data for one additional station is added. The $RMSE_{dest}$ error is presented separately for three parts of the day (morning, midday and afternoon). When data for all destinations has been added, the error is of course zero.

Surprisingly, including a relatively small number of destinations increases the error for both the midday and afternoon periods. Only after a certain amount of destination data has been added does the error decrease. Adding the average travel distance data in the estimation further reduces the relative error (up to 50%). For the midday and afternoon periods, just adding the average travel distance decreases the error by as much as adding data for almost all destinations but without the average distance.

These results suggest that having data only for a subset of destinations is sometimes not enough—in fact, it may even increase the overall error, e.g., midday and afternoon periods. Having enough exit counts seems to be important to get better estimates. However, systems for collecting such larger amounts of data are often expensive, e.g., design, operation and maintenance. Thus, the importance of comparing the data collection costs and value for alternative types of data, e.g., average trip distances which seems to be highly valuable here.

With a focus on the O-F model, Fig. 8 shows how $RMSE_{dest}$ varies when incrementally adding link flow data from 12 of the most crowded links. The error is presented for the three-time periods of the day, and links are sorted and added according to their passenger flows. The order of added links may differ between the different periods, see later in Fig. 10, for the specific added links.

Including link flow data does not seem to always reduce the destination error. It increases during midday peak hours whereas it remains almost constant in the afternoon. The exception is the morning period as the error is slightly reduced when more links are added.

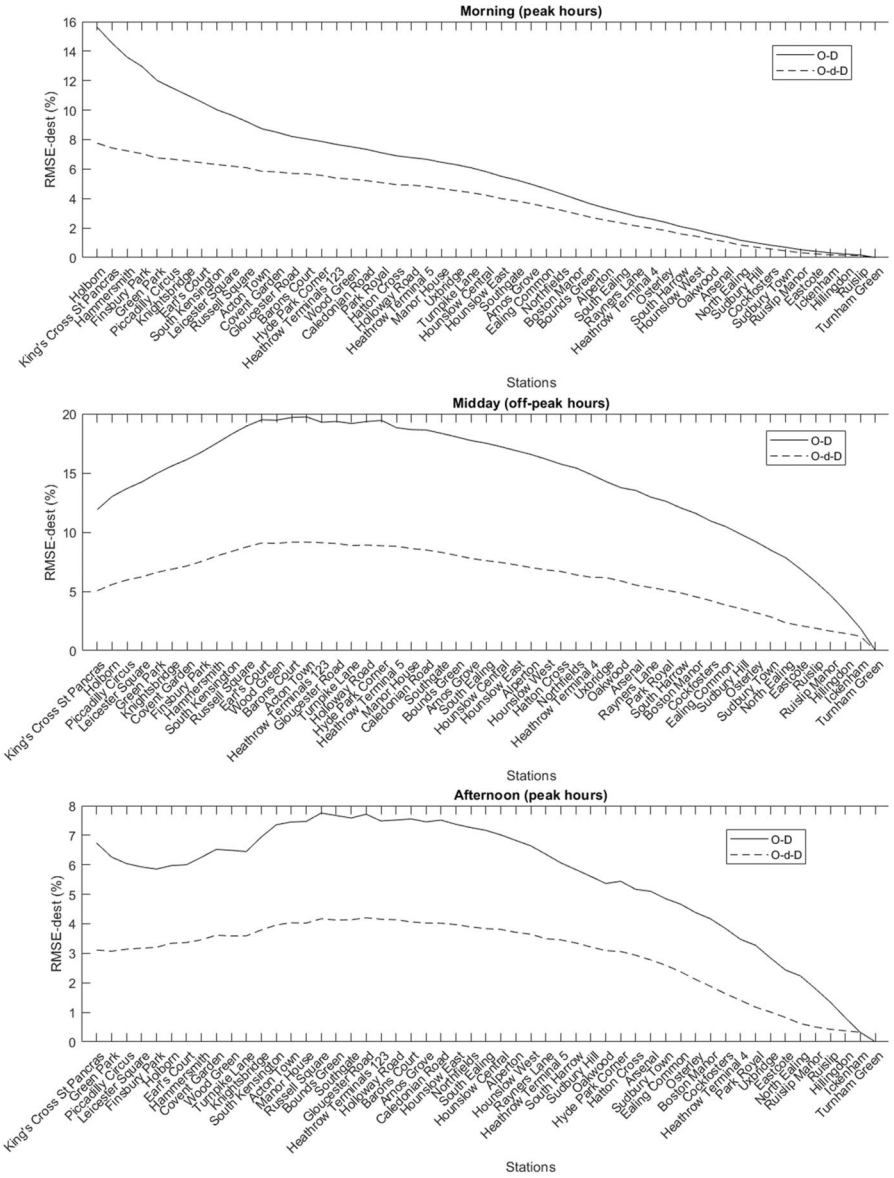

Fig. 7 Variation of $\mathrm{RMSE}_{\mathrm{dest}}$ as destination data is incrementally added

The results indicate that detailed (often expensive) data such as link flows are not always valuable. It is therefore important to investigate (cheaper) complementary data types that can improve the quality of the estimates. Later in Fig. 10, we study the combination with other additional data, i.e., average travel distance and exit counts.

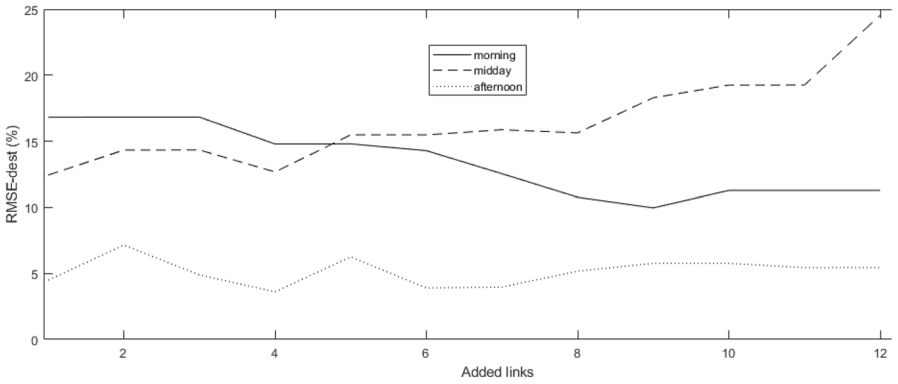

Fig. 8 Variation of $RMSE_{dest}$ as link data is incrementally added to the O-F model

4.2 Estimating passenger flows per link

Figure 9 (similar to Fig. 7) shows how the link flow error $RMSE_{link}$ varies when additional destination data is incrementally added, i.e., O-D and O-d-D models. Unlike Fig. 7 which presents the deviation errors of the (more aggregate) exit counts, Fig. 9 shows the error for the (more detailed) link flows between the studied OD pairs. Thus, the errors are higher in Fig. 9.

Even after all exit counts have been included, the link flow error is far from zero. In fact, adding destination data hardly improves the link flow estimation, apart from the first few destinations. As above, information about the average trip distance greatly decreases the error. Surprisingly, however, adding more destination data tends to increase the error in this case. Therefore, better (and more economical) estimates of link flows can be reached by combining different types of data and by using the right (amount of) data points.

To get decent precision in the link flow estimation, link flow data seems to be necessary. Figure 10 shows the change in $RMSE_{link}$ when incrementally including flow data for 12 of the most crowded links (as in Fig. 8). The figures show estimations without destination data (O-F), with all destination data (O-D-F), and with destination data and average travel distance (O-D-F-d).

In the O-F model, the error decreases with more data, as expected. However, although the relative error is lower (than O-F), in models with destination data (O-D-F) and average travel distance (O-d-D-F), it remains almost constant when link flow data is added. This is the case for all the periods except for the morning peak hours where the relative error decreases after adding the 4th link but remains constant after.

These results indicate that the marginal value of additional link flow data is high only when data such as exit counts and average travel distance are absent. If such data exists and is used in the estimation, the value of link flow data becomes almost insignificant. It is therefore important to compare the cost of collecting the additional data with its marginal value. Including more link data can provide further

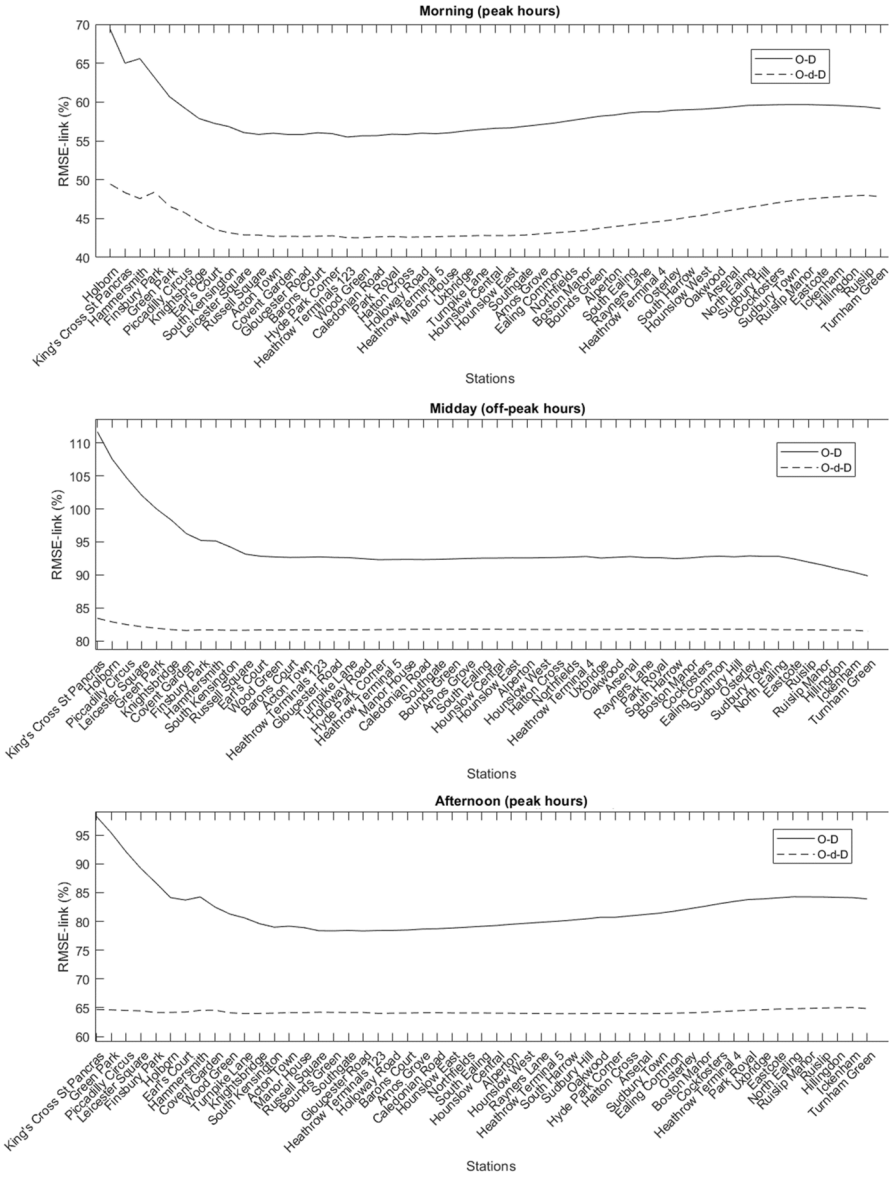

Fig. 9 Variation of $\mathrm{RMSE}_{\mathrm{link}}$ as destination data is incrementally added

insights, however, the estimation models tend to become more computationally expensive.

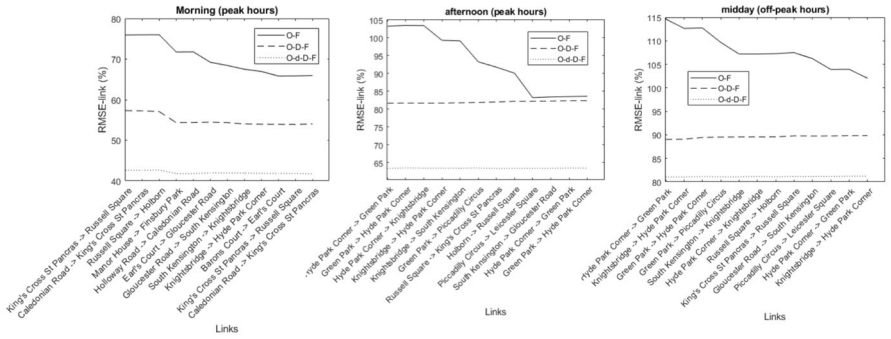

Fig. 10 Variation of $\mathrm{RMSE}_{\mathrm{link}}$ as link flow data is added incrementally

5 Conclusions and future works

Even if the literature includes several studies on (dynamic) OD-matrix estimation, this work attempts to assess the marginal value of additional data in terms of estimation precision. The additional data we have studied is arrival rates per station (which may be collected through AFC systems or specialized equipment), link flows (which may be collected by a vehicle weighing system) and average trip distances (from travel surveys). We explore this through a case study based on the London Piccadilly line in 2018, separating three time periods of the day (i.e., morning and afternoon peak hours, midday). We focus on the precision of the estimated time-dependent arrival rates and link flows, rather than on individual cells in the time-dependent OD matrix.

The results indicate that arrival rates per destination station (if enough) may improve the estimation, but in two cases of three, including data for a subset of destinations made the estimation worse. Perhaps contrary to expectations, it turns out to be valuable to have data for a very large share of destinations: the marginal value of acquiring more data, even for the last stations, is surprisingly high. Arrival rates (exit counts) can be collected easily for AFC systems which are based on "tap-in/tap-out" (such as London), but for entry-only AFC systems (such as Stockholm), special data collection equipment needs to be installed at exit gates. Our results suggest that installing such equipment may only lead to marginal improvements of the estimated OD-data unless a large (enough) share of stations is equipped. If such equipment is costly, it might be more cost-efficient to consider other forms of data collection, and to study the value of the collected data.

Similarly, the study of a subset of added link flows indicates that link estimation may improve but only if no prior additional data is already added. Otherwise, the estimates are better (than with no prior data) but do not improve with added link data. Thus, the marginal value of such detailed data may be insignificant if specific prior data is already included. These results show that detailed (often expensive) data may have a lower marginal value for the demand estimation and can therefore lead to less accurate demand-sensitive policy decisions, e.g., setting welfare-optimal line frequencies.

Based on a study case, the paper highlights that collecting additional, more detailed data (often more expensive) is not always leading to more accurate estimates, i.e., lower marginal value. Thus, the results emphasize the importance of considering both the costs of collecting such additional data and its marginal value.

There are a number of possible future works that can further validate these results, e.g., using other estimation models, metrics for the valuation of the estimation quality, and by studying additional data sources in other case studies. For instance, we used the relative RMSE to quantify the precision of these estimates, but other metrics can be tested in future work, such as the implied optimal service frequencies (Ait-Ali et al. 2020), or levels of in-vehicle crowding (Çelebi and İmre 2020). Full-day estimation instead of per period can also be tested when additional data is lacking. However, assuming that the time-aggregated OD matrix is symmetric is a strong assumption, and is for example violated in our data set. Furthermore, such full-day estimation also requires additional computational power and can be intractable for large networks.

Overall, information about average trip distances gives by far the greatest improvement of the estimation. Acquiring such estimates, from travel surveys, link flow measurements or other means is hence a priority. In this study, we have only used one average distance (per time period) for the whole line, but obviously, getting more detailed data (for parts of the line) would be highly valuable. Furthermore, instead of gradually including data based on the magnitude of the counts or flow, other orders can also be tested, e.g., based on job or home locations during peak hours. Closely related, the model can be adapted to find more valuable data collection strategies, e.g., data types and their spatiotemporal locations.

Appendix 1: Solution formulation

The Lagrangian relaxation of the different constraints with the corresponding multipliers leads to the following Lagrange function

$$\mathcal{L}(n,\lambda,\mu,\theta,\varphi) = E(n) - \sum\nolimits_{it}\lambda_{it}\left(\sum\nolimits_{j} n_{ij}^{t} - O_i^t\right) - \sum\nolimits_{j\in\Delta,t}\mu_{jt}\left(\sum\nolimits_{i} n_{ij}^{t-\tau_{ij}} - D_j^t\right)$$
$$- \theta\left(\sum\nolimits_{ijt} d_{ij} n_{ij}^{t} - \bar{d}.O\right) - \sum\nolimits_{l\in L,t}\varphi_{lt}\left(\sum_{\substack{i<l\\ j>l}} n_{ij}^{t-\tau_{ij}} - F_l^t\right)$$

The first-order optimality condition for the function $\mathcal{L}$ in terms of the variable n_{ij}^t is as follows

$$\frac{\partial\mathcal{L}}{\partial n_{ij}^t} = 0 \Rightarrow \mathcal{L}(\lambda,\mu,\theta,\varphi) = \log\left(n_{ij}^t\right) - \lambda_{it} - \mu_{j,t+\tau_{ij}} - d_{ij}\theta - \varphi_{lt}$$

Note that the multiplier $\mu_{j,t+\tau_{ij}}$ is only included if $j \in \Delta$, i.e., known data on the exit counts at station j. Similarly, φ_{lt} is also only included if $l = (i,j) \in L$, i.e., known flow at link l. Thus, we have the following general solution formulation

$$n_{ij}^t = e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} = e^{\lambda_{it}} e^{\theta d_{ij}} e^{\varphi_{lt}} e^{\mu_{j,t+\tau_{ij}}}$$

To sum up, depending on the studied additional data, we have different variants of the solution formulation as presented in Table 2.

Appendix 2: Iterative algorithm

The iterative algorithm aims at estimating the multipliers, i.e., λ, μ, θ and φ. For that, each iteration of the algorithm attempts to balance the different constraints until these are satisfied (up to a certain error tolerance ϵ). To derive the algorithms, we first use the (hard) constraints for the origin counts ($O_i^t \neq 0$) to estimate λ_{it} as follows

$$\sum\nolimits_j n_{ij}^t = O_i^t \Rightarrow \sum\nolimits_j e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} = O_i^t$$

$$\Rightarrow e^{\lambda_{it}} = \frac{O_i^t}{\sum_j e^{\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}}}$$

With smart card data on counts ($D_j^t \neq 0$) at large destination stations, we use the corresponding constraints to estimate μ_{jt} as follows

$$\sum\nolimits_i n_{ij}^{t-\tau_{ij}} = D_j^t \Rightarrow \sum\nolimits_i e^{\lambda_{i,t-\tau_{ij}}+\theta d_{ij}+\varphi_{l,t-\tau_{ij}}+\mu_{jt}} = D_j^t$$

$$\Rightarrow e^{\mu_{jt}} = \frac{D_j^t}{\sum_i e^{\lambda_{i,t-\tau_{ij}}+\theta d_{ij}+\varphi_{l,t-\tau_{ij}}}}$$

Similarly, the constraints for additional data on the average travel distance $\bar{d}$ can be used to estimate θ by finding the solution (root) of the following equation

$$\sum\nolimits_{ijt} d_{ij} n_{ij}^t = \bar{d}O \Rightarrow \sum\nolimits_{ijt} d_{ij} e^{\lambda_{it}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} \left(e^{d_{ij}}\right)^{\theta} = \bar{d}O$$

When we include additional data on flows ($F_{l=(i,j)}^t \neq 0$) at crowded links, we estimate φ_{lt} by solving the following system of linear (in $e^{\varphi_{lt}}$) equations

$$\sum_{\substack{s<l \\ e>l}} n_{se}^{t-\tau_{se}} = F_l^t$$

$$\Rightarrow e^{\varphi_{lt}} e^{\lambda_{it}+\theta d_{ij}+\mu_{j,t+\tau_{ij}}} + \sum_{\substack{s<l \\ e>l \\ l^*=(s,e)\in L \\ l^*\neq l}} e^{\varphi_{l^*,t-\tau_{si}}} e^{\lambda_{s,t-\tau_{si}}+\theta d_{se}+\mu_{e,t+\tau_{se}-\tau_{si}}} = F_l^t$$

The iterative algorithm stops either after a certain number of iterations or when all the constraints are satisfied, e.g., using RMSE and an error tolerance ϵ.

Acknowledgements This research is part of the project Socio-economically efficient allocation of railway capacity, SamEff (Samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar) which is funded by a grant from the Swedish Transport Administration (Trafikverket). The authors are grateful to Jan Lundgren for improvement suggestions on an earlier version.

Author contribution AA-A: conceptualization, methodology, software, data curation, formal analysis, investigation, writing—original draft, visualization. JE: supervision, conceptualization, project administration, funding acquisition, resources, validation, writing—review and editing.

Funding Open access funding provided by Swedish National Road and Transport Research Institute (VTI). The research leading to these results received funding from the Swedish Transport Administration (Trafikverket) as part of the project Socio-economically efficient allocation of railway capacity, SamEff (Samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Abareshi M, Zaferanieh M, Safi MR (2019) Origin-destination matrix estimation problem in a Markov chain approach. Netw Spat Econ 19:1069–1096

Abrahamsson T (1998) Estimation of origin-destination matrices using traffic counts—a literature survey. IIASA, Laxenburg

Ait-Ali A, Warg J, Eliasson J (2020a) Pricing commercial train path requests based on societal costs. Transp Res Part A Policy Pract 132:452–464

Ait-Ali A, Eliasson J, Warg J (2020b) Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding? Working Papers, Swedish National Road & Transport Research Institute. VTI, Stockholm

Alsger AA (2017) Estimation of transit origin destination matrices using smart card fare data. PhD Thesis, The University of Queensland

Alsger AA, Mesbah M, Ferreira L, Safi H (2015) Use of smart card fare data to estimate public transport origin-destination matrix. Transp Res Rec 2535:88–96

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 Emerg Technol 68:490–506

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

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

Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge

Bera S, Krishna Rao K V (2011) Estimation of origin-destination matrix from traffic counts: the state of the art. Eur Transp 49:2–23

Brenninger-Göthe M, Jörnsten KO, Lundgren JT (1989) Estimation of origin-destination matrices from traffic counts using multiobjective programming formulations. Transp Res Part B Methodol 23:257–269

Carvalho L (2014) A Bayesian statistical approach for inference on static origin-destination matrices in transportation studies. Technometrics 56:225–237

Cascetta E, Nguyen S (1988) A unified framework for estimating or updating origin/destination matrices from traffic counts. Transp Res Part B Methodol 22:437–455

Çelebi D, İmre Ş (2020) Measuring crowding-related comfort in public transport. Transp Plan Technol 43:735–750

Cesario FJ (1973) A note on the entropy model of trip distribution. Transp Res 7:331–333

Chen SP, Liu DZ (2016) Bus passenger origin-destination matrix estimation using available information from automatic data collection systems in Chongqing, China. Adv Mater Res 779–780:878–889

Cho HJ, Jou YJ, Lan CL (2009) Time dependent origin-destination estimation from traffic count without prior information. Netw Spat Econ 9:145–170

Deng Q, Cheng L (2013) Research review of origin-destination trip demand estimation for subnetwork analysis. Proc Soc Behav Sci 96:1485–1493

Djukic T, Flötteröd G, Lint HV, Hoogendoorn S (2012) Efficient real time OD matrix estimation based on Principal Component Analysis. In: 15th international IEEE conference on intelligent transportation systems. IEEE, pp 115–121

Doblas J, Benitez FG (2005) An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix. Transp Res Part B Methodol 39:565–591

Fisk C (1988) On combining maximum entropy trip matrix estimation with user optimal assignment. Transp Res Part B Methodol 22:69–73

Fu GJ (2012) Study of solving crossing origin-destination matrix based on entropy maximizing model. Appl Mech Mater 182–183:970–974

Ge Q, Fukuda D (2016) Updating origin–destination matrices with aggregated data of GPS traces. Transp Res Part C Emerg Technol 69:291–312

Gordillo F (2006) The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation. Massachusetts Institute of Technology, Cambridge

Hora J, Dias TG, 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

Hörcher D, Graham DJ, Anderson RJ (2017) Crowding cost estimation with large scale smart card and vehicle location data. Transp Res Part B Methodol 95:105–125

Li T, Sun D, Jing P, Yang K (2018) Smart card data mining of public transport destination: a literature review. Information 9:18–21

Mishra S, Wang Y, Zhu X, Moeckel R, Mahapatra S (2013) Comparison between gravity and destination choice models for trip distribution in Maryland. In: 92nd annual meeting of the Transportation Research Board (TRB), Washington, DC, 13–17 January 2013

Morphet R (1975) A note on the calculation and calibration of doubly constrained trip distribution models. Transportation 4:43–53

Mosallanejad M, Somenahalli S, Mills D (2019) Origin-destination estimation of bus users by smart card data. In: Geertman S, ZhAN Q, Allan A, Pettit C (eds) Computational urban planning and management for smart cities. Springer, Cham, pp 305–320

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

Nassir N, Khani A, Lee SG, Noh H, Hickman M (2011) Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transp Res Rec 2263:140–150

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

Peterson A (2007) The origin–destination matrix estimation problem—analysis and computations. PhD Dissertations, University of Linköping

Reilly WJ (1931) The law of retail gravitation. W.J. Reilly, New York

Shen G, Aydin SG (2014) Origin–destination missing data estimation for freight transportation planning: a gravity model-based regression approach. Transp Plan Technol 37:505–524

Sun L, Jin JG, Lee D-H, Axhausen KW, Erath A (2014) Demand-driven timetable design for metro services. Transp Res Part C Emerg Technol 46:284–299

TFL (2018) A comprehensive multi-rail demand data set for London. Project NUMBAT

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

van Zuylen HJ, Willumsen LG (1980) The most likely trip matrix estimated from traffic counts. Transp Res Part B Methodol 14:281–293

Wang H, Zhang XN (2016) Estimation of origin-destination matrix with tolling data. Appl Mech Mater 50–51:239–244

Wang W, Attanucci J, Wilson N (2011) Bus passenger origin-destination estimation and related analyses using automated data collection systems. J Public Transp 14(4):131–150

Wang N, Gentili M, Mirchandani P (2012) Model to locate sensors for estimation of static origin-destination volumes given prior flow information. Transp Res Rec 2283:67–73

Wang M-H, Schrock SD, vander Broek N, Mulinazzi T (2013) Estimating dynamic origin-destination data and travel demand using cell phone network data. Int J Intell Transp Syst Res 11:76–86

Wilson AG (1967) A statistical theory of spatial distribution models. Transp Res 1(3):253–269

Wong SC, Tong CO (1998) Estimation of time-dependent origin–destination matrices for transit networks. Transp Res Part B Methodol 32:35–48

Xie C, Kockelman KM, Waller ST (2010) Maximum entropy method for subnetwork origin-destination trip matrix estimation. Transp Res Rec 2196:111–119

Xie C, Kockelman KM, Waller ST (2011) A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation. Procedia Soc Behav Sci 17:189–212

Yap M, Cats O, van Arem B (2018) Crowding valuation in urban tram and bus transportation based on smart card data. Transportmetrika A Transport Sci 16:1–20

Zúñiga F, Muñoz JC, Giesen R (2021) Estimation and prediction of dynamic matrix travel on a public transport corridor using historical data and real-time information. Public Transport 13:59–80