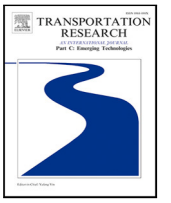
Transportation Research Part C 165 (2024) 104694

Contents lists available at ScienceDirect

Transportation Research Part C

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



Origin–destination matrix estimation for public transport: A multi-modal weighted graph approach

Dong Zhao [a,*], Adriana-Simona Mihăiţă [a], Yuming Ou [a], Hanna Grzybowska [b,c], Mo Li [d]

[a] University of Technology Sydney, 61 Broadway Str, Sydney, NSW, Australia
[b] Data61, CSIRO, Eveleigh, NSW 2015, Australia
[c] School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia
[d] Nanyang Technological University (NTU), Singapore

ARTICLE INFO

Keywords:
Gravity Model
Deterrence function calibration
Multi-modal PT
Dynamic origin–destination matrices
Entropy
Smart card and GTFS data

ABSTRACT

Estimating the large-scale Origin–Destination (OD) matrices for multi-modal public transport (PT) in different cities can vary largely based on the network itself, what modes exist, and what traffic data is available. In this study, to overcome the issue of traffic data unavailability and effectively estimate the demand matrix, we employ several data sets like the total boarding and alighting, smart card as well as the General Transit Feed Specification (GTFS) in order to capture the PT dynamic patronage patterns.

First, we propose a new method to model the dynamic large-scale stop-by-stop OD matrix for PT networks by developing a new enhancement of the Gravity Model via graph theory and Shannon's entropy. Second, we introduce a method entitled "Entropy-weighted Ensemble Cost Features" that incorporates diverse sources of costs extracted from traffic states and the topological information in the network, scaled appropriately. Last, we compare the efficiency of a single travel cost versus various combinations of travel costs when using traditional methods like the Traverse Searching and the Hyman's method, alongside our proposed "Entropy-weighted" method; we demonstrate the advantages of using topological features as travel costs and prove that our method, coupled with multi-modal PT OD matrix modelling, is superior to traditional methods in improving estimation accuracy, as evidenced by lower MAE, MAPE and RMSE, and reducing computing time.

1. Introduction

1.1. Background and motivation

Public Transport Origin–Destination matrices (PT OD), also known as trip matrices, serve as a foundational element for traffic demand estimation, providing insights into how the travel demand is distributed in space and time. This information is crucial for planning, designing and managing transportation systems in a way that is responsive to the actual needs and behaviours of the population. Therefore, several techniques for trip matrix estimation have been attracting scientific attention since the last century (Balcombe et al., 2004).

The choice of what OD matrix estimation model should be adopted largely depends on the data availability. Smart card data (Hussain et al., 2021), mobile phone tracking (Ge and Fukuda, 2016; Wismans et al., 2018) or Bluetooth and Wi-Fi tracking

* Corresponding author.
E-mail address: Dong.Zhao@student.uts.edu.au (D. Zhao).
URL: https://www.fmlab.org (D. Zhao).

https://doi.org/10.1016/j.trc.2024.104694
Received 29 November 2022; Received in revised form 30 May 2024; Accepted 31 May 2024
Available online 1 July 2024

data (Ta et al., 2018; Huang et al., 2022) provide comprehensive information on each trip from an origin stop to a destination stop following a certain PT route, which enables it to be one of the ideal data resources for PT demand modelling. However, such extensive data may not be accessible in many cities. In cases where full trip matrix data are unavailable, the total trip generation and attraction (GA) data are often derived from census data or surveys (Tamblay et al., 2016). Alternatively, aggregated data from smart card transactions is often utilised, as demonstrated in this paper for practical considerations.

Given the total trip GA from smart card transactions, there is a need for an estimation model to derive the OD matrix. The Gravity Model is a conventional modelling approach suitable for this purpose (Ortuzar et al., 2011). Originating from the law of gravity, this model was first introduced by Reilly (1931) to delineate the connection between trip distribution and microscopic zonal demand (GA). The accuracy of the estimated OD matrix hinges on the model's calibration; therefore, subsequent to the introduction of the Gravity Model, numerous endeavours have been made to enhance the calibration of this model.

Past studies on the Gravity Model have two main categories for model calibration: (a) the first involves the calibration of balancing factors, which are weights against the total generation and attraction in the model. These factors aim to ensure that the model accurately reflects the real-world travel volume; (b) the second category focuses on calibrating the deterrence function, which is a model that represents the resistance or travel cost associated with travelling between OD pairs. By calibrating this function, transportation planners can consider various factors influencing travel behaviour, such as travel time, distance, and speed. However, due to limitations in the travel cost data, different calibration methods have been proposed to compensate for biases in the travel cost features, reflect travel preferences, and enhance the model performance. This paper delves into the relationship between the form of the deterrence function and the travel cost features, aiming to deepen our understanding of this crucial aspect of the model calibration.

In prior research studies, the deterrence function was typically associated with a single type of travel cost due to the challenges of determining the optimal weightings for multiple cost features (Tamblay et al., 2016; Suprayitno, 2018). Some studies incorporated multiple travel costs but without assigning weights to each cost feature (Rubio-Herrero and Muñuzuri, 2021). Expanding upon previous research, in this research paper, we introduce a new innovative approach to rank and assign weights to various travel cost features more efficiently. By incorporating multiple travel cost features into the Gravity Model, we aim to significantly enhance the model's overall performance and show its efficiency as compared to more traditional techniques.

1.2. Challenges and contributions summary

In order to obtain accurate public-transport-related OD matrices that are multi-modal, we encounter several challenges (open questions):

- How to dynamically model the spatial and temporal travel patterns of PT passengers?
- How to effectively obtain accurate OD matrices for multi-modal PT networks using a minimum data availability?
- What factors have a critical influence on the PT OD estimation across various types or modes of the PT networks?

To address these challenges, in our current work we propose:

- a new framework to dynamically estimate the stop-by-stop PT OD matrix for multiple transport modes;
- an extended version of the Gravity Model for our multi-modal PT OD estimation. The model relies on a novel model calibration method using Entropy-weighted, which considers both the traffic characteristics (travel time, travel distance, and fare cost) as well as the graph topological features (connections, closeness and straightness);
- and finally, we demonstrate the advantages of using topological features as travel costs and show that our new proposed PT OD calibration method outperforms the traditional approach, namely Hayman's and Traverse Searching methods, in terms of RMSE, MAPE and MAE.

1.3. Organisation of the paper

The rest of this paper is organised as follows: a literature review regarding PT OD estimation, the Gravity Model and model calibration is presented in Section 2, in Section 3 we first present the framework of the large-scale stop-by-stop OD estimation for PT. Following the framework, the details of the extended Gravity Model are further highlighted in Section 3.2, where we show the modelling method for aggregated GA vectors (see Section 3.2.2) and deterrence function considering "single travel cost feature" and "fused travel cost features". Then, the classical deterrence function calibration processes are discussed in Section 3.2.5 against the proposed method of "Entropy-weighted" in Section 3.2.6. Finally, the topological features' modelling methods are included in Section 3.2.7. The performance of an OD estimation is validated using MAE, RMSE and MAPE, shown in Section 3.3. The application of the estimation methods involving the three deterrence function calibration methods to a real network is presented in Section 4 with results shown in Section 4.4, where The ideal parameter configurations for the deterrence function and the most appropriate form of the function tailored to various travel cost features are presented in this section. Finally, Section 5 is provided to clarify the research limitations and offer future directions in this field.

2. Literature review

2.0.1. PT OD matrix estimation by the gravity model

According to the literature, historical stop-based passenger trajectory data (such as smart card, mobile phone or Bluetooth tracking data) in some areas can be hard to obtain. Given the prevailing circumstances, it is widely acknowledged that obtaining data on the total departure and arrival from PT stops, or zones can be achieved through efforts of conducting census or field surveys. In certain locations, aggregated data presenting the PT stop-based GA vector can be available due to safety concerns when using smart card data. With this amount of data, the Gravity Model is one of the most suitable approaches. The Gravity Model is derived from the law of gravity and was introduced in 1931 by Reilly to capture the relationship between towns and retail trade. Such application of the gravitation law triggered much research to focus on the trip estimation between a pair of nodes, as introduced by Dieter, who built a model based on the gravity law and indicated that the number of trips between two areas was proportional to the total number of trips at origin and destination, which was the total number of production and attraction, and revering proportion to the distance between groups. Considering to solve the Gravity Model, Bouchard and Pyers and Ben et al. treated the model as a singly constrained model, and only the zonal attraction parameter was considered to ensure that the modelled number of trips attracted by a zone matches the survey result. The main drawback of using such an abridged constraint was that the model could hardly ensure that the calculated total trips between an OD equal the sum of the trip departing and trip arriving. This limitation encouraged the practice of the doubly constrained Gravity Model, where both parameters for trip origins and trip destinations are considered (see Evans (1973), Ortuzar et al. (2011), Abdel-Aal (2014), Pitombo et al. (2017) and Rubio-Herrero and Muñuzuri (2021)). In this paper, we adopt a doubly-constrained Gravity Model introduced by Ortuzar et al., and a detailed modelling method is presented in Section 3.

Since its appearance in the scientific community, the Gravity Model has maintained its attractiveness to researchers due to its ease of computing and convergence; only the total generation and attraction data and the travel cost data are required when estimating the OD matrix using the Gravity Model. In reality, travellers tend to avoid expensive routes; the deterrent effect of travel costs in the model reflects this phenomenon. Such travel costs are generally derived from the traffic features, such as utility (see Sun et al. (2021)), travel distance (see Tamblay et al. (2018) and Thompson et al. (2019)) or time (see Delgado and Bonnel (2016) and Zhao et al. (2022)). However, for PT users, the dynamic travel features generated from the network structure and operation plans also matter, as they mirror the performance of the PT services and the level of convenience for travellers as well. In previous literature, many studies considered the graph theory as a tool for network vulnerability analysis (see Chen et al. (2021), Wang et al. (2021) and Zhang et al. (2021)), which can also be used when modelling the deterrent effect on network performance considering network structures. Following this, one of the original contributions of this work is that instead of using the traditional travel cost features in cost matrix estimation, we propose a boosting of the Gravity Model via graph theory and Shannon's entropy targeting the large-scale PT networks to embrace the impact of the network structure on the travel behaviour.

2.0.2. Gravity model calibration

As mentioned, when estimating the OD matrix for PT by using the Gravity Model, the travel cost which influences the number of trips is vital. Such travel cost is also known as the friction factor (see Evans (1973)), friction or cost matrix when considering the independent friction by OD pairs (see Wilson (2013)). The travel cost has been found to perform better by fitting it to a deterrence function — the most classical being exponential, power and Tanner functions (as summarised by Ortuzar et al., Wilson).

When addressing the deterrence function, the process of searching for function parameters becomes crucial. If a reliable travel cost matrix is readily available, such information can be directly utilised without calibration. However, when such data is unavailable, additional parameters are necessary to ensure the accuracy of calibrating the Gravity Model. An early approach involves a Traverse Searching, wherein various parameters, typically ranging between 0 and 1, are explored during the estimation process using the Gravity Model. The findings of Feldman et al. expanded the scope of optional parameters to include negative values, and alternative deterrence function forms beyond the Tanner function were examined, such as the Lognormal, WebTag Cost Damping and Low/High-cost functions. Feldman et al. concentrated on three conventional deterrence function forms and evaluated parameter performance within the range of 0 to ± 1.2. Rubio-Herrero and Muñuzuri further investigated additional function forms, including Weibull, Box–Cox, March, and Richards, setting the initial parameter range between 0 and 0.1. Drawing from the author's experience, parameter ranges were expanded to $[0, 10]$ accordingly, with the exception of the Richards function form, for which parameters were selected from $[0, 500]$. The performance of these parameters is then evaluated based on the accuracy between the real and estimated OD matrix or trip length distribution if the historical OD matrix is unknown. The traverse search method focusing on three traditional methods, namely, the power, exponential and Tanner functions are employed in this paper for comparison purposes.

A technique introduced by Hyman has proven to be more effective due to its fast convergence; such a method initialises the parameter by:

$$\beta_0 = \frac{3}{2\bar{C}} \tag{1}$$

where β_0 is the parameter of the deterrence function and $\bar{C}$ is the mean travel cost.

In a new cycle of calibration, the parameter is found by:

$$\beta_\varsigma = \frac{\beta_0 c_0}{\bar{C}} \tag{2}$$

where c_0 is the mean estimated travel cost.

For the following run of the calibration with $\zeta \geqslant 1$, the parameter is calculated by:

$$\beta_{\zeta+1} = \frac{(\bar{C} - \bar{C}_{\zeta-1})\beta_{\zeta} - (\bar{C} - \bar{C}_{\zeta})\beta_{\zeta-1}}{\bar{C}_{\zeta} - \bar{C}_{\zeta-1}} \tag{3}$$

where c_0 is the mean estimated travel cost. Hyman's method is investigated in this paper for comparison purposes, as well.

The selection of travel cost elements is another impact factor that determines the performance of estimation, where the most prevalent factor in the literature comes to the travel distance (see Simini et al. (2012), Shen and Aydin (2014), Thompson et al. (2019), He and Chow (2021) and Rubio-Herrero and Muñuzuri (2021)) followed by the travel time (see Feldman et al. (2012) and Delgado and Bonnel (2016)). In practice, the number of trips that passengers make on PT also depends on features such as fare cost, waiting time, transfer time or the level of occupancy in the PT (see Van Acker et al. (2021)). In the existing literature, travel cost features primarily focus on traffic-related factors. Only one study, conducted by Rubio-Herrero and Muñuzuri, incorporated fuel, driver, tyres, and maintenance costs into OD estimation for freight networks. However, this study did not assign weights to each of these travel cost features. The inclusion of multiple travel cost features presents another calibration challenge: determining the weight assigned to each feature. To address this gap, we introduced an entropy-weighting method to determine the weight for each travel cost feature effectively. This enables the model to derive maximum benefit from multiple travel cost features, thereby enhancing overall accuracy.

In our work presented in this paper, we also consider, in addition to the traffic states, that the trip distribution is also associated with the network topology because the geometric properties can affect the node and the link capacity, as well as the passenger accessibility in the network; this can further influence the link travel capacity and the travel efficiency. Especially when considering a dynamic cost matrix, the pre-defined timetable can interfere with timely access for travellers in specific locations. Therefore, this intuition also raises our initiative to employ the topological features for the cost matrix estimation, as this will reflect the network accessibility and the inter-modality between all modes inside an interconnected PT graph. The graph theory-oriented features have not been explored extensively in the past, the exceptions being some recent studies regarding network vulnerability analysis. However, only one recent study has explored the potential of graph theory by separately using the betweenness and the travel time as travel costs (see Lu (2018)). Furthermore, by using the Gravity Model, we conduct a comparative analysis to assess the efficacy of various travel cost features in OD estimation. This evaluation encompasses the utilisation of both traditional Traverse Searching and Hyman's method, as well as our proposed entropy-weighting method for calibrating the deterrence function.

2.0.3. Calibrating by entropy ranking

In line with the method of weighting by the friction factor, the entropy measurement has the potential to estimate the friction factor accurately (see Ai (2017)). Entropy was initially proposed in the information theory as a measure of the uncertainty or randomness in a system, it can be used in order to quantify the degree of variability or diversity for each travel cost feature. The entropy measures the average amount of information required to draw an outcome from a probability distribution. Therefore, by ranking the travel cost features based on their entropy, it is possible for us to identify the factors that are most important in shaping travel behaviour. Once we have ranked the travel cost features by entropy, we can integrate various impacts into one index, which can be used to calibrate travel patterns in a Gravity Model. This approach allows us to capture the overall impact of different travel cost factors on the travelling pattern and to develop more accurate and reliable models for estimating travel demand. So far, the entropy measure has only been used in ranking or for evaluation purposes in the current literature (see Ai (2017), Qi et al. (2021), Zhang and Ng (2021) and Wei et al. (2022)). To the best of our knowledge, no publication indicates the feasibility of using entropy to rank the travel cost features and the advantage of using the entropy-weighted cost matrix in the OD estimation.

To fill the gap in PT OD estimation, our paper introduces a framework for dynamically estimating the stop-by-stop OD matrix for large-scale PT networks. We enhance the traditional Gravity Model by incorporating the entropy weighting on travel cost features. Our study investigates the influence of various PT network travel cost features on OD estimation, encompassing traffic characteristics such as travel time, travel distance and fare cost, as well as graph topological features such as connections, closeness and straightness. We evaluate the performance of different forms of deterrence function used in the Gravity Model, and we compare traditional calibration methods like Traverse searching and Hyman's method against our proposed Entropy-weighted method. The results demonstrate significant improvements in terms of RMSE, MAPE and MAE.

3. Methodology

3.1. Modelling framework

The framework presented in Fig. 1 illustrates our approach to dynamically estimating the stop-by-stop OD matrix. This framework comprises three stages. In *Stage 0*, we collect, filter, and clean all input datasets, including smart card data and Global Transit Feed Systems (GTFS) data. At this stage, we also integrated total counts of boarding and alighting. This integrated data serves as an ideal alternative for GA vector generation in *Stage 1*; *Stage 2* illustrates the deterrence function calibration process considering three types of data resources: the "Deterrence function considering single travel cost features", "Deterrence function considering fused travel cost feature" and "Deterrence function considering Entropy-weighted fused travel cost features". When it comes to individual travel cost features, we employ two calibration methods, namely Traverse Searching and Hyman's methods. Conversely,

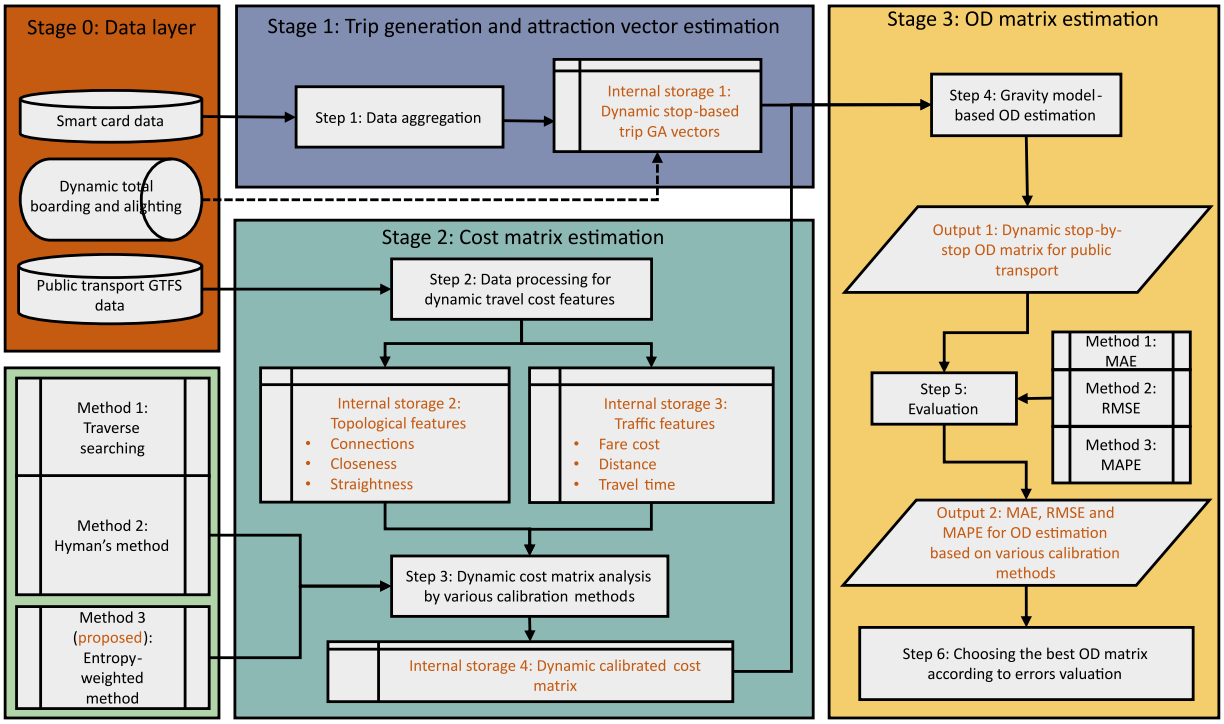


Fig. 1. Framework of our proposed dynamic stop-by-stop OD Estimation for a large-scale multi-modal PT network.

when dealing with multiple costs, we adopt a fusion approach by utilising the Hyman method alongside our proposed Entropy-weighted method to demonstrate the effectiveness of OD estimation. Our proposed methodology leverages entropy ranking to assess and weigh various travel cost features, aiming to optimise the accuracy of the deterrence function and, consequently, OD estimation. For each calibration process, we evaluate the efficacy of three function forms: power, exponential, and Tanner; at the *Stage 3*, we employ the Gravity Model for the OD estimation and further validate the feasibility of the proposed deterrence function calibration method. The best cost matrix estimation method and the most accurate OD matrix are selected at this final stage according to MAE, RMSE and MAPE.

3.2. PT OD estimation using the gravity model

3.2.1. The gravity model

For PT such as buses or trains, the unit of its OD matrix is defined as the number of passengers, and the content of the OD matrix is the total number of passenger trips. Unlike the OD matrix for cars, the origins and destinations for the PT network are the PT stops.

Adjacency Matrix: The existence of trips between each OD pair depends on the predefined paths for routing the PT in the network. Therefore, the PT network is defined as a Space L' graph, where the nodes are represented by the PT stops, and the links between nodes are the routes between any PT stops. In a graph of Space L', different edges represent different links used in different networks with different directions (Yang et al., 2014). To capture the nature of dynamic timetables, such information is recorded in a time-dependent adjacency matrix:

$$H(t) = [h_{m_i,n_j}(t)], i, j = \{1 \dots J\}, m, n = \{1 \dots N\} \tag{4}$$

where J is the total number of statistical areas inside the sub-network; N represents the total number of PT stops in the network; and

$$h_{i,j}(t) = \begin{cases} 1, & \text{if } (i,j) \text{ is an accessible link at t,} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The Gravity Model: The number of time-dependent trips made by a public transit mode pt is obtained based on the placement and existing routing between PT stops, where pt is considered as buses (b), train and metro (tn). Therefore, the total number of trips departing from an origin stop $m \in \{1 \dots N\}$ to a destination stop $n \in \{1 \dots N\}$ can be calculated based on the Gravity Model with a given total number of trips departing from the origin stop m by a mode pt, denoted by $o^{pt}_{m_i}$, and that of trips arriving at a destination

stop n by mode pt, denoted by $d^{pt}_{n_j}$. The PT trip matrix can be represented by $OD_{pt}(t) = [od^{pt}_{m_i,n_j}(t)], i,j = \{1 \dots J\}, m,n = \{1 \dots N\}$. The total number of trips departing is also known as the trip generation. In contrast, the total number of trips arriving is known as the trip attraction in the traditional four-step OD estimation. The most common form of the Gravity Model is expressed as:

$$od^{pt}_{m_i,n_j}(t) = A^{pt}_{m_i}(t)\ o^{pt}_{m_i}(t) B^{pt}_{n_j}(t) d^{pt}_{n_j}(t) f\left(c^{pt}_{m_i,n_j}(t)\right) \tag{6}$$

where $A^{pt}_{m_i}$ and $B^{pt}_{n_j}$ represent the time-dependent weights towards the total number of origins ($o^{pt}_{m_i}$) by PT and the total destinations ($d^{pt}_{n_j}$), respectively; $A^{pt}_{m_i}$ and $B^{pt}_{n_j}$ also known as the balancing factors. $f\left(c^{pt}_{m_i,n_j}\right)$ is the travel cost function that represents the time-dependent travel cost between two zones by a mode pt. Three cost matrix calibration methods are compared in this paper, namely Traverse searching, Hyman's method and our proposed Entropy-weighted method, as detailed in Sections 3.2.5 and 3.2.6. The origin stop m belongs to the origin zone i while the destination stop n belongs to the arrival zone j.

The constraints: Two constraints are employed in the Gravity Model to enable the estimated trips to match the real number of trips, two constraints are employed in the Gravity Model. Firstly, the total number of departures by a PT mode from a public stop m should be equal to the sum of trips that originate from that particular stop to each possible destination stop n:

$$o^{pt}_{m_i}(t) = \sum_{j=1}^{N} od^{pt}_{m_i,n_j}(t) \tag{7}$$

and secondly, the total number of trips taken by PT arriving at a stop n at the time interval t equals the sum of trips that terminate at that particular destination from all possible origin stops m:

$$d^{pt}_{n_j}(t) = \sum_{i=1}^{N} od^{pt}_{m_i,n_j}(t) \tag{8}$$

The parameters: The time-dependent weights ($A^{pt}_{m_i}$ and $B^{pt}_{n_j}$) are the parameters of the Gravity Model that are estimated iteratively. The estimation equations can be transformed from Eqs. (6) and (7) to:

$$A^{pt}_{m_i}(t) = \frac{1}{\sum_{j=1}^{N} B^{pt}_{n_j}(t) d^{pt}_{n_j}(t) f\left(c^{pt}_{m_i,n_j}(t)\right)} \tag{9}$$

and from Eqs. (6) and (8) to:

$$B^{pt}_{n_j}(t) = \frac{1}{\sum_{i=1}^{N} A^{pt}_{m_i}(t) o^{pt}_{m_i}(t) f\left(c^{pt}_{m_i,n_j}(t)\right)} \tag{10}$$

The criterion: The criterion of the convergence follows either the maximum number of iterations that have been reached:

$$\zeta \in \{1,\ \dots, \zeta_{max}\} \tag{11}$$

or the functions of acceptable distance between iterative number ζ and $\zeta + 1$:

$$cc^{pt}\ \geq \max_{i,j}\left(\max_{i}\left(\frac{A^{pt,\zeta+1}_{m,i} - A^{pt,\zeta}_{m,i}}{A^{pt,\zeta+1}_{m,i}}\right), \max_{i}\left(\frac{B^{pt,\zeta+1}_{n,j} - B^{pt,\zeta}_{n,j}}{B^{pt,\zeta+1}_{n,j}}\right)\right) \tag{12}$$

where cc^{pt} is the criteria of convergence that defines the acceptable distance of the last two time-dependent weights, and ζ is the count of iterative calculations. The ζ_{max} used in this research is defined as shown in Assumptions (see Section 4.2).

3.2.2. GA vector estimation

The trip generation and attraction (GA) estimation is known as the first step in the traditional four-step demand estimation. This step aims to produce a GA vector that can be used as the total number of departing trips ($d^{pt}_{n_j}$) and total arriving trips ($o^{pt}_{m_i}$) in the Gravity Model-based OD matrix estimation. As introduced in Ortuzar et al. (2011), in practice, the GA vector is initially obtained from the demographic data through a regression analysis. For our research study, the total generation from a stop is the total number of departing trips; thus, the vector of generation is the total number of tap-ons at each stop. Meanwhile, the total attraction of a stop is the total number of arriving trips; thus, the vector of attraction is the total number of tap-offs at each stop. Therefore, the GA vector can be captured from historical smart card tap-on/tap-off data as follows:

$$o^{pt}_{m_i}(t) = \sum_{j=1}^{N} od^{pt,historical}_{m_i,n_j}(t), \tag{13}$$

$$d^{pt}_{n_j}(t) = \sum_{i=1}^{N} od^{pt,historical}_{m_i,n_j}(t) \tag{14}$$

and the GA vector can be represented as $GA^{pt}_{m_i,n_j} = [o^{pt}_{m_i}, d^{pt}_{n_j}]$.

The paper aims to demonstrate the effectiveness of using the Gravity Model for estimating the PT OD matrix with the most accessible data: total generation and attraction and travel cost data. The data provided for this research study is stop-based smart card data, we process it as stop-based GA vectors. The travel cost data obtained from both traffic states and topological graphs is

calibrated following the deterrence function-based Gravity Model by the Traverse searching method (Ortuzar et al., 2011), Hyman's method (Hyman, 1969) and the new proposed Entropy-weighted method. We compare the performance of travel cost features following "single travel cost feature", "fused travel cost features" and the "entropy-weighted fused travel cost features".

3.2.3. Deterrence function considering single travel cost features

In the OD estimation method via the Gravity Model, the deterrence function, $f\left(c^{pt}_{m_i,n_j}\right)$, is the negative function limiting the number of trips generated from origins to destinations in a network. Following three classical forms of the deterrence function, the friction elements can be described as:

Power function:

$$f\left(c^{pt}_{m_i,n_j}\right) = \gamma(c^{r,pt}_{m_i,n_j})^{\alpha} \tag{15}$$

Exponential function:

$$f\left(c^{pt}_{m_i,n_j}\right) = \gamma e^{-\beta c^{r,pt}_{m_i,n_j}} \tag{16}$$

Tanner function:

$$f\left(c^{pt}_{m_i,n_j}\right) = \gamma(c^{r,pt}_{m_i,n_j})^{\alpha}\ e^{-\beta c^{r,pt}_{m_i,n_j}} \tag{17}$$

where γ is the weight of the travel cost, in the case study of this paper, $\gamma = 1$, in order to examine the performance of raw travel cost features in OD estimation. $i, j = \{1 \dots J\}, m, n = \{1 \dots N\}$; α and β are parameters of the deterrence function that are required to be estimated.

3.2.4. Deterrence function considering fused travel cost feature

However, in cases where travel costs such as travel time and connection are used in our situation, due to the property of non-linearity with respect to travel distance, a simple summation, as employed in Rubio-Herrero and Muñuzuri (2021), is not suitable. Therefore, after fitting using a deterrence function(Eq. (15), Eq. (16) or Eq. (17)), we standardise the travel cost matrix by dividing it by the mean fitted travel cost ($\overline{C^r}$). This normalisation ensures that all travel cost features have a consistent mean value of 1. Consequently, we can aggregate all travel cost matrices to create a fused travel cost matrix that encompasses the effects of multiple types of travel costs. The fusion travel cost can be represented by:

$$f\left(c^{pt}_{m_i,n_j}\right) = \sum_{i=1}^{R} \frac{f\left(c^{r}_{m_i,n_j}\right)}{\overline{C^r}} \tag{18}$$

3.2.5. Deterrence function calibration

The calibration of the deterrence function is significant as it directly impacts the subsequent performance of OD estimation. In our effort to verify this idea and determine the most suitable calibration approach for our specific case study, we conduct the calibration analysis involving Hyman's method.

Deterrence function calibration by the Hyman method: Following Eq. (1), Eq. (2) and Eq. (3) presented in Section 2.0.2, we engage in an iterative adjustment of the parameters α and β in three common forms of the deterrence function: power (request a calibration on α in Eq. (15)), exponential (request a calibration on β in Eq. (16)), and Tanner (request a calibration on both α and β in Eq. (17)). The calibrated deterrence function, which accurately models travel cost and influences trip distribution, will be utilised in Eq. (6) for iteratively estimating the OD matrix until the specified objective is achieved (as defined in Eq. (11) and Eq. (12)).

According to Hyman (1969) and Williams (1976), the solution of Eqs. (7) and (8) follows an iterative calculation process. Since the mean observed travel cost should equal the estimated one, we introduce another constraint:

$$\sum_{i,j}^{J} od_{i,j} c_{i,j} = \sum_{i,j}^{J} od^{*}_{i,j} c_{i,j} \tag{19}$$

where the exponential function is given as the deterrence function. For the power function, the constraint should be defined as:

$$\sum_{i,j}^{J} od_{i,j} \log c_{i,j} = \sum_{i,j}^{J} od^{*}_{i,j} \log c_{i,j} \tag{20}$$

for Tanner function, both Eqs. (19) and (20) need to be considered.

To adjust the travel cost matrix using the deterrence function, it requires an initial β_0; according to Hyman (1969),

Step 0:

$$\beta^0 = \frac{3}{2\overline{C}} \tag{21}$$

where the mean observed travel cost $\overline{C}$ is defined as:

$$\overline{C} = \frac{1}{J^2} \sum_{i,j}^{J} c_{i,j} \tag{22}$$

Step 1: Therefore, the initial friction factor matrix can be converted from the cost matrix to Eq. (15), Eq. (16) or Eq. (17). According to Ortuzar et al. (2011), the initial trip matrix can be obtained by:

$$od_{i,j}^{0} = \frac{\sum_{i,j}^{J} f\left(c_{i,j}\right) o_i}{\sum_{i,j}^{J} f\left(c_{i,j}\right)} \tag{23}$$

and the time-dependent weights ($A_{m_i}^{pt}$ in Eq. (9) and $B_{n_j}^{pt}$) in Eq. (10) can be updated accordingly.

Step 2: With the base matrix, we can update the initial mean estimated travel cost, following the method proposed by Williams (1976):

$$C^{0} = \frac{\sum_{i,j}^{J} f\left(c_{i,j}\right) o_i d_j}{\sum_{i,j}^{J} f\left(od_{i,j}\right)^2} \tag{24}$$

and the mean observed travel cost becomes:

$$\overline{C^{s}} = \frac{\sum_{i,j}^{J} od_{i,j}^{s} c_{i,j}}{\sum_{i,j}^{J} od_{i,j}} \tag{25}$$

where s represents the iterative number in gravity estimation.

Step 3: According to Hyman (1969), the parameter for the deterrence function can be computed by:

$$\beta^{1} = \frac{\overline{C^{0}} * \beta^{0}}{\overline{C^{0}}} \tag{26}$$

then the cost matrix can be re-adjust using the β^1, returning to *Step 1*. However, for $s \geq 1$, the discrepancy of deterrence factor between two iterations should be proportional to that of mean travel costs to avoid calculation error according to Williams (1976); therefore:

$$\beta^{s+1} = \beta^{s} + \frac{(\bar{C} - C^{s})(\beta^{s} - \beta^{s-1})}{C^{s} - C^{s-1}} \tag{27}$$

Deterrence function calibration by the Traverse Searching method: The decision to employ the traverse method for determining optimal parameters in research depends on the complexity of the problem and the characteristics of the dataset (Suprayitno, 2018). In light of our experience and the outcomes of applying the Hyman method, we observed that the parameters' impact on the OD estimation can be unpredictable. Hence, we conducted a traverse sensitivity test to explore the optimal parameter that best estimates the OD matrix. The range of parameters used in this research is defined in Assumptions (see Section 4.2).

Based on the outcomes derived from Hyman's methodology, we confirm the most suitable form of the deterrence function for fitting purposes. This selected function is then utilised for Traverse Searching. Subsequently, guided by the suggested parameters of the deterrence function following Hyman's method, we define a range of potential parameters for exploration. Within this range, our objective is to identify the optimal set of parameters that enhance estimation accuracy to a greater extent.

3.2.6. Deterrence function considering entropy-weighted fused travel cost features

Assigning weights to travel cost features ensures their appropriate impact on OD estimation, thereby enhancing the accuracy of the estimation. However, weighting requires considerable information, which has historically posed a challenge. Our results from testing different travel cost features individually in the deterrence function (as discussed in Section 3.2.3) reveal that various travel costs perform differently when fitted to the same deterrence functions, which underscores the significance of selecting suitable travel cost features in OD estimation.

Our proposed Entropy-weighted method is the initiative to consider weighing various travel cost features to maximise the accuracy of OD estimation. Such a feature resemble method is derived from the principles of Shannon's entropy (Shannon, 1948; McClean, 2003) which has been used for feature ranking (see Nie et al. (2016), Ai (2017), Qi et al. (2021) and Wei et al. (2022)). In this way, our current research study takes a further step and applies entropy when weighing various travel costs and evaluating their true importance for the final trip demand estimation process.

In the following, we detail our proposed weighting method via Shannon's entropy:

Network representation: The transport network is captured by an unweighted non-directed graph which we denote $G = (V, E)$, and which follows the Space L' representation. The set of vertices is represented by $V(G) = \{v_1, v_2, \ldots, v_N\}$, where N is the total number of PT stops, while the elements of E are the edges following $E(G) = \{e_1, e_2, \ldots, e_l\}$. For a PT mode pt, let G^{pt} be the graph where V^{pt} is the set of vertices, and E^{pt} is the set of edges.

Travel cost feature representation: Each network has its unique topological feature reflected by centrality and global characteristics (see Lin and Ban (2013)). Currently, the most used centrality measures in the literature are connection, closeness and network straightness. The details of these topological features are described in Section 3.2.7. In this research, the travel cost feature representations also include the traffic features such as the fare costs, the travel distance and time. Therefore, the travel cost features of the graph G include both topological and traffic features, which are further represented by $C(G) = \{c_1, c_2, \ldots, c_R\}$, where the total number of travel cost features is R.

According to the graph features, the matrix following travel cost features by nodes ($c_{m,r}$) can be expressed as:

$$S = \begin{bmatrix} s(v_1, c_1) & s(v_1, c_2) & \dots & s(v_1, c_R) \\ \dots & \dots & \dots & \dots \\ s(v_N, c_1) & s(v_N, c_2) & \dots & s(v_N, c_R) \end{bmatrix} \tag{28}$$

where for each value of the cell, s is the travel cost value defined by the location, which is the PT stop v (ordered as $1 \dots N$), and the travel cost feature c, either topological or traffic features (represented by $1 \dots R$). In this way, for each row in this matrix, the row number represents the stop name; for each column, the column name represents either topological or traffic features. Therefore, the $s(v_1, c_1)$ means the value of the first travel cost feature for the first stop, and $s(v_1, c_2)$ is the value of the second type of travel cost feature for the first stop. To simplify computation, we convert origin–destination travel costs (c_{mn}), like travel time, into one-dimensional factors. Each value is assigned to its origin stop (c_m), aligning its dimension with other one-dimensional features such as closeness. We employ ($s_{m,r}$) to represent the origin stop-based travel cost.

Standardised travel cost matrix: To standardise a feature r for each node (PT stop), the ratio is estimated by using the mathematical formula below:

$$u_{m,r} = \frac{s_{m,r} - min(s_{m,r})}{max(s_{m,r}) - min(s_{m,r})} \tag{29}$$

The standardisation technique described is called max–min scaling, also known as max–min normalisation. Through normalisation, each feature contributes to the final distance roughly in proportion to its range. The scaling range following max–min scaling is set to [0, 1].

Thus, the standardised travel cost matrix is denoted as:

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,R} \\ \dots & \dots & \dots & \dots \\ u_{N,1} & u_{N,2} & \dots & u_{N,R} \end{bmatrix} \tag{30}$$

Entropy measure: According to Shannon's entropy (see Shannon (1948), McClean (2003) and Ai (2017)), the ratio of each standardised travel cost feature $u_{m,r}$ is denoted by impact probability $p_{m,r}$, where:

$$p_{m,r} = \frac{u_{m,r}}{\sum_{r=1}^{R} u_{m,r}} \tag{31}$$

which helps us to estimate further the entropy of each travel cost measure, which is denoted by:

$$I_r = -\frac{1}{N} \sum_{m=1}^{N} p_{m,r} \log(p_{m,r}) \tag{32}$$

We apply the max–min scaling again to standardise the weights, as can be expressed by:

$$w_r = \frac{I_r - min(I_r)}{max(I_r) - min(I_r)} \tag{33}$$

Considering that the weights are entropy-based factors representing uncertainty, where features with higher entropy exhibit greater uncertainty and unpredictability, we aim to prioritise features that are stable and predictable in OD estimation. Therefore, we calculate the importance of each feature as $1 - w_r$. Therefore, according to Eq. (18), the scaled Entropy-weighted ensemble cost features becomes:

$$f\left(c^{pt}_{m_i,n_j}\right) = \sum_{i=1}^{R} \frac{(1 - w_r) f\left(c^{r}_{m_i,n_j}\right)}{(1 - w_r) f\left(c^{r}_{m_i,n_j}\right)} \tag{34}$$

where each importance factor is assigned to the travel cost feature accordingly.

Once the weighted travel cost features are constructed, they form a calibrated friction matrix that can be directly applied in the Gravity Model. To simplify the calibration process, we set the values of alpha and beta in the deterrence function to 1. This Entropy-weighted method eliminates the need to calibrate the deterrence function, resulting in significant time savings. Consequently, utilising the Gravity Model to explore the optimal balancing factors ($A^{pt}_{m_i}$ and $B^{pt}_{n_j}$ in Eq. (6)) that maximises estimation accuracy, intending to minimise disparities between the estimated and historical OD matrices, as demonstrated in this paper, or trip length distributions, as utilised in prior literature.

The code for the method of Entropy-weighting can be found by the link: https://github.com/Future-Mobility-Lab/Entropy-weighting-method.

3.2.7. Topological features

In the discussion in Sections 3.2.4 and 3.2.6, we include multiple travel cost features; in this section, we discuss those features in detail. Apart from the connection derived from the adjacency matrix, additional topological cost features captured from the network graph, including closeness and straightness, are also selected to enhance the accuracy of OD matrix estimation in this study. These features capture attributes related to travel distance that influence passengers' route choices and, consequently, impact

trip distribution. The definition of each feature is expressed through the following equations, the same as described in Lin and Ban (2013):

Connection: property indicates the number of edges connected to a node, also known as degree in graph theory, which is determined based on an adjacency matrix as follows:

$$c_{m_i,n_j}^{cn,pt} = \sum_{m_i=1}^{J} \sum_{m_j=1}^{J} ed_{m_i,n_j} \tag{35}$$

Closeness: the characteristic defining the total travel distance from a given node to all other accessible nodes in the network and is expressed as:

$$c_{m_i,n_j}^{cl,pt} = \frac{1}{\sum_{m_j=1}^{J} d_{m_i,n_j}} \tag{36}$$

where d_{m_i,n_j} indicates the travel distance on predefined bus routes, which is the shortest path travelled by bus.

Straightness: the feature displaying the ratio of the Euclidean distance (d_{m_i,n_j}^{Eucl}) over the shortest travel distance following the bus routes.

$$c_{m_i,n_j}^{st,pt} = \sum_{m_j=1}^{J} \frac{d_{m_i,n_j}^{Eucl}}{d_{m_i,n_j}} \tag{37}$$

3.3. OD matrix evaluation

Assuming that the estimated OD matrix using our proposed approach is denoted as $[O\hat{D}_t]$, while the observed one is $[OD_t]$, then the OD estimation accuracy in this research is measured by using the following three performance metrics:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{N} \sum_{n=1}^{N} \left| [O\hat{D}_t] - [OD_t] \right| \tag{38}$$

where n represents the PT stop and N is the total number of stops in the network.

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left([OD_t] - [O\hat{D}_t] \right)^2} \tag{39}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{[OD_t] - [O\hat{D}_t]}{[O\hat{D}_t]} \right| \times 100\% \tag{40}$$

where t is the iteration number and T represents the total number of iteration calculation.

Computing time: To evaluate the effectiveness of the estimation, we also take into account the computing time required for iteratively estimating the OD matrix using the Gravity Model, which includes the time spent calibrating the deterrence function.

4. Case study

4.1. Geography and data information

The zones covered in the study are located to the Northwest of Sydney, along the M2 motorway, which includes several major residential and business areas, as shown in Fig. 2 upper-left. This area is defined following the digital mapping according to the Statistical Area Level 2 (SA2) (Australian Bureau of Statistics, 2021). The zones used by the car network are defined in the software Aimsun according to a Voronoi diagram (Xiao et al., 2016; Nikolić and Bierlaire, 2018), which is denoted as Z_j, $j \in \{1 \dots J\}$, as shown in Fig. 2 (up left). As of 2017, there are 76 bus routes, two train routes and a metro route spread within this area. There are 3799 bus stops and seven train and metro stations, as shown in Fig. 2 (down). To simplify notations, we will further refer to our case study as the M2 area in the following sections. The smart card data used for this paper includes records for both trains and metros. Consequently, this paper treats trains and metros as a single mode of transport, referred to as "train", throughout this study.

The trip distribution is captured from the local smart card data. The raw smart card data has been processed and filtered in advance to eliminate outliers and anomalies. The trip distribution in Fig. 3 is drawn by using as an example one month of smart card data (June of 2017) for the M2 area in Sydney. The upper figure in Fig. 3 illustrates the distribution for trips by bus, and the rest figure in Fig. 3 illustrates the distribution of train trips. As shown in Fig. 3, for both transport services, the morning peak hour starts from 7:00 and lasts until around 11:00. In contrast, the afternoon peak hour spreads from 16:00 to 20:00. In the case study

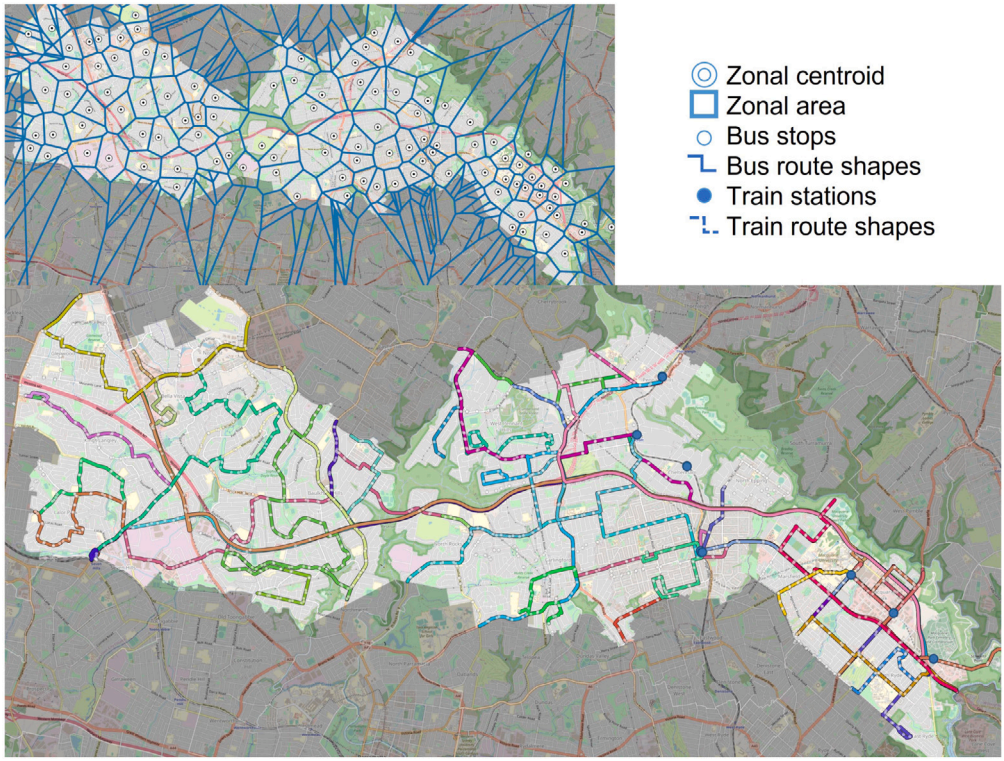


Fig. 2. Zones in M2 Area (up left) and the 2017 Sydney M2 Area PT, including Buses, Trains and Metro networks (down).

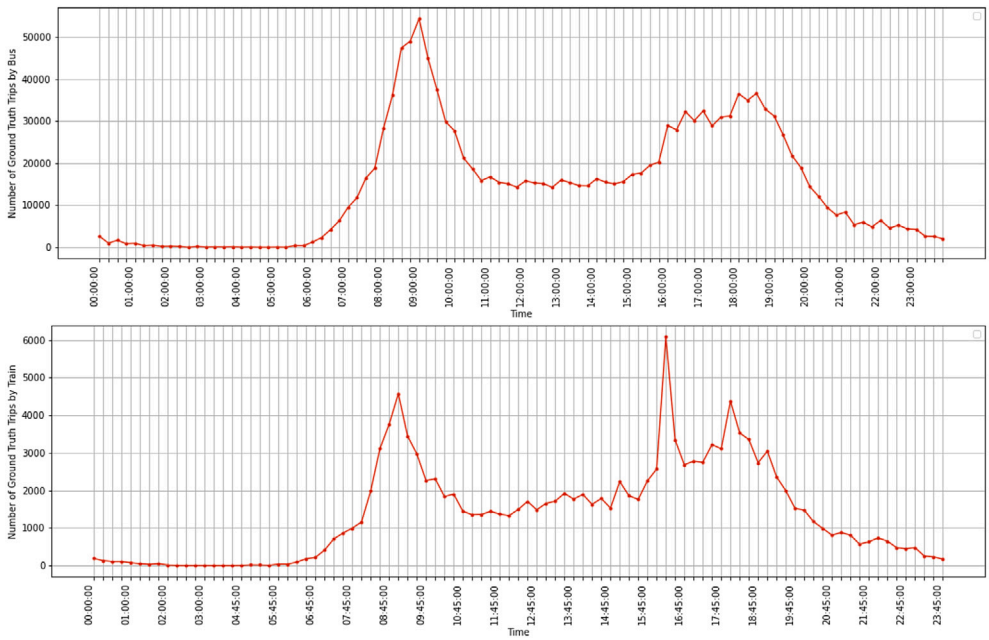


Fig. 3. Ground truth PT trip distributions (shown by time) for buses (up) and trains (down).

exemplified in this paper, we focus mainly on the morning peak hour (as the afternoon can follow a similar approach); therefore the data for 7:00 to 11:00 is collected and used for the estimation method.

Table 1
Scenario and experiment settings.

Network	Calibration Method	Travel cost type	Fitting function	Travel cost features	Notation
Bus/train	Traverse Searching	Single cost	Optimal form (Exponential/Power/Tanner)	Single cost (Connection/closeness/straightness/fare cost/travel time/travel distance)	S1E1
	Hyman's	Single cost	Optimal form (Exponential/Power/Tanner)	Single cost (Connection/closeness/straightness/fare cost/travel time/travel distance)	S2E1
		Multiple costs	Optimal form (Exponential/Power/Tanner)	Fused Costs	S2E2
			Optimal form (Exponential/Power/Tanner)	Optimal Fused	S2E3
	Entropy-weighted	Multiple costs	Optimal form (Exponential/Power/Tanner)	Fused Costs	S3E1
			Optimal form (Exponential/Power/Tanner)	Optimal Fused	S3E2

The topological feature data for the cost matrix estimation is captured from historical PT GTFS data provided by the OpenData (2017) and another open source TransitFeeds (2017). The data include the information of the PT agency, its calendar, the routes information, the PT stop times and stop location, as well as all the information regarding the PT trips and stops. The data for June 2017 is collected and used in this section as an exemplification.

4.2. Assumptions

External nodes: In this research study, an external node is inserted in the OD estimation approach using the Gravity Model detailed in *Stage 3* to balance the total generation and attraction trips. Since the M2 area is only a tiny part of the Great Sydney area, apart from the trips completed within the study area, there are still trips that solely start in the M2 area, which requires a match of an external node to attract those trips. In contrast, for those trips that simply end in the M2 area, an exterior node is also required to generate these trips. This way, we can reach a balanced GA vector that matches the constraints of Eqs. (7) and (8).

Convergence criteria: In the case study, the maximum iteration number for each experiment is 20, and the convergence criteria are reached when the gap between estimated and ground-truth total generation and attraction is less than 0.01%. In this research study, we aim to identify the most suitable method for model calibration—one that is accessible, adaptable, and converges quickly. To achieve this, we strive to set the value of r_{max} as high as possible, ensuring that the model satisfies Eq. (12) while also keeping the computational requirements for iterative calculations manageable on our computer. Consequently, we conducted all experiments until they met the acceptable criteria multiple times. The maximum number of iterations required for convergence in all experiments is 18. Hence, we round up this number to 20 for input as r_{max}.

Hyman method using Tanner function: In the literature, Hyman (1969) demonstrated deterrence function calibration using only the exponential function form, with a single parameter, β, defining the function. Following the same approach, we apply the calibration method to a power function (see Eq. (15)). To streamline calculations and demonstrate the performance of the Tanner function (refer to Eq. (17)), we utilise identical values for α and β when considering the Tanner function.

Traverse Searching method limitations: When exploring values of α and β using the Traverse Searching method (refer to Section 3.2.5), according to the results following Hyman's calibration method, a range from 0 to 3 in intervals of 0.05 is used for the train network in this paper. For the bus network, in instances where a distinct pattern was not apparent, we extended the range to encompass values from 0 to 0.2 with intervals of 0.01. For the sake of time and computational considerations, we keep these parameter pairs constant throughout the iterative estimation processes.

4.3. Scenario settings and experiment settings

When estimating the deterrence function, we set up three main scenarios that match the three cost matrix calibration methods presented in *Stage 2*, namely Traverse searching *(S1)*, Hyman's method *(S2)*, and our proposed Entropy-weighted method *(S3)*. Under *(S1)*, we consider the "single travel cost feature" only, and we intend to find out the optimal type of single cost that suits a certain form of fitting function. For *(S2)*, all "single travel cost feature", "fused travel cost features" and the "entropy-weighted fused travel cost features" are considered, and for *(S3)*, we include "fused travel cost features" and the "entropy-weighted fused travel cost features", as shown in Table 1.

In *S1*, all six travel cost features are individually incorporated into three forms of deterrence functions using the Traverse Searching method, yielding 18 outcomes. The range of potential parameters is determined based on the detail explained in *Traverse Searching method limitations* of Section 4.2.

In *S2*, following Hyman's method, all six travel cost features are incorporated into three forms of deterrence functions, as well. Under this setup, we incorporate three sub-experiments, each testing the efficacy of a single cost, fused costs, or the cost of an optimal combination of features (i.e., "entropy-weighted fused travel cost features"). Specifically, the experiments of using single

Table 2
Example of a list of combinations among the travel cost features.

Combination No.	Travel cost features included	Combination No.	Travel cost features included
1	connection,closeness	39	connection,closeness,fare_cost,TravelTime
2	connection,straightness	40	connection,closeness,fare_cost,TravelDistance
3	connection,fare_cost	41	connection,closeness,TravelTime,TravelDistance
4	connection,TravelTime	42	connection,straightness,fare_cost,TravelTime
5	connection,TravelDistance	43	connection,straightness,fare_cost,TravelDistance
...	...	...	...
15	TravelTime,TravelDistance	53	connection,closeness,straightness,TravelTime,TravelDistance
16	connection,closeness,straightness	54	connection,closeness,fare_cost,TravelTime,TravelDistance
17	connection,closeness,fare_cost	55	connection,straightness,fare_cost,TravelTime,TravelDistance
18	connection,closeness,TravelTime	56	closeness,straightness,fare_cost,TravelTime,TravelDistance
19	connection,closeness,TravelDistance	57	connection,closeness,straightness,fare_cost,TravelTime,TravelDistance

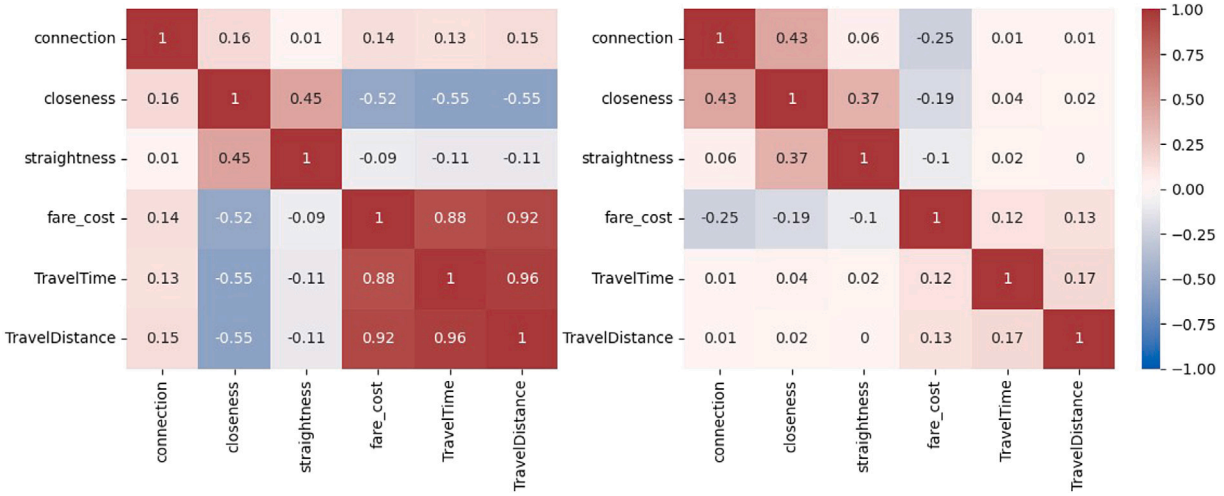


Fig. 4. Correlation matrices across six travel cost features for the bus networks (left) and the train networks (right).

costs *(S2E1)* cover all travel cost features individually, whereas fused costs *(S2E2)* involve a fusion of these features into a unified representation. Conversely, optimally fused cost *(S2E3)* follows a comparative evaluation across all feasible combinations of cost features. This evaluation aims to select the most optimal combination for utilisation in OD estimation. Such combination test results in 57 outcomes.

A total of 6 travel cost features results in 64 (2^6) possible combinations across these features. However, our method focuses on pre-ranking the importance of multiple features, with at least 2 features we can compare and rank. We cannot rank individual features in isolation. Therefore, we must eliminate all single-feature elements and the empty subset from the combinations. Consequently, we are left with a total of 57 combinations (64-6-1). An example of the combination list is shown in Table 2.

In *S3*, we assess the performance of each combination of travel cost features, resulting in 57 outcomes. This analysis provides evidence regarding the optimal selection of travel cost features that maximise benefits in OD estimation. Both the fusion of all travel cost features *(S3E1)* and the optimal combined fusion costs *(S3E2)* are tested under this setup.

4.4. Results

4.4.1. Correlation across travel cost features

A correlation matrix helps identify relationships between variables and is useful for understanding patterns and dependencies in the data. In our case study, we consider six travel cost features that have a deterrence effect on trip distribution to estimate the OD matrix for PT networks. Three of these six features are traffic characteristic-related, while the other four features are graph topological features. They are all related to the PT networks, but the internal relationships between each feature are still unknown. We want to identify the overlapping features or any new correlation between features that have never been found. In order to see the correlation between features, we visualise the assessment using a correlation matrix.

In the correlation matrix illustrated in Fig. 4, each cell represents a Pearson correlation coefficient. A coefficient of 1 indicates a strong positive relationship between variables, 0 indicates a neutral relationship, and −1 suggests a strong negative relationship. The Pearson correlation coefficient quantifies the linear correlation between two sets of feature data.

Table 3
Deterrence function calibration results when using a single travel cost feature via the Hyman's method applied for bus networks.

Exponential	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	38.50%	0.401	27.744	NaN	0.004	128
Closeness	36.66%	0.401	27.892	NaN	0.033	149
Straightness	36.72%	0.401	27.886	NaN	0.033	147
FareCost	38.50%	0.401	27.744	NaN	0.004	129
TravelTime	38.50%	0.401	27.744	NaN	0.004	128
TravelDistance	38.50%	0.401	27.745	NaN	0.004	129
Power	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	36.59%	0.401	28.194	0.025	NaN	136
Closeness	36.20%	0.402	28.285	0.033	NaN	134
Straightness	36.48%	0.401	28.218	0.002	NaN	137
FareCost	36.25%	0.402	28.297	0.067	NaN	138
TravelTime	36.97%	0.401	28.127	0.196	NaN	140
TravelDistance	36.59%	0.401	28.194	0.030	NaN	151
Tanner	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	38.49%	0.401	27.744	0.004	0.004	167
Closeness	36.74%	0.400	27.871	0.029	0.029	204
Straightness	36.51%	0.400	27.884	0.030	0.030	201
FareCost	38.49%	0.401	27.744	0.004	0.004	154
TravelTime	NaN	NaN	NaN	NaN	NaN	NaN
TravelDistance	38.49%	0.401	27.744	0.004	0.004	149

In the correlation matrix for bus networks, see Fig. 4 left, we observe a notable correlation among traffic state features, particularly between travel time and route distance, with a correlation coefficient of 0.96. This correlation arises because both variables capture aspects of the spatial relationship between two points within the network.

However, such a strong correlation is not as evident in the train network. This discrepancy can be attributed to the small size of the train network that we have available for this case study. In addition, train networks, unlike bus networks, operate more according to fixed schedules dictated by passenger demand and constrained by infrastructure limitations (such as capacity, speed, maintenance requirement, and crew schedule). These factors lead to variability in travel time relative to geographical factors, thus weakening the correlation between travel time and route distance in train networks.

Regarding the topological features in the bus network, we can observe that closeness (which reflects the reciprocal of the sum of distances from a node to all other nodes in the network), is related to straightness (which reflects the ratio of the Euclidean distance over the shortest path), with a coefficient of 0.45. Similarly, for a train network, the closeness is related to the straightness, with a coefficient of 0.37. One possible explanation for this correlation could be that a well-connected network with multiple routes may allow PT passengers to travel more in shorter distances.

Additionally, in small train networks, closeness shows a higher relation to connection (reflecting the total number of links connected to a given node), where the correlation coefficient is 0.43. This indicates that the train stations with more connections are likely to be centrally located and accessible to other stations, resulting in shorter average distances to reach other destinations in the network. However, bus networks often have a more decentralised structure, with multiple routes and stops dispersed throughout the network. This decentralised nature reduces the correlation between connection and closeness in bus stations, as there may not be a clear hub-and-spoke pattern as observed in the train network.

4.4.2. Calibration results of deterrence function considering single travel cost

Following the Hyman's method S2: In this section, we assess the performance of six distinct travel cost features across three different forms of the deterrence function, as per Hyman's approach within the Gravity Model framework. The results based on MAPE, indicate that the power forms considering closeness as the travel cost feature in a large bus network work the best, with average computation times of 134 s, as highlighted in red in Table 3. By using the exponential function, closeness and straightness perform well. These two features also suit Tanner's function better than the rest when fitting to Tanner's function. The analysis of RMSE results reveals more prominent outliers in the estimated matrix when employing the power function. According to the MAE results, the magnitude of errors between the estimated and historical matrices is relatively low when considering a Tanner function. Notably, the Tanner function does not align with the distribution of travel time. Based on the results, the optimal feature varies depending on the metric used. However, considering overall performance across all metrics, the exponential function using straightness emerges as the best scenario.

For a small network such as the train network examined in this case study, we note the remarkable effectiveness of all functions, particularly the exponential function, which displays rapid convergence and calculation. Evidently, all function forms prove suitable for estimating OD superior performance, depicted in Table 4, travel time demonstrates superior performance when following an exponential function, as highlighted in red. However, obtaining travel time by stops necessitates the timetable travel time for all connected trips along public transport routes based on GTFS data, which consumes significant computing time compared to the other features, posing a major drawback to its utilisation.

Following Traverse Searching method S1: Based on the results obtained from Hyman's method, we gain insights into the optimal fitting function and the ideal range of parameters. Subsequently, we conduct a thorough investigation into parameter performance through a traverse search. We select the exponential function for fitting purposes for both bus and train networks.

Table 4
Deterrence function calibration results when using single travel cost feature via Hyman's method for train networks.

Exponential	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	1.68%	0.412	0.796	NaN	1.55	0.69
Closeness	1.21%	0.397	0.722	NaN	3.57	0.40
Straightness	0.93%	0.362	0.661	NaN	0.03	0.75
FareCost	1.05%	0.357	0.681	NaN	0.27	0.72
TravelTime	0.74%	0.329	0.642	NaN	0.44	0.72
TravelDistance	0.90%	0.340	0.663	NaN	0.08	0.72
Power	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	24.87%	0.339	0.711	96625.0	NaN	0.31
Closeness	53.49%	0.535	1.103	518.7	NaN	0.66
Straightness	30.00%	0.432	0.858	285656.7	NaN	0.30
FareCost	42.06%	0.421	0.881	-13646.6	NaN	0.36
TravelTime	40.05%	0.798	1.600	30.7	NaN	0.28
TravelDistance	39.91%	0.399	0.838	-11831.0	NaN	0.32
Tanner	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Connection	4.95%	0.395	0.755	-16.2	-16.2	0.70
Closeness	39.91%	0.399	0.838	609.5	609.5	0.33
Straightness	31.64%	0.465	0.909	128.9	128.9	0.71
FareCost	12.84%	0.471	0.868	19.8	19.8	0.67
TravelTime	35.30%	0.441	0.830	3799.7	3799.7	0.56
TravelDistance	9.04%	0.414	0.809	9.8	9.8	0.80

Table 5
Deterrence function calibration results when using fused travel cost features via the Hyman's method and our proposed entropy-weighted method for the bus networks.

Hyman Method	MAPE	MAE	RMSE	Alpha	Beta	Computing Time (second)
Exponential	20.27%	0.534	4.698	NaN	-13.06	190
Power	NaN	NaN	NaN	NaN	NaN	138
Tanner	23.16%	0.637	6.448	-10.58	-10.58	161
Entropy Method	MAPE	MAE	RMSE	Alpha	Beta	Computing Time (second)
Exponential	23.109%	0.631	4.96	NaN	1.00	152
Power	3.038%	0.03996	3.07	1.00	NaN	136
Tanner	23.118%	0.632	4.92	1.00	1.00	150

Based on the results of our experiments for the bus network, we have found that according to the results obtained based on Hyman's method, we can locate the optimal range of parameters for further traverse searching. The results of traverse searching provide us with an average MAPE of 35% across all three forms of functions, which is approximately a 6.4% improvement in accuracy. However, the average computing time to find the optimal estimation results is about 7862 s across three forms of function and all six travel cost features. It is crucial to recognise that traverse searching has a significant drawback in terms of time consumption, and we must also account for the pre-processing time required by Hyman's method.

For train networks, traverse searching helps us find the optimal parameters, resulting in a 73% improvement in accuracy. However, the time required to achieve this result is approximately 1244 s, compared to only about 0.67 s when using Hyman's method. This highlights, on the one hand, the importance of incorporating Hyman's method (or another pre-search method) for preliminary parameter range exploration. On the other hand, it is crucial to recognise that traverse searching has a significant drawback in terms of time consumption.

4.4.3. Calibration results of deterrence function considering fused travel cost features

Following the Hyman's method S2: Acknowledging the effectiveness of integrating multiple travel costs in prior OD estimation studies, we introduce a scenario considering multiple travel costs. We standardise the costs by normalising their mean values to streamline the integration process. Thus, all travel cost features share an equal mean value of 1, facilitating fusion. As shown in Table 5, this fusion approach enhances the accuracy of OD estimation for the bus network, as indicated by MAPE and RMSE. However, the fused travel costs do not significantly reduce MAE. For a small train network, the fused travel costs improve accuracy according to both MAE and RMSE, as shown in Table 6. In this case, the accuracy improvement, according to MAPE, is achieved only when using the power function. This observation highlights the significance of optimal cost selection or appropriately weighing the costs. Notably, when adopting a fusion cost approach, the exponential form emerges as particularly effective for both types of networks, while the power function proves suitable when using the Entropy-weighted method.

Following the Entropy-weighted method S3: S3 follows the Entropy-weighted method outlined in Section 3.2.6, where we determine the weight assigned to each travel cost feature and construct an ensemble cost for fitting directly to deterrence functions. The results presented in Table 6 reveal that entropy ranking is effective for enhancing OD estimation accuracy in small train networks, facilitating convergence and improvement according to all metrics. However, in the case of bus networks examined in this study,

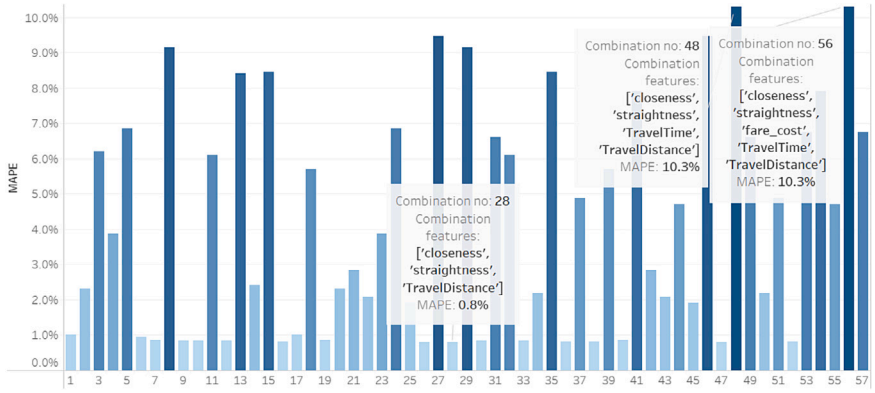


Fig. 5. MAPE error comparison for the combination of travel cost features in the train networks, fitted by an exponential function when using our proposed entropy-weighted method.

Table 6
Deterrence function calibration results when using fused travel cost features via the Hyman's method and our proposed entropy-weighted method for the train networks.

Hyman Method	MAPE	MAE	RMSE	Alpha	Beta	Computing Time (second)
Exponential	5.88%	0.432	0.776	NaN	6.42	0.57
Power	27.88%	0.279	0.555	-22142.24	NaN	0.53
Tanner	31.07%	0.477	0.882	38.02	38.02	0.61
Entropy Method	MAPE	MAE	RMSE	Alpha	Beta	Computing Time (second)
Exponential	1.35%	0.386	0.613	NaN	1.00	0.5
Power	1.26%	0.385	0.614	1.00	NaN	0.49
Tanner	1.27%	0.386	0.612	1.00	1.00	0.52

Table 7
Deterrence function calibration results when using fused travel cost features via the Hyman's method and our proposed entropy-weighted method for the bus networks.

Network	Calibration Method	Travel cost type	Fitting function	Travel cost features	Notation	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Bus	Hyman method	Multiple costs	Exponential	Closeness and Straightness	B_S2E3	1.50%	0.01503	0.34	NaN	0.75	128
	Entropy method (proposed)	Multiple costs	Power	Closeness and Straightness	B_S3E2	0.94%	0.015	0.47	NaN	1.00	111

as shown in Table 5, while Hyman's method already yields favourable outcomes, further enhancements achieved through entropy weighting are comparatively modest when fitting the data by a Tanner function. However, using the Power function vastly increases the estimation accuracy, and the computation time for matrix estimation following the Gravity Model is significantly saved.

4.4.4. Calibration results of deterrence function considering entropy-weighted fused travel cost features

Determination of optimal combination of travel cost features: To further investigate the performance of various travel cost features and their combinations on OD matrix estimation, we conducted a performance comparison of various combinations of travel cost features, as shown in Table 2. Such an experiment transversely estimates the OD matrix using all possible combinations of travel cost features based on the Gravity Model.

Due to space constraints, we solely present the comparison results for the train network utilising the Entropy-weighted approach and fitting by the exponential function, and their performance is assessed based on the MAPE, as shown in Fig. 5. The bars in the figure are arranged in ascending order based on combination numbers ranging from 1 to 57. Notably, the most favourable outcomes are achieved when utilising a combination of closeness, straightness and travel distance (Combination number 28); see evaluation results, which are summarised in Table 7. The performance of combined travel cost features exhibits variability across different combinations, incorporating additional travel cost features does not always improve the accuracy. Notably, outcomes are less favourable when considering travel time, as evidenced by Combination numbers 48 and 56, which yield comparatively poorer results.

According to the results, in the case of train networks, when relying on Hyman's method for calibration, the optimal feature combination comprises closeness, straightness and fare cost. When using the Entropy-weighted method, a combination of closeness and straightness becomes the best option. For the bus network, under both the deterrence function utilising Hyman's method and the Entropy-weighted approach, the cost of an optimal combination of features consist of closeness and straightness. Evaluation results for these optimal fusions are presented in Table 8.

Table 8
Deterrence Function calibration results when using fused travel cost features via the Hyman's method and our proposed entropy-weighted method for the train networks.

Network	Calibration Method	Travel cost type	Fitting function	Travel cost features	Notation	MAPE	MAE	RMSE	Alpha	Beta	Computing time (second)
Train	Hyman method	Multiple costs	Exponential	Closeness, Straightness and Fare Cost	T_S2E3	0.82%	0.37	0.67	NaN	0.049	0.66
	Entropy method (proposed)	Multiple costs	Power	Closeness and Straightness	T_S3E2	1.03%	0.33	0.55	1.00	NaN	0.23

Table 9
Summary of the optimal calibration method with suitable/optimal travel cost features.

Network	Calibration Method	Travel Cost Type	Optimal Fitting Function	Optimal (Combinated) Travel Costs	Notation
Bus	Traverse Searching	Single Cost	Exponential	TravelTime	B_S1E1
	Hyman's	Single Cost	Power	Straightness	B_S2E1
		Multiple Costs	Exponential	Fused Costs	B_S2E2
			Exponential	Closeness, Straightness	B_S2E3
	Entropy-weighted	Multiple Costs	Power	Fused Costs	B_S3E1
			Power	Closeness, Straightness	B_S3E2
Network	Calibration Method	Travel Cost Type	Optimal Fitting Function	Optimal (Combinated) Travel Costs	Notation
Train	Traverse Searching	Single Cost	Exponential	TravelTime	T_S1E1
	Hyman's	Single Cost	Exponential	TravelTime	T_S2E1
		Multiple Costs	Exponential	Fused Costs	T_S2E2
			Exponential	Closeness, Straightness and Fare Cost	T_S2E3
	Entropy-weighted	Multiple Costs	Power	Fused Costs	T_S3E1
			Power	Closeness, Straightness	T_S3E2

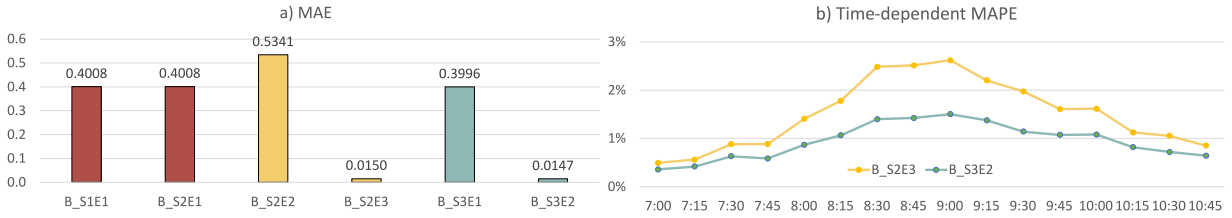


Fig. 6. (a) MAE errors for the bus networks evaluated between the ground truth and various estimated OD matrices and (b) Time-dependent MAPE evolution for our proposed entropy-weighted method (S3E2) versus the Hyman's method (S2E3).

4.4.5. Comprehensive comparison of calibration methods performance on OD estimation

Drawing from the analyses conducted above, we consolidate the optimal deterrence function forms, calibration method and cost features in Table 9. This table is derived from Table 1 but divided for bus or train networks. The following section overviews each experiment's performance metrics, including MAE, MAPE, and RMSE.

Comparison of performance using various calibration methods when optimal travel costs are adopted for bus network: By comparing the bar charts in Fig. 6, we observe that the method of *S3E2* (*Entropy-weighted* considering the cost of an optimal combination of features) performs the best among others in an accurate OD estimation for the bus network. Examining the MAE, it is observed that the inclusion of fusion travel costs (results of *S2E2* and *S3E1*) leads to increases in the magnitude of errors compared to results by single travel cost. This highlights the importance of selecting the right travel cost features, as choosing the wrong attributes can reduce the accuracy of estimations. Building on this concept, we tested various combinations of travel cost features and discovered that using Hyman's calibration method, the combination of closeness and straightness (*S2E3*) yields good performance. Similarly, our proposed Entropy-weighted method *S3E2* also demonstrates that this combination excels in accurately calibrating the OD matrix. The Entropy-weighted method not only improves estimation accuracy but also reduces computing time, as demonstrated in Table 7.

Fig. 6 (b) presents the time-dependent error distribution, showcasing the MAPE for both Hyman's method and the *Entropy-weighted* method considering the cost of an optimal combination of features. The figure illustrates consistently low MAPE values across all time slots for both scenarios. However, a slight peak is evident from 8:30 to 9:30, aligning with the peak patronage hour.

Comparison of performance using various calibration methods when optimal travel costs are adopted for train network: In this comparison, single travel costs, specifically, travel time *S2E1*, outperform the utilisation of multiple travel costs within the proposed method, as indicated by lower MAE and MAPE values, as shown in Fig. 7. However, according to RMSE, the performance of the proposed *Entropy-weighted* method stands out, indicating that there is a less large error when using such an estimation method. Apart from *S2E1*, the performance of the proposed Entropy-weighted method performs well according to MAE and, on average, is 19% more accurate than that for others. By comparing the results of using fused or optimally combined travel costs, we observe that selecting the optimal combination of features effectively reduces errors across all metrics. Based on the results following RMSE, the use of the Entropy-weighted method increases the accuracy by 23% compared to the average error measure across the rest of the five scenarios.

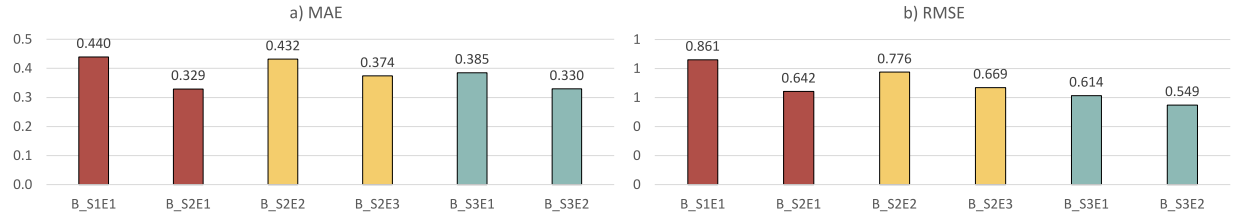


Fig. 7. Train networks errors evaluated between the ground truth and various estimated OD matrices when using: (a) MAE and (b) RMSE.

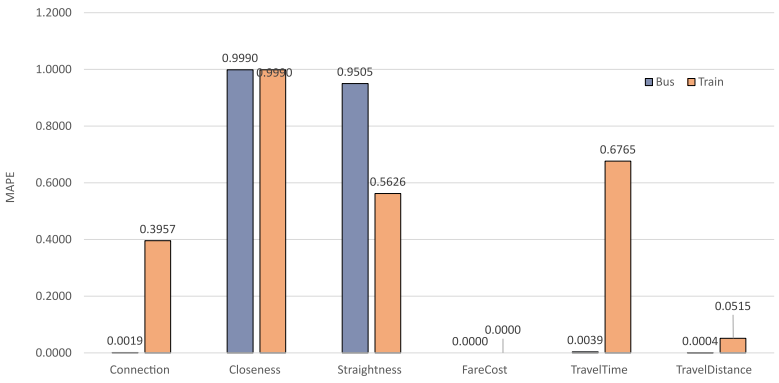


Fig. 8. Normalised importance of each travel cost feature in bus and train OD estimation based on the entropy-weighted method.

4.4.6. Importance of travel cost features in demand estimation

We introduced Shannon's entropy-weighting method to evaluate the weights for certain travel cost features. As detailed in Section 3.2.6, each travel cost feature is ranked by Shannon's entropy derived from the impact probability of the cost. We then treat the entropy as a measure of the importance of assessing different travel cost features in estimating the demand matrix. The features with higher impact probabilities and lower entropy scores would be more significant in determining demand, and their inclusion would lead to more accurate demand matrix estimations. As shown in Fig. 8 below, we display the entropy of travel cost features.

The importance of each travel cost feature depends on the entropy calculated based on Shannon's entropy. The entropy calculated for each feature is then processed to determine the percentage of the total entropy contributing to the overall cost. This percentage can be used to determine the relative importance of each feature in influencing the overall cost. Since entropy is a measure of uncertainty, at this stage, the importance determined by the entropy means the uncertainty of the cost feature adds to the overall cost when estimating the demand pattern. Thus, the larger the importance calculated, the higher the influence on the demand estimation.

As shown in Fig. 8, blue bars represent the importance of each travel cost feature for buses captured by entropy. It is observed that closeness and straightness play a significant role in estimating demand patterns for buses. Network closeness, as introduced in Section 3.2.7, is determined by the connectivity between each node in the network. Therefore, a more connected network is likely to have a higher demand as more effective travel is allowed. The straightness refers to the effectiveness of travelling between each node pair. High straightness in a bus network indicates that routes between stops are relatively direct and straightforward. The high importance of these features matches that passengers are often more willing to use a bus system that provides a direct and short route between their OD. This is the major reason for including a bus system citywide: provide direct and short trips, improve accessibility and offer reliable and affordable transport for more residents.

The orange bars depict the significance of each travel cost feature for trains as determined by entropy analysis. In a small network such as a train network, a broader array of travel cost features influences OD estimation, with closeness, straightness and travel time emerging as the top three influential factors. In contrast to bus networks, the stability of travel time in train networks is enhanced by limited stops, designed schedules, pre-defined routes and adjusted travel speed. Consequently, travel time becomes an essential factor to consider when estimating the OD matrix for train networks.

5. Conclusion

This paper provides a new framework for a dynamic large-scale stop-by-stop OD estimation model for PT. In this model, we emphasise a microscopic stop-based OD matrix, yet in order to simplify the computing workloads, we assume that the time interval is 15 min and calculate the number of trips between any OD pair for every 15 min to mimic the dynamic condition. The proposed framework shows the ability of our model to examine the effects of the travel cost matrix, namely the "single travel cost feature",

"fused travel cost features" and the "entropy-weighted fused travel cost features", reflected by various performance metrics (MAE, RMSE and MAPE) between the ground truth matrix and the estimated matrix. The proposed large-scale OD estimation model is established based on the Gravity Model with total generation and attraction (GA vectors) by stops, network physical configuration data and transport services operation data on inputs. In terms of the input data, the smart card data that enables the GA vectors' estimation is used, and the PT GTFS data is processed for the cost matrix data from topological-level cost features, including the connection, closeness, straightness and efficiency, and the traffic-level cost features, such as the travel distance or the travel time as well as the travel distance-based fare costs.

This research study proposed a novel deterrence function calibration method, where Shannon's entropy is employed for weighting the feature for each node (represented by a PT stop). These are due to the fact that the method has pre-weighted the cost features before combining the impact of the cost feature together and applying them in the process of iterative parameter calibration, and the process of weighing the cost features can be separated from the iterative OD matrix estimation, which reduces the load of iterative computing. The performance of the Entropy-weighted method is contrasted with traditional Hyman's and Traverse Searching methods. In bus networks, the Traverse Searching method demonstrates the capability to identify optimal parameters, consequently enhancing estimation accuracy. Moreover, the fusion of travel cost features achieved through both Hyman's and Entropy-weighted methods contributes to further improvements in accuracy. It is worth noting that the cost of an optimal combination of features (closeness and straightness) holds the potential to enhance the accuracy of bus OD estimation further. Optimal deterrence function calibration for small train networks is achieved when utilising a single travel cost, such as travel time. While the incorporation of fusion travel costs decreases accuracy in OD estimation, the effective combination of practical travel cost features (closeness, efficiency and fare cost when calibrating by Hyman's method or closeness and straightness when calibrating by Entropy-weighted method) leads to a notable increase in accuracy.

Additionally, after illustrating the mean errors by time, the time-dependent tendency of error fluctuation matches the timely number of trips in the network: the peak errors occur when the maximum number of trips occurs in the network (at about 8:30–9:30). This indicates that the network over-crowding is associated with the accuracy of the OD estimation by using the proposed model.

This paper also highlights the benefits of using topological features as travel costs. Unlike travel speed or travel time, deriving topological features only requires fixed route shapes and public transport stop location data, making the process simpler. Our results show that incorporating topological features as travel costs enhances the accuracy of OD matrix estimation. For train networks, the combination of closeness and straightness yields the best performance. Similarly, for large bus networks, utilising closeness and straightness also improves estimation accuracy. Additionally, when evaluating the performance of single travel cost features, we found that using the connection feature for bus networks achieved the highest estimation accuracy.

Limitations and Future Directions: In our research study, a framework for a dynamic stop-by-stop OD matrix estimation for large-scale PT is provided. However, the model does not directly include the impact of general traffic, such as the delay time. Although the travel time is captured by using the scheduled time, we could have compared it with the historical travel time captured from smart card data.

The proposed PT OD matrix estimation model has the potential to be integrated into other transport networks, such as car networks, to form a large-scale multi-modal OD matrix estimation model. In this research, due to the data availability issue, we only establish the model for bus and train networks. The same modelling process can be applied to estimate the OD matrix for cars, light-rails or on-demand solutions.

Furthermore, an area for future research could involve examining transfers between bus and train networks subsequent to OD estimation. Although our historical smart card data contains limited information on transfers. Specifically, we can identify transfer trips but lack details on matching upstream and downstream transfer trips. Subsequent research endeavours may concentrate on aligning upstream and downstream transfer trips and exploring the number of transfers.

In regard to the deterrence function calibration, there is room for further exploration of parameters. However, due to time constraints in this paper, we have limited our focus to the Hyman and Traverse Searching methods. Many other estimation models, such as the Competing Destination Model (see Thorsen and Gitlesen (1998)), the Self Deterrence Model with Quadratic Cost (SDMQC) (see Fang et al. (1995)), or a combination of these models (see de Grange et al. (2010)), hold potential for investigation. Beyond the modelling approach, the form of the deterrence function can also be expanded to include functions like Weibull, Box–Cox, or March (see Rubio-Herrero and Muñuzuri (2021)), among others. Using smart card data, various other methods can be tested to generate OD matrices based on historical data. These methods include regression models, maximum likelihood estimation, Bayesian estimation, and machine learning algorithms.

Regarding the findings from the traverse search results, we have identified a tipping point where the MAE remains constant regardless of changes in parameter values. Investigating the underlying reasons for this phenomenon could be a direction for future research. Additionally, exploring the tipping point when considering travel cost features beyond those examined in this paper may provide insights into the relationship between travel cost and the accuracy of estimated OD matrices.

CRediT authorship contribution statement

Dong Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adriana-Simona Mihăiţă:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Yuming Ou:** Writing – review & editing, Supervision, Software, Project administration, Funding acquisition, Conceptualization. **Hanna Grzybowska:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Mo Li:** Writing – review & editing, Supervision, Funding acquisition.

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.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used the Grammarly and ChatGPT-3.5 in order to improve language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgements

This research is supported by the Australia Research Council (ARC), under scheme of the Linkage Projects (Grant ID: LP180100114) and the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

Abdel-Aal, M.M.M., 2014. Calibrating a trip distribution gravity model stratified by the trip purposes for the city of Alexandria. Alexandria Eng. J. 53 (3), 677–689. http://dx.doi.org/10.1016/j.aej.2014.04.006.

Ai, X., 2017. Node importance ranking of complex networks with entropy variation. Entropy 19 (7), 303. http://dx.doi.org/10.3390/E19070303, 2017, Vol. 19, Page 303. URL: https://www.mdpi.com/1099-4300/19/7/303/htm https://www.mdpi.com/1099-4300/19/7/303.

Australian Bureau of Statistics, 2021. Australian bureau of statistics. URL: https://www.abs.gov.au/.

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M., White, P., 2004. The demand for public transport: a practical guide. Transp. Policy 13, 295–306. http://dx.doi.org/10.1016/j.tranpol.2005.12.004.

Ben, C., Bouchard, R.J., Sweet, Jr., C.E., 1965. An evaluation of simplified procedures for determining travel patterns in a Small Urban Area. Highw. Res. Rec. 1 (88).

Bouchard, R.J., Pyers, C.E., 1965. Use of gravity model for describing urban travel. Highw. Res. Rec. 1 (88).

Chen, H., Zhang, L., Liu, Q., Wang, H., Dai, X., 2021. Simulation-based vulnerability assessment in transit systems with cascade failures. J. Clean. Prod. 295, 126441. http://dx.doi.org/10.1016/J.JCLEPRO.2021.126441.

de Grange, L., Fernández, E., de Cea, J., 2010. A consolidated model of trip distribution. Transp. Res. E 46 (1), 61–75. http://dx.doi.org/10.1016/j.tre.2009.06.001.

Delgado, J.C., Bonnel, P., 2016. Level of aggregation of zoning and temporal transferability of the gravity distribution model: The case of Lyon. J. Transp. Geogr. 51, 17–26. http://dx.doi.org/10.1016/j.jtrangeo.2015.10.016.

Dieter, K.H., 1962. Distribution of work trips in Toronto. J. City Plan. Div. 88 (1), 9–28. http://dx.doi.org/10.1061/JCPEAW.0000043, URL: https://ascelibrary.org/doi/abs/10.1061/JCPEAW.0000043.

Evans, S.P., 1973. A relationship between the gravity model for trip distribution and the transportation problem in linear programming. Transp. Res. 7 (1), 39–61. http://dx.doi.org/10.1016/0041-1647(73)90005-1.

Fang, S., Science, H., Tsao - Transportation and undefined 1995, 1995. Linearly-constrained entropy maximization problem with quadratic cost and its applications to transportation planning problems. pubsonline.informs.org 29 (4), 353–365. http://dx.doi.org/10.1287/trsc.29.4.353, URL: https://pubsonline.informs.org/doi/abs/10.1287/trsc.29.4.353.

Feldman, O., Forero-Martinez, J., Coombe, D., 2012. Alternative gravity modelling approaches for trip matrix synthesis. Transp. Res. Rec..

Ge, Q., Fukuda, D., 2016. Updating origin-destination matrices with aggregated data of GPS traces. Transp. Res. C 69, 291–312. http://dx.doi.org/10.1016/j.trc.2016.06.002.

He, B.Y., Chow, J.Y., 2021. Gravity model of passenger and mobility fleet origin–destination patterns with partially observed service data. Transp. Res. Board 2675 (6), 235–253. http://dx.doi.org/10.1177/0361198121992074, 2021. URL: https://journals.sagepub.com/doi/10.1177/0361198121992074.

Huang, Z., De Villafranca, A.E.M., Sipetas, C., 2022. Sensing multi-modal mobility patterns: A case study of helsinki using bluetooth beacons and a mobile application. In: Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022. IEEE, pp. 2007–2016. http://dx.doi.org/10.1109/BigData55660.2022.10020578, arXiv:2209.13537.

Hussain, E., Bhaskar, A., Chung, E., 2021. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. Transp. Res. C 125, 103044. http://dx.doi.org/10.1016/j.trc.2021.103044, URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000759.

Hyman, G.M., 1969. The calibration of trip distribution models. Environ. Plan. A: Econ. Space 1 (1), 105–112. http://dx.doi.org/10.1068/a010105.

Lin, J., Ban, Y., 2013. Complex network topology of transportation systems. Transp. Rev. 33 (6), 658–685. http://dx.doi.org/10.1080/01441647.2013.848955, URL: https://www.tandfonline.com/doi/abs/10.1080/01441647.2013.848955 http://www.tandfonline.com/doi/abs/10.1080/01441647.2013.848955.

Lu, Q.C., 2018. Modeling network resilience of rail transit under operational incidents. Transp. Res. A 117, 227–237. http://dx.doi.org/10.1016/J.TRA.2018.08.015.

McClean, S.I., 2003. Data mining and knowledge discovery. Encyclopedia Phys. Sci. Technol. 229–246. http://dx.doi.org/10.1016/B0-12-227410-5/00845-0.

Nie, T., Guo, Z., Zhao, K., Lu, Z.M., 2016. Using mapping entropy to identify node centrality in complex networks. Phys. A 453, 290–297. http://dx.doi.org/10.1016/J.PHYSA.2016.02.009.

Nikolić, M., Bierlaire, M., 2018. Data-driven spatio-temporal discretization for pedestrian flow characterization. Transp. Res. C 94, 185–202. http://dx.doi.org/10.1016/J.TRC.2017.08.026.

OpenData, 2017. Open data | TfNSW open data hub and developer portal.

Ortuzar, S., de Dios, J., Willumsen, L.G., 2011. Modelling Transport. Wiley, http://dx.doi.org/10.1002/9781119993308, URL: https://onlinelibrary-wiley-com.ezproxy.library.sydney.edu.au/doi/book/10.1002/9781119993308. arXiv:arXiv:1011.1669v3.

Pitombo, C.S., de Souza, A.D., Lindner, A., 2017. Comparing decision tree algorithms to estimate intercity trip distribution. Transp. Res. C 77, 16–32. http://dx.doi.org/10.1016/J.TRC.2017.01.009.

Qi, X., Ni, Y., Xu, Y., Tian, Y., Wang, J., Sun, J., 2021. Autonomous vehicles' car-following drivability evaluation based on driving behavior spectrum reference model. Transp. Res. Rec. 2675 (7), 129–141. http://dx.doi.org/10.1177/0361198121994857, URL: https://journals.sagepub.com/doi/10.1177/0361198121994857.

Reilly, W.J., 1931. The Law of Retail Gravitation. New York.

Rubio-Herrero, J., Muñuzuri, J., 2021. Indirect estimation of interregional freight flows with a real-valued genetic algorithm. Transportation 48 (1), 257–282. http://dx.doi.org/10.1007/s11116-019-10050-6.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. http://dx.doi.org/10.1002/J.1538-7305.1948.TB01338.X.

Shen, G., Aydin, S.G., 2014. Origin–destination missing data estimation for freight transportation planning: a gravity model-based regression approach. Transp. Plan. Technol. 37 (6), 505–524. http://dx.doi.org/10.1080/03081060.2014.927665, URL: http://www.tandfonline.com/doi/abs/10.1080/03081060.2014.927665.

Simini, F., González, M.C., Maritan, A., Barabási, A.L., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100. http://dx.doi.org/10.1038/nature10856.

Sun, W., Schmöcker, J.D., Fukuda, K., 2021. Estimating the route-level passenger demand profile from bus dwell times. Transp. Res. C 130 (May), http://dx.doi.org/10.1016/j.trc.2021.103273.

Suprayitno, H., 2018. Searching the correct and appropriate deterrence function general formula for calculating gravity trip distribution model. IPTEK J. Eng. 4 (3), http://dx.doi.org/10.12962/joe.v4i3.3762.

Ta, V.C., Dao, T.K., Vaufreydaz, D., Castelli, E., 2018. Smartphone-based user positioning in a multiple-user context with wi-fi and bluetooth. In: IPIN 2018 - 9th International Conference on Indoor Positioning and Indoor Navigation. IEEE, pp. 206–212. http://dx.doi.org/10.1109/IPIN.2018.8533809, arXiv:1807.05716.

Tamblay, S., Galilea, P., Iglesias, P., Raveau, S., Muñoz, J.C., 2016. A zonal inference model based on observed smart card transactions for santiago de Chile. Transp. Res. A 84, 44–54. http://dx.doi.org/10.1016/j.tra.2015.10.007.

Tamblay, S., Muñoz, J.C., de Dios Ortúzar, J., 2018. Extended methodology for the estimation of a zonal origin-destination matrix: A planning software application based on smartcard trip data. Transp. Res. Rec. 2672 (8), 859–869. http://dx.doi.org/10.1177/0361198118796356, URL: https://journals.sagepub.com/doi/10.1177/0361198118796356.

Thompson, C.A., Saxberg, K., Lega, J., Tong, D., Brown, H.E., 2019. A cumulative gravity model for inter-urban spatial interaction at different scales. J. Transp. Geogr. 79, 102461. http://dx.doi.org/10.1016/j.jtrangeo.2019.102461, URL: https://www.sciencedirect.com/science/article/pii/S0966692318304289.

Thorsen, I., Gitlesen, J.P., 1998. Empirical evaluation of alternative model specifications to predict commuting flows. J. Reg. Sci. 38 (2), 273–292. http://dx.doi.org/10.1111/1467-9787.00092.

TransitFeeds, 2017. Greater sydney GTFS - TransitFeeds.

Van Acker, V., Sandoval, S., Cools, M., 2021. Value-based approach to assess the impact of lifestyles on mode shares. Transp. Res. Rec. 2675 (3), 313–325. http://dx.doi.org/10.1177/0361198120971261.

Wang, B., Su, Q., Chin, K.S., 2021. Vulnerability assessment of China–Europe railway express multimodal transport network under cascading failures. Phys. A 584, 126359. http://dx.doi.org/10.1016/J.PHYSA.2021.126359.

Wei, J., Cheng, Y., Chen, K., Wang, M., Ma, C., Hu, X., 2022. Nonlinear model-based subway station-level peak-hour ridership estimation approach in the context of peak deviation. Transp. Res. Rec. 2676 (6), 549–564. http://dx.doi.org/10.1177/03611981221075624.

Williams, I., 1976. A comparison of some calibration techniques for doubly constrained models with an exponential cost function. Transp. Res. 10 (2), 91–104. http://dx.doi.org/10.1016/0041-1647(76)90045-9.

Wilson, A.G., 2013. Entropy in urban and regional modelling. Entropy in Urban and Regional Modelling. Taylor and Francis, pp. 1–166. http://dx.doi.org/10.4324/9780203142608, URL: https://ebookcentral.proquest.com/lib/uts/reader.action?docID=1111668.

Wismans, L.J., Friso, K., Rijsdijk, J., de Graaf, S.W., Keij, J., 2018. Improving a priori demand estimates transport models using mobile phone data: A rotterdam-region case. J. Urban Technol. 25 (2), 63–83. http://dx.doi.org/10.1080/10630732.2018.1442075.

Xiao, Y., Gao, Z., Qu, Y., Li, X., 2016. A pedestrian flow model considering the impact of local density: Voronoi diagram based heuristics approach. Transp. Res. C 68, 566–580. http://dx.doi.org/10.1016/J.TRC.2016.05.012.

Yang, Z., Bing, Q., Lin, C., Yang, N., Mei, D., 2014. Research on short-term traffic flow prediction method based on similarity search of time series. Math. Probl. Eng. 2014, http://dx.doi.org/10.1155/2014/184632.

Zhang, N., Graham, D.J., Hörcher, D., Bansal, P., 2021. A causal inference approach to measure the vulnerability of urban metro systems. Transportation 48 (6), 3269–3300. http://dx.doi.org/10.1007/S11116-020-10152-6/TABLES/6, arXiv:2007.05276. URL: https://link.springer.com/article/10.1007/s11116-020-10152-6.

Zhang, Y., Ng, S.T., 2021. A hypothesis-driven framework for resilience analysis of public transport network under compound failure scenarios. Int. J. Crit. Infrastruct. Prot. 35, 100455. http://dx.doi.org/10.1016/J.IJCIP.2021.100455.

Zhao, D., Mihaita, A.-S., Ou, Y., Shafiei, S., Grzybowska, H., Qin, K., Tan, G., Li, M., Dia, H., 2022. Traffic disruption modelling with mode shift in multi-modal networks. In: IEEE International Conference on Intelligent Transportation. Institute of Electrical and Electronics Engineers (IEEE), pp. 2428–2435. http://dx.doi.org/10.1109/ITSC55140.2022.9921763.