TRB Annual Meeting

Large-scale Public Transport Origin-destination Matrix Estimation via Weighted Graph and Traffic Features Integration

--Manuscript Draft--

Full Title:	Large-scale Public Transport Origin-destination Matrix Estimation via Weighted Graph and Traffic Features Integration
Abstract:	Estimation of the large-scale demand estimation for public transport in different cities can vary depending on the public transport network, public transport modes, and traffic data. To overcome the issue of traffic data shortage and effectively estimate the Origin-Destination (OD) matrix, we use the most accessible data: total boarding and alighting and public transport timetable, to capture the public transport dynamic patronage, and from this perspective, we establish a dynamic and microscopic OD matrix for public transport. In this paper, we propose a new method to model the dynamic large-scale stop-by-stop OD demand for public transport by developing a boosting of the gravity model via graph theory and Shannon's entropy. First, we propose a novel cost matrix estimation method that considers various sources of travel cost features extracted from both the traffic flow information in the traffic network and topological information in the graph network. Secondly, we proposed an ``Ensemble Cost Matrix Weighted by Entropy'' method to estimate the best weights of importance for each feature using Shannon's Entropy to maximise the performance of the cost matrix in OD matrix estimation. Third, we validate the proposed approach using real smart-card data. Last, by comparing the effectiveness of our proposed method with the traditional deterrence function-oriented methods, we prove that our proposed cost matrix estimation method cooperated OD matrix modelling method is superior in accurately OD matrix estimation to traditional methods by almost 54.46% according to RMSE, 84.44% according to MAPE, and 85.09% according to MAE.
Manuscript Classifications:	Data and Data Science; Urban Transportation Data and Information Systems AED20; Origin and Destination Data; Planning and Analysis; Transportation Demand Management AEP60; Travel Demand Modeling; Policy and Organization; Executive Management Issues; Strategic Management AJE10; Technology Demand and Performance; Public Transportation; Planning and Development AP025; Accessibility; Model; Network; Transformative Trends in Transit Data AP090; GTFS; Smart card data; Sustainability and Resilience; Transportation and Society; Community Resources and Impacts AME80; Travel Demand Models
Manuscript Number:	TRBAM-23-03979
Article Type:	Presentation
Order of Authors:	Dong Zhao
	Simona Mihaita
	Yuming Ou
	Hanna Grzybowska
	Sajjad Shafiei
	Kai Qin
	Hussein Dia
	Gary Tan
Additional Information:	
Question	**Response**
The total word count limit is 7500 words including tables. Each table equals 250 words and must be included in your count. Papers exceeding the word limit may be rejected. My word count is:	6490

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Is your submission in response to a Call for Papers? (This is not required and will not affect your likelihood of publication.)	No

Under Review

Manuscript

LARGE-SCALE PUBLIC TRANSPORT ORIGIN-DESTINATION MATRIX ESTIMATION VIA WEIGHTED GRAPH AND TRAFFIC FEATURES INTEGRATION

Dong Zhao
University of Technology Sydney, Ultimo, NSW 2007, Australia
Dong.Zhao@student.uts.edu.au

Adriana-Simona Mihăiţă
University of Technology Sydney, Ultimo, NSW 2007, Australia
adriana-simona.mihaita@uts.edu.au

Yuming Ou
University of Technology Sydney, Ultimo, NSW 2007, Australia
yuming.ou@uts.edu.au

Hanna Grzybowska
Data61, CSIRO, Eveleigh, NSW 2015, Australia and and University of New South Wales, Sydney, NSW 2052, Australia

Sajjad Shafiei
Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Kai Qin
Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Hussein Dia
Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Gary Tan
National University of Singapore, Singapore

Word Count: 6490 words + 0 table(s) × 250 = 6490 words

Submission Date: August 2, 2022

Under Review

ABSTRACT

Estimation of the large-scale demand estimation for public transport in different cities can vary depending on the public transport network, public transport modes, and traffic data. To overcome the issue of traffic data shortage and effectively estimate the Origin-Destination (OD) matrix, we use the most accessible data: total boarding and alighting and public transport timetable, to capture the public transport dynamic patronage, and from this perspective, we establish a dynamic and microscopic OD matrix for public transport. In this paper, we propose a new method to model the dynamic large-scale stop-by-stop OD demand for public transport by developing a boosting of the gravity model via graph theory and Shannon's entropy. First, we propose a novel cost matrix estimation method that considers various sources of travel cost features extracted from both the traffic flow information in the traffic network and topological information in the graph network. Secondly, we proposed an "Ensemble Cost Matrix Weighted by Entropy" method to estimate the best weights of importance for each feature using Shannon's Entropy to maximise the performance of the cost matrix in OD matrix estimation. Third, we validate the proposed approach using real smart-card data. Last, by comparing the effectiveness of our proposed method with the traditional deterrence function-oriented methods, we prove that our proposed cost matrix estimation method cooperated OD matrix modelling method is superior in accurately OD matrix estimation to traditional methods by almost 54.46% according to RMSE, 84.44% according to MAPE, and 85.09% according to MAE.

Keywords: Public transport, Dynamic origin-destination estimation, Entropy, Cost matrix estimation, Gravity model

Under Review

INTRODUCTION

Background

Demand estimation aims to understand the current and future usage of transport modes so that the designed transport management and operation schemes can be leveraged well to fit the realistic traveller's demand in the network. Therefore, techniques for the demand estimation modelling have become a critical task that has been attracting attention since the last century (*1*). One of the essential parts of the travel demand estimation is the origin and destination (OD) estimation or the travel pattern estimation, which demonstrates the trip distribution and the travel pattern by OD pairs in the network. Among the biggest research contributions regarding the OD estimation method, the gravity model, derived from the gravity law, was introduced in 1931 by (*2*) to capture the relationship between the trip distribution and the microscopic zonal demand (*1*). The model follows the rule that the number of trips between an OD pair is direct proportional to the total number of trips at an origin or destination, and negative correlated to the travel costs between groups (detailed modelling method is presented in the Section of Methodology). Since its appearance in the scientific community, the gravity model has maintained its attractiveness to researchers due to its ease of computing and converging.

Challenges for public transport OD estimation: Most related applications to the OD estimation have focused on private vehicles rather than on public transport (*3–6*). This is due to the large share of car travelling, as it is commonly believed that driving a personal car can offer a flexible, comfortable and efficient travelling experience. In addition, driving also matches the major personal, cultural and psychological significance (*7*). The benefits of driving individual-owned cars and the negative experiences related to public transport utilisation, such as long waiting times, transfer distances, delays or crowding, have the travellers give up on using public transport across multiple countries.

In addition, the data source for estimating car OD is more accessible than that for public transport. Data such as traffic counts (*3, 4, 8*) has been applied in demand modelling; nevertheless, the data for estimating the public transport demand, such as the number of boarding and alighting or loading (*9*), has only been collected and released since the recent introduction and use of smart-cards for tapping on and off from public transport modes. In the field of demand estimation for public transport, it is more reliable to estimate at an individual trip level, though there is a common alternative which is to estimate using the average number of in-vehicle travellers based on vehicles(*1, 10*).

Another problem that discourage work around an accurate estimation for the public transport OD estimation is the network complexity. Unlike travelling by car, public transport trips are limited by pre-defined routes and timetables, where the network graph is dynamically changing in time and location. When considering the public transport OD matrix for a mesoscopic or microscopic transport modelling, the stop-based OD matrix is required. However, the use of stops at the OD matrix is usually translated into a large OD matrix (reaching millions of OD pairs) which inherently drags a significant increase in computing.

The other complexity is that each public transport mode in a city (trains, buses, ferries light-rails) might have different operators or different ticketing systems: for example, people need to tap-on and tap-off on train trips but they need to only tap-on for bus trips or ferries which can lead to missing the last part of travel segment when people get off without tapping off. Integrating together multiple data sets of various public transport modes represents a great challenge across multiple cities around the world, and often transport agencies need to make estimations using

Under Review

missing and incomplete information; this severely affects the future planning of public transport services, as it leads to a wrong demand across the city. Having an integrated approach for multiple modes is one of our major expectations, however, due to the data availability issue, we only include the bus network in this paper. The dynamic OD matrix estimation method has the potential to be extended to include more transport networks. Details will be explained in Section of Conclusion.

Related works: When estimating the OD matrix for public transport, the travel cost which influences the number of trips is also vital (*1*). Such travel cost is also known as the friction factor (*11*) or the friction and travel matrix cost matrix if we consider the independent friction by OD pairs (*12*). The travel cost has then been found to perform better by fitting it to a deterrence function - the most popular being the exponential, the power and the Tanner functions (*1, 12, 13*). However, the parameter estimation for the deterrence function appears challenging as such estimation also relies on an iterative updating in the same way as the OD estimation method (*13, 14*). In terms of features, the most popular features in the literature are the travel cost, the travel distance (*3–6*) followed by the travel time (*15*). The cost matrix derived from a single travel cost feature is known as a single feature cost matrix.

In practice, the number of trips by public transport also depends on the features such as the fare cost, the waiting time to catch the service, the transfer time in between stops or the level of occupancy in the public transport service (*7*). The cost matrix used in an OD estimation then becomes a fusion cost matrix that includes multiple cost features (*7, 14, 16, 17*). The above cost features are all related to the traffic states; in our work presented in this paper, we consider in addition to these traffic states, the trip distribution is also associated with the network topology because the geometric properties affect the node and the link capacity as well as the passenger accessibility in the network, which can influence the link travel capacity and travel efficiency. Especially when considering a dynamic cost matrix, the pre-defined timetable can interfere with a timely access for travellers in specific locations. Therefore, this intuition raises our initiative to employ the topological features as well for the cost matrix estimation, as this will reflect the network accessibility and the inter-modality between all modes inside an interconnected public transport graph. The graph theory-oriented features have not been explored extensively in the past, exceptions being some recent studies regarding the analysis of the network vulnerability (*18–20*); only one recent study has explored their potential by separately using the betweenness and the travel time as travel costs (*15*). However, in this work, we present an integrated modelling approach based on the graph theory and traffic state features to capture all structural and behavioural aspects of public transport utilisation.

Originality of this work: Following the integration of multiple cost features, estimating their weights becomes another challenge. To address this challenge, different methods have been used in previous works, as summarised below:

- *Weighted by a friction factor*: this was derived by directly weighting the travel cost by pre-defined factors, such as early or late arriving time (*16*), or wight by a ratio of the distance between origin to the centre of the zone over the distance between zonal centre to destination (*21*).
- *Weighted by deterrence function fitting*: this method was triggered by converting the travel cost into the form of a deterrence function and iteratively finding the suitable parameter for each feature's deterrence function (*13*).
- *Weighted by data-driven methods*: have originated from historical data processing by using further optimisation functions, regressions, machine learning models or deep learning

Under Review

algorithms (*9, 22–27*).

In line with the method of weighting by the friction factor, the entropy measurement has the potential to estimate accurately the friction factor (*28*). Entropy was initially proposed in the information theory in order to quantify a system's uncertainty. It measures the average amount of information required to draw an outcome from a probability distribution. In the cost matrix estimation, the entropy measures the average level of the travel cost feature (information) required to possibly indicate that such a feature can actually discourage trips. However, the entropy measure has only been used in ranking or for evaluation purposes in the current literature (*18, 28–30*); to the best of our knowledge, no publication indicates the feasibility of using the entropy to rank the travel cost features and the advantage of using the entropy weighted cost matrix in OD estimation.

Contributions summary:

In this paper, we propose an approach to dynamically estimate the OD matrix for public transport based on the gravity model boosted by an entropy-weighting of the feature cost matrix. The proposed model is established at a microscopic level, and the OD trip matrix is constructed at a public transport stop-by-stop level. The required inputs of our model are represented by the Generation and Attraction (GA) vectors and the cost matrix, where the GA vectors are acquired from historical smart-card data, and the cost matrix is captured from the General Transit Feed Specification (GTFS) data. The calibration of the cost matrix requires a selection of methods that are able to ease the computing time and consider multiple key travel cost features. We propose an entropy-based ensemble cost matrix estimation algorithm that incorporates multiple travel cost features (derived from the traffic state and graph theory modelling), and weights them by their importance in the network; such importance is measured by the entropy calculation of each feature. Another two cost matrix estimation methods, namely the single feature cost matrix and the deterrence function-based cost matrix, are also provided in this paper in order to display the impact of the cost matrix on the accuracy of our proposed new OD estimation approach.

To summarise, the main theoretical and methodological contributions of this paper are the following:

- we propose a new framework for the dynamic stop-by-stop OD estimation for large-scale public transport by using the boosted gravity model,
- we construct a new entropy-based ensemble cost matrix by considering the impact of multiple features from the public transport network, including traffic features and graph topological features (such as connections, closeness, straightness and efficiency),
- we validate and evaluate the performance of the new cost matrix and the cost matrix calibration method by using large-scale smart-card data,
- we showcase the significant improvement in terms of RMSE, MAPE and MAE of our approach which outperforms classical methods.

Paper Structure

The rest of this paper is organised as follows: in the Section of Methodology we present the framework of the large-scale stop-by-stop OD estimation for public transport. The details of the boosted gravity model-based are further highlighted in Section of Public transport OD Estimation using the Gravity Model, followed by the methods of the cost matrix estimation, including the traditional methods of Single Feature Cost Matrix and Deterrence function-based Fusion Cost Matrix from multiple features against the proposed method of Entropy-based Ensemble Cost Matrix from

Under Review

multiple features. The application of the proposed Entropy-based Ensemble Cost Matrix from multiple features algorithm to a real network is presented in Section of Case Study with the Results of the case study to demonstrate the performances of different cost matrix estimations. Finally, the research Conclusion is provided to clarify the research limitations and offer future directions in this field.

METHODOLOGY

Modelling framework

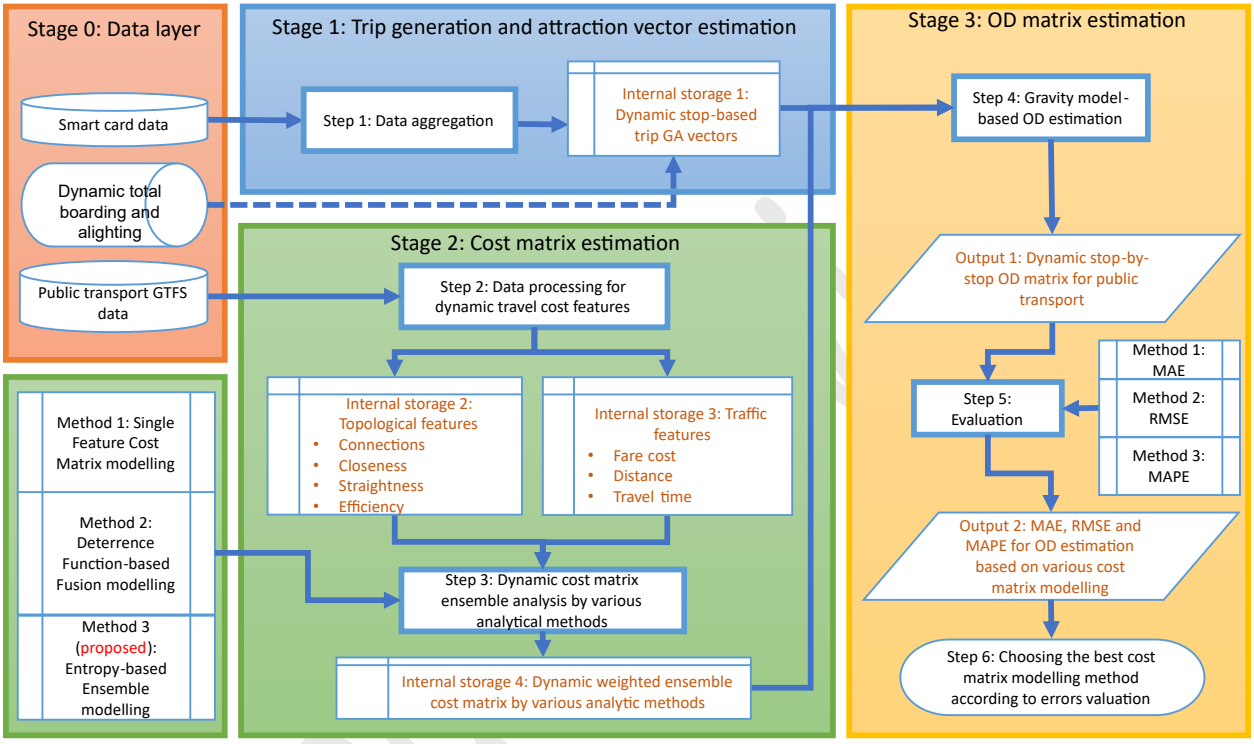

FIGURE 1 The framework of our proposed dynamic stop-by-stop OD estimation for large-scale public transport.

Figure 1 showcases our new proposed modelling framework for dynamically estimating the stop-by-stop OD matrix. The framework consists of three stages: at *Stage 0* we collect, filter and clean all the input data-sets (such as the smart-card data and the public transport GTFS data - Global Transit Feed Systems data). At this stage, we include not only the smart-card data but also the integrated total number of boarding and alighting, which indicates the data of the total number of tap-on and tap-off at each stop. The boarding and alighting data is an ideal alternative to smart-card data that can be available for all situations because the total loading data can be collected either by automatic ticket collection systems or on-site counting. With the total boarding and alighting data, it is also able to establish a microscopic stop-by-stop OD estimation for public transport by using our proposed modelling method. Therefore, a solid arrow is used in the figure to indicate the processing sequence of using the smart-card data, while a dashed arrow from total boarding and alighting data indicates that this is an alternative; at *Stage 1* we aggregate the smart-card data into trip generation and attraction vectors, which is the total loading information at all stops; at

Stage 2 we propose a cost estimation by the entropy algorithm *(method 3)* to combine and evaluate the impacts of various travel cost features, where three ensemble analysing methods are contained, namely: the Single Feature Cost Matrix, the Deterrence function-based Fusion Cost Matrix from multiple features and the Entropy-based Ensemble Cost Matrix from multiple features; at the final *Stage 3* we employ the gravity model for the OD estimation and further validate the feasibility of the proposed cost matrix calibration method inside the section Entropy-based Ensemble Cost Matrix from multiple features.

Public transport OD Estimation using the Gravity Model

For public transport 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 public transport network are the public transport stops.

Adjacency Matrix: The existence of trips between each OD pair depends on the available paths that are predefined for routing the public transport in the network. Therefore, the public transport network is defined as a Space L' graph, where the nodes are represented by the public transport stops and the links between nodes are the routes between any public transport stops. In a graph of Space L', different edges represent different links used in different networks with different directions (*31*). To capture the nature of dynamic timetables, such information is recorded in a time-dependent adjacency matrix:

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

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

$$h_{i,j}(t) = \begin{cases} 1, & \text{if } (i,j) \text{ is an accessible link at t,} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

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 public transport stops. 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 public transport 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, while 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:

$$od^{pt}_{m_i,n_j}(t) = A^{pt}_{m_i}(t)\ o^{pt}_{m_i}(t) B^{pt}_{n_j}(t) d^{pt}_{n_j}(t) f\left(c^{pt}_{m_i,n_j}(t)\right), \tag{3}$$

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 public transport and the total destinations ($d^{pt}_{n_j}$), respectively; $f\left(c^{pt}_{m_i,n_j}\right)$ is the deterrence function that represents the time-dependent travel cost between two zones by a mode pt; the origin stop m belongs to the origin zone i while the destination stop n belongs to the arrival zone j.

The constraints: To enable the estimated trips to match the real number of trips, there are two constraints employed in the Gravity model. Firstly, the total number of departures by a public transport mode from a public stop m should be equal to the sum of trips that originate from that

Under Review

particular stop to each possible destination stop n:

$$o_{m_i}^{pt}(t) = \sum_{j=1}^{N} od_{m_i,n_j}^{pt}(t), \tag{4}$$

and secondly, the total number of trips taken by public transport 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:

$$d_{n_j}^{pt}(t) = \sum_{i=1}^{N} od_{m_i,n_j}^{pt}(t) \tag{5}$$

The parameters: The time-dependent weights ($A_{m_i}^{pt}$ and $B_{n_j}^{pt}$) are the parameters of the Gravity model that are estimated iterativly. The estimation equations can be transformed from Equation 3 and Equation 4 to:

$$A_{m_i}^{pt}(t) = \frac{1}{\sum_{j=1}^{N} B_{n_j}^{pt}(t) d_{n_j}^{pt}(t) f\left(c_{m_i,n_j}^{pt}(t)\right)} \tag{6}$$

and from Equation 3 and Equation 5 to:

$$B_{n_j}^{pt}(t) = \frac{1}{\sum_{i=1}^{N} A_{m_i}^{pt}(t) A_{m_i}^{pt}(t) f\left(c_{m_i,n_j}^{pt}(t)\right)} \tag{7}$$

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

$$r \in \{1, ..., r_{max}\}, \tag{8}$$

or the functions of acceptable distance between iterative count r and $r+1$:

$$cc^{pt} \geq \max_{i,j}\left(\max_{i}\left(\frac{A_{m,i}^{pt,r+1} - A_{m,i}^{pt,r}}{A_{m,i}^{pt,r+1}}\right), \max_{i}\left(\frac{B_{n,j}^{pt,r+1} - B_{n,j}^{pt,r}}{B_{n,j}^{pt,r+1}}\right)\right) \tag{9}$$

where cc^{pt} is the criteria of convergence that defines the acceptable distance of the last two time-dependent weights, and r is the count of iterative calculations. The r_max used in this research is defined as shown in *Iterative OD estimation configuration* of Section 4.3.

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 ($d_{n_j}^{pt}$) and total arriving trips ($o_{m_i}^{pt}$) in the gravity model-based OD matrix estimation. As introduced in (*1*), 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. While the total attraction of a stop is the total 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 following:

$$o_{m_i}^{pt}(t) = \sum_{j=1}^{N} od_{m_i,n_j}^{pt,historical}(t), \tag{10}$$

$$d_{n_j}^{pt}(t) = \sum_{i=1}^{N} od_{m_i,n_j}^{pt,historical}(t) \tag{11}$$

and the GA vector can be represented as $GA_{m_i,n_j}^{pt} = [o_{m_i}^{pt}, d_{n_j}^{pt}]$.

Under Review

Single Feature Cost Matrix

In the OD estimation method via the Gravity model, the friction factors for each OD pairs are required. Such factors work as the negative function that limits the number of trips generated from origins to destinations in a network. Since the factors' values vary at each OD pair, it often appears as a matrix form and is normally known as the friction matrix. In practice, the friction matrix can be derived from travel costs such as trip length ($c_{m_i,n_j}^{tl,pt}$) or travel time ($c_{m_i,n_j}^{tt,pt}$):

$$C_{pt}(t) = \left[c_{m_i,n_j}^{pt}(t)\right] \tag{12}$$

where $i,j = \{1 \dots J\}, m,n = \{1 \dots N\}$. When calibrating the friction matrix that is subjected to the trip length, the common method is to iteratively adjust the parameters until the observed, and estimated travel cost distribution match each other (*11*). However, since the availability of General Transit Feed Specification (GTFS) (*32*) data, the public transport routing physical features such as travel time and travel distance is known and can be used to define a single feature cost matrix directly.

Deterrence function-based Fusion Cost Matrix from multiple features

In line with the travel cost-affected OD estimation, the deterrence function is proposed to better estimate the disincentives of travelling between any OD pair. We first employ a deterrence function with the Tanner form in this section: $f\left(c_{m_i,n_j}^{pt}\right)$, where the disincentives are related to: the public transport cost regarding the traffic features containing a) the fare cost ($c_{m_i,n_j}^{f,pt}$), b) the travel speed ($c_{m_i,n_j}^{ts,pt}$), c) the travel distance ($c_{m_i,n_j}^{td,pt}$) as well as the topological features including d) connection ($c_{m_i,n_j}^{cn,pt}$), e) closeness ($c_{m_i,n_j}^{cl,pt}$), f) straightness ($c_{m_i,n_j}^{st,pt}$) and g) efficiency ($c_{m_i,n_j}^{ef,pt}$) in this research study.

Therefore, the deterrence function taking a Tanner function form that combines the impact of multiple types of travel costs can be represented by:

$$f\left(c_{m_i,n_j}^{pt}\right) = \sum_{i=1}^{R} (c_{m_i,n_j}^{r,pt})^{\alpha}\, e^{-\beta c_{m_i,n_j}^{r,pt}} \tag{13}$$

where $r \in \{1 \dots R\}$ represents different travel costs.

Entropy-based Ensemble Cost Matrix from multiple features

In a cost matrix estimation by using the deterrence function, the function parameters are optimised iteratively, which costs a lot of computing time. Our proposed method is a first initiative to consider both the topological features (the degree which is the connection between nodes, the closeness between nodes, the straightness of links and the travel efficiency in the network) and the traditional traffic features (fare cost, travel speed and travel distance) for effectively estimating the cost matrix. Such a feature fusion method is derived from the principles of Shannon's entropy (*33, 34*) which has been used for feature ranking (see (*28–30, 35*)). In this way our current research study makes a further step and applies the entropy when weighting various travel costs and evaluating their true importance for the final trip demand estimation process.

In the following, we detail our proposed boosted estimation 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 $V(G) = \{v_1, v_2, \dots, v_m\}$, while the elements of E are the edges following $E(G) = \{e_1, e_2, \dots, e_l\}$. For a public transport mode pt, let G^pt be the graph where V^pt is the set of vertices, and E^pt is the set of edges.

Under Review

Travel cost feature representation: Each network has its unique topological feature reflected by centrality and global characteristics (see (*36*)). Currently, the most used centrality measures in the literature are the degree of each node, also known as the connection between each OD pair, closeness between each OD pair, straightness of each OD pair and network efficiency in relation to the shortest travel distance between each OD pair. The details of the topological features are described in Section 3.2.5. 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 includes both topological and traffic features, which are further represented by $C(G) = \{c_1, c_2, \ldots, c_q\}$.

According to the graph features, the matrix following travel cost features (c_j) by nodes (v_i) can be expressed as:

$$S = \begin{bmatrix} s(v_1,c_1) & s(v_1,c_2) & \ldots & s(v_1,c_q) \\ \ldots & \ldots & \ldots & \ldots \\ s(v_m,c_1) & s(v_m,c_2) & \ldots & s(v_m,c_q) \end{bmatrix} \tag{14}$$

where for each value of the cell, s is the travel cost value defined by the location, which is the public transport stop v (ordered as $1 \ldots m$), and the travel cost feature c, either topological or traffic features (represented by $1 \ldots q$). 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.

Standardised topological feature-based matrix: To standardised a feature r for each node (public transport stop), the ratio is estimated by using the mathematical formula below:

$$u_{i,r} = \frac{s_{i,r} - min(s_{i,r})}{max(s_{i,r}) - min(s_{i,r})} \tag{15}$$

Thus, the standardised topological feature matrix is denoted as:

$$U = \begin{bmatrix} s(v_1,c_1)u_{1,1} & s(v_1,c_2)u_{1,2} & \ldots & s(v_1,c_q)u_{1,q} \\ \ldots & \ldots & \ldots & \ldots \\ s(v_m,c_1)u_{m,1} & s(v_m,c_2)u_{m,2} & \ldots & s(v_m,c_q)u_{m,q} \end{bmatrix} \tag{16}$$

Entropy weighted topological measure: According to Shannon's entropy (see (*28, 33, 34*)), the ratio of each standardised topological feature $s(v_m,c_q)u_{m,q}$ is denoted by $p_{i,q}$, where:

$$p_{i,q} = \frac{s(v_m,c_q)u_{m,q}}{\sum_{q=1}^{Q} s(v_m,c_q)u_{m,q}} \tag{17}$$

which helps us to further estimate the entropy of each topological measure is denoted by:

$$I_q = -K \sum_{i=1}^{N} p_{i,q} \log(p_{i,q}) \tag{18}$$

Therefore, by updating Equation 16, the weight of each topological feature can be expressed as:

$$w_q = \frac{1 - I_q}{\sum_{q=1}^{n} (1 - I_{i,q})} \tag{19}$$

and the weighted standardized topological feature matrix now becomes:

$$V = \begin{bmatrix} s(v_1,c_1)u_{1,1}w_1 & s(v_1,c_2)u_{1,2}w_2 & \ldots & s(v_1,c_q)u_{1,q}w_q \\ \ldots & \ldots & \ldots & \ldots \\ s(v_m,c_1)u_{m,1}w_1 & s(v_m,c_2)u_{m,2}w_2 & \ldots & s(v_m,c_q)u_{m,q}w_q \end{bmatrix} \tag{20}$$

Under Review

Topological Features

As mentioned in the Sections 3.2.3 and 3.2.4, with the exception of the connection derived from the adjacency matrix, other topological cost features that are normally used in the graph theory are further selected to boost the accuracy of the cost matrix. These topological cost features are typical network characteristics such as the connection, closeness, straightness and efficiency are used in this research study. Such features reflect the travel distance-related characteristics that influence the route choice of passengers. The definition of each feature is expressed via the following equations, as discussed in (*36*):

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

$$c_{m_i,n_j}^{cl,pt} = \frac{1}{\sum_{m_j=1}^{J} d_{m_i,n_j}} \tag{21}$$

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

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

$$c_{m_i,n_j}^{st,pt} = \sum_{m_j=1}^{J} \frac{d_{m_i,n_j}^{Eucl}}{d_{m_i,n_j}} \tag{22}$$

Efficiency: is the property calculated by using the shortest travel length between each node pairs.

$$c_{m_i,n_j}^{st,pt} = \frac{1}{N(N-1)} \sum_{m_j=1}^{J} \frac{1}{d_{m_i,n_j}} \tag{23}$$

OD Matrix Evaluation

Assuming that the estimated OD matrix using our proposed approach is denoted as $[\hat{OD}_t]$, 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):

$$MAE = \frac{1}{N} \sum_{n=1}^{N} \left| [\hat{OD}_t] - [OD_t] \right| \tag{24}$$

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

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left([OD_t] - [\hat{OD}_t] \right)^2} \tag{25}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{[OD_t] - [\hat{OD}_t]}{[\hat{OD}_t]} \right| \tag{26}$$

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

Under Review

CASE STUDY

Geography and Data Information

The zones covered in the study are located to the North-West of the city of Sydney, along the M2 motorway which includes several major residential and business areas, as shown in Figure 2. This area is defined following the digital mapping according to the Statistical Area Level 2 (SA2) (*37*), which is denoted as $Z_j,\ j \in \{1 \dots J\}$. As of 2017, there are 89 bus routes that spread between 3799 stops. To simplify notations we will further refer to our case study as the M2 area in the following sections.

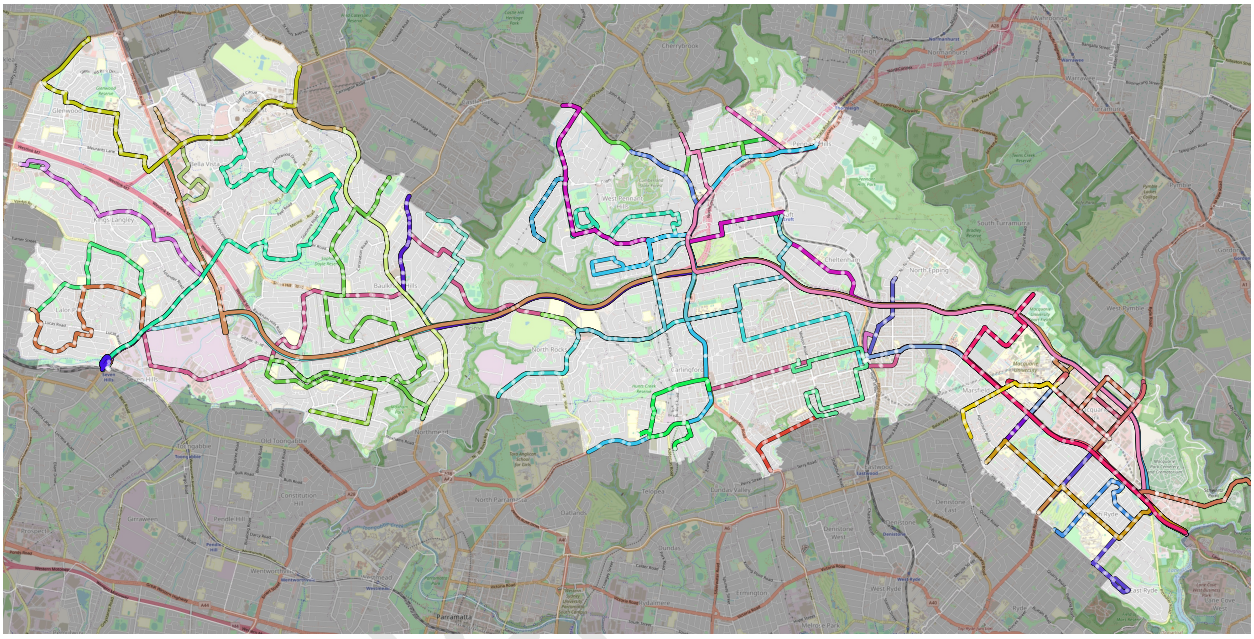

FIGURE 2 The 2017 Sydney M2 area bus network.

The trip distribution is captured from the local smart-card data. The raw smart-card data has been processed and filtered in advance for eliminating outliers and anomalies. The trip distribution in Figure 3 is drawn by using as an example one-month of smart-card data (June of 2017) for the M2 area in Sydney. As shown in the Figure 3, the morning peak hour starts from 7:00 and lasts until around 11:00, while the afternoon peak hour spreads from 16:00 to 20:00. In our case study 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.

The topological feature data for the cost matrix estimation is captured from historical bus General Transit Feed Specification (GTFS) data provided by the OpenData (*32*) and another open source TransitFeeds (*38*). The data include the information of the bus agency, its calendar, the routes information, the bus stop times and stop location, as well as all the information regarding the bus trips and stops. The data for June 2017 is collected and used in this section as an exemplification.

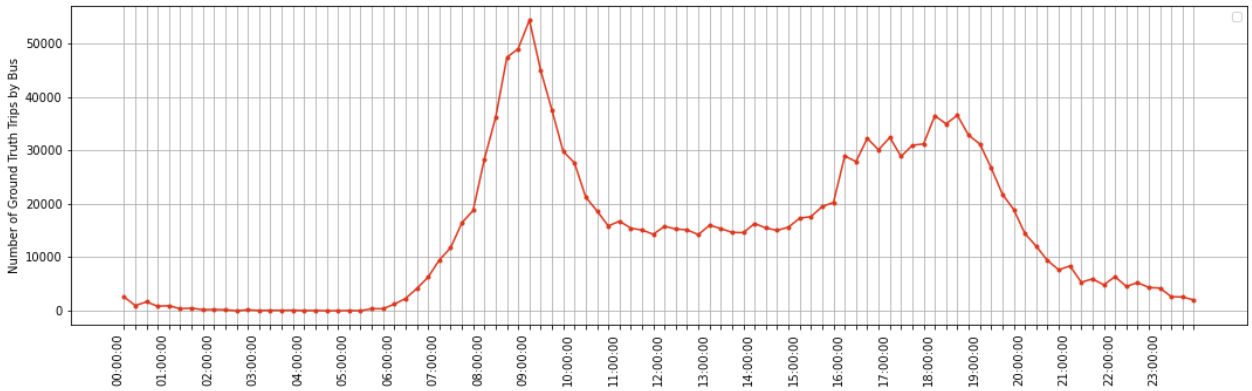

FIGURE 3 Ground Truth Bus Trip Distribution By Time.

Assumptions:

External Node: In this research study, an external node is inserted in the OD estimation approach by using the gravity model detailed in *Stage 3* in order to balance the total generation and attraction trips. Since the M2 area is only a small 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; while for those trips that simply end in M2 area, an external node is also required to generate these trips. In this way, we can reach a balanced GA vector that matches the constraints of Equation 4 and Equation 5.

Scenario and Experiments Configuration:

When estimating the cost matrix, we set up three scenarios that match the proposed three dynamic cost matrix estimation methods presented in *Stage 2*. In the first scenario, the Single Feature Cost Matrix is applied, where we select the top two commonly used features, namely the travel distance and the travel time, as our study matrix. For each cost feature, the experiments are split by either fitting to a deterrence function or not. Therefore, the settings of the experiment in the first scenario are:

- *Scenario 1 Experiment 1 (S1E1):* travel distance matrix
- *Scenario 1 Experiment 2 (S1E2):* travel distance matrix fitted by the deterrence function
- *Scenario 1 Experiment 3 (S1E3):* travel time matrix
- *Scenario 1 Experiment 4 (S1E4):* travel time matrix fitted by the deterrence function.

The second scenario considers a Deterrence function-based Fusion Cost Matrix from multiple features, and only one experiment is conducted (named as Fusion Cost Matrix Fitted by the Deterrence Function - S2). In this S2 scenario, both the traffic and the topological features are included to play a role in the trip estimation. The features are separately fitted to the Tanner function and then combined to establish a fusion cost matrix following Equation 13.

The Entropy-based Ensemble Cost Matrix from multiple features is then utilised in the third scenario (which we name as the Ensemble Cost Matrix Weighted by Entropy - S3), where instead of using the deterrence function to weigh each cost feature, we employ the entropy ranking algorithm following Equation 18, Equation 19 and Equation 20.

Iterative OD estimation configuration: for each experiment, the maximum iteration time is 20, and the convergence criteria are reached when the gap between estimated and ground-truth

Under Review

total generation and attraction is less than 1%. In this research study, we attempt to uncover the best method of cost matrix estimation, which is accessible, adaptable and convenient to use and fast to converge. Therefore, the r_max should be as large as possible to allow the model to satisfy Equation 9 while the processing requirements for the iterative calculation should also be acceptable for our computer. Following these, we conduct all our experiments (set up with different methods of cost matrix estimation, as introduced in subsection 4.3) until they meet the acceptable distance criteria as shown in Equation 9 multiple times, and the maximum number of the iterative counts towards convergence for all experiments is 18; therefore, we round the number of iterative counts to 20 as the input of the r_max, this is how we choose the r_max for this section.

Results

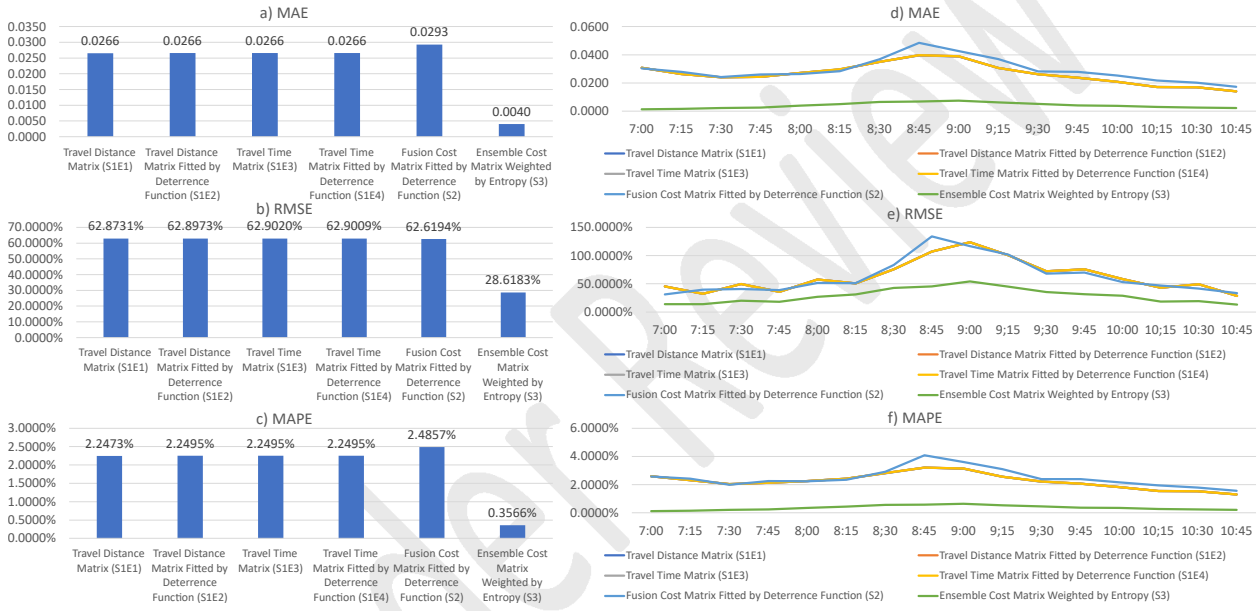

FIGURE 4 Errors evaluated between the ground truth and various cost matrix estimations by using: a) Mean Absolute Error (MAE) b) Root Mean Square Error (RMSE) c) Mean Absolute Percentage Error (MAPE), and Time-dependent error distribution by using d) MAE e) RMSE f) MAPE.

According to the experiments defined previously in Section of Assumptions:, Figure 4 presents the results following the evaluation methods from the Section of OD Matrix Evaluation (see *Step 5* in *Stage 3*). By comparing the bar charts in Figure 4a), b) and c), we observe that, after the same number of iterative OD estimations, the estimated *Ensemble Cost Matrix Weighted by Entropy (S3)* performs the best among others in an accurate OD estimation. The MAE value of *S3* is 0.0040, which means that by using the *Ensemble Cost Matrix Weighted by Entropy (S3)* method, we can obtain an estimated OD matrix that is approximately 85.0949% more accurate than the mean MAPE results of other methods.

According to the MAE, the proposed method is superior to the *Fusion Cost Matrix Fitted by Deterrence Function (S2)* by 86.2129%. The same superiority can be found in Figure 4c) when comparing the MAPE results. The MAPE from our proposed method is down to 0.35%, which is almost 84.44% less than the MAPE value of all the rest of approaches; for example, the MAPE

Under Review

value of the proposed method is considerably lower than that of the *Fusion Cost Matrix Fitted by Deterrence Function (S2)* by almost 85.65%.

The RMSE values provide another proof that our *Ensemble Cost Matrix Weighted by Entropy (S3)* method is about 28.61%, which is 54.45% less than the mean RMSE value estimated in other experiments. However, when estimating using RMSE, the worse estimation method is to use a single *Travel Time Matrix (S1E3)* as the cost matrix, where the RMSE for this method is 62.90%, which is 54.50% more than that of our proposed method.

The comparison of results between *S3* and *S2* shows that the entropy algorithm is also superior in performing the impacts of the ensemble cost features; by further comparing the results of *S1E1*, *S1E2*, *S1E3* and *S1E4* with *S2*, we are able to see that the deterrence function weighted fusion cost matrix cannot provide a relatively accurate OD estimation. This means that compared with using *Fusion Cost Matrix Fitted by Deterrence Function (S2)*, a single feature cost matrix could provide a better OD estimation. Nevertheless, when comparing the result for *S1E1* and *S1E3*, as well as *S1E2* and *S1E4*, we can observe that the travel distance-based cost matrix works better than the travel time in an OD estimation for public transport.

The above results obtained as average values for the whole modelling period, while the results that are shown in Figure 4 d), e) and f) are depicted dynamically by time to picture the timely influence of using different cost matrices for an accurate public transport OD estimation. The time period for this plotting is the morning peak hour, from 7:00 to 11:00, derived from the above Figure 3. From Figure 4 d), e) and f), overall, the errors calculated by all three methods for the experiments following *Ensemble Cost Matrix Weighted by Entropy (S3)* are significantly lower than the results of all other estimations (see the green curve of our proposed methodology which is significantly lower than the other curves). The standard deviation (SD) of MAE for the proposed *Ensemble Cost Matrix Weighted by Entropy* method is approximately 0.20%, whereas the mean SD for the rest of the results is 0.76%. Similarly, the SD of the RMSE for *(S3)* is almost 12%, while the mean SD of the rest of approaches is more than 28%; last the SD of MAPE results for *(S3)* is 0.16%, but the mean SD for the MAPE of other methods is nearly 0.59%.

The phenomenon that the tendency of errors matches the ground truth number of trips shows the overcrowding of the network can also be related to the accuracy of the OD estimation, which provides the same observation as shown in (*25*). According to the standard deviations of each curve, the timely change of errors for *Fusion Cost Matrix Fitted by the Deterrence Function (S2)* is larger than that of errors estimated based on a single travel cost matrix. This reminds us that the method of the cost matrix calibration is vital for the OD estimation, as an inferior calibration method could increase data noise and reduce the accuracy of the final public transport OD estimation.

CONCLUSION

In this paper, we provide a new framework for a dynamic large-scale stop-by-stop OD estimation model for public transport. In this model, we emphasise on a microscopic stop-based OD matrix, yet in order to simplify the computing workloads, we assume that the time interval is 15 minutes and calculate the number of trips between any OD pair for every 15 minutes to mimic the dynamic condition. The proposed framework shows the ability of our model to examine the effects of the cost matrix, namely the Single Feature Cost Matrix, Deterrence function-based Fusion Cost Matrix from multiple features and the Entropy-based Ensemble Cost Matrix from multiple features, reflected by various performance metrics (MAE, RMSE and MAPE) between the ground truth

Under Review

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, a 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 public transport 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. All the estimations in this model are time-dependent, and the time interval of estimations is 15 minutes.

This research study also proposed a novel cost matrix estimation method, where the Shannon's entropy is employed for weighting the feature for each node (represented by a public transport 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 in the process of iterative parameter calibration, and the process of weighting the cost features can be separated from the iterative OD matrix estimation, which reduces the load of iterative computing. The performance of the entropy weighting method is compared with traditional cost matrix weighting methods, namely *(Fusion Cost Matrix Fitted by Deterrence Function (S2), Travel Distance Matrix (S1E1), Travel Distance Matrix Fitted by Deterrence Function (S1E2), Travel Time Matrix (S1E3) and Travel Time Matrix Fitted by Deterrence Function (S1E4))*. According to the result reflected by errors, the performance of such weights fusion by deterrence function method *(S2)* is even inferior to using a single cost feature in an OD estimation. The results also show that the mean errors for experiments of *S1E1, S1E2, S1E3 and S1E4* are similar in their performance. Additionally, after illustrating the mean errors by time for each experiments, the time-dependent tendency of error fluctuation matches the timely number of trips in the network: the peak errors occurs when the maximum number of trips occurs in the network (at about 8:45-9:00). This indicates that the network over-crowding is associated with the accuracy of the OD estimation by using the proposed model.

Limitations and Future Directions In our research study, a framework for a dynamic stop-by-stop OD estimation for large-scale public transport is provided. However, the model does not directly include the impact of general traffic on the trip distribution. The travel time as the cost feature is obtained from public transport timetable data, where we assume that the timetable design considers the impact on the general traffic. But from the result of errors from experiments, the travel time method and the travel distance method do not lead to any clear differences. Thus, the impact of the general traffic states is unclear. As a next step we would like to combine the cost features such as the travel delay time or the level of congestion in order to uncover such impacts directly.

The proposed public transport OD estimation model has the potential to be integrated into other transport networks, such as car networks or other public transport networks, to form a large-scale OD estimation model. In this research, due to the data availability issue, we only establish the model for bus networks. The same modelling process can be applied to estimate OD matrix for trains, light-rails or on-demand solutions. In the proposed model, the public transport OD matrix is defined at the stops level, and by using an aggregation processing, the matrix can easily be converted into a zonal or centroid-based matrix that matches the configuration of the matrix for cars or other public transport networks. The OD matrix estimation demonstrates the travel pattern, and such matrix can be used as a test-bed to examine the network accessibility or the vulnerability analysis. For example, by degrading the node or the link capacity, the impacted trip distribution can

Under Review

be simulated, and a further direction from here could be an impacted journey recovery or further network optimisation.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: D. Zhao, S Mihăiţă and Y. Ou; data collection: D. Zhao, H. Grzybowska, S Mihăiţă and Y. Ou; analysis and interpretation of results: D. Zhao, S Mihăiţă and Y. Ou; draft manuscript preparation: D. Zhao, S Mihăiţă and Y. Ou. All authors reviewed the results and approved the final version of the manuscript.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article

ACKNOWLEDGMENTS

This research is supported by the ARC LP project LP180100114.

Under Review

REFERENCES

1. Ortuzar S., J. d. D. and L. G. Willumsen, *Modelling transport*. WILEY, 2011.
2. Reilly, W. J., *The law of retail gravitation*. New York,, 1931.
3. Shen, G. and S. G. Aydin, Origin–destination missing data estimation for freight transportation planning: a gravity model-based regression approach. *Transportation Planning and Technology*, Vol. 37, No. 6, 2014, pp. 505–524.
4. Thompson, C. A., K. Saxberg, J. Lega, D. Tong, and H. E. Brown, A cumulative gravity model for inter-urban spatial interaction at different scales. *Journal of Transport Geography*, Vol. 79, 2019, p. 102461.
5. Gonzalez-Calderon, C. A., J. J. Posada-Henao, and S. Restrepo-Morantes, Temporal origin–destination matrix estimation of passenger car trips. Case study: Medellin, Colombia. *Case Studies on Transport Policy*, Vol. 8, No. 3, 2020, pp. 1109–1115.
6. He, B. Y. and J. Y. Chow, Gravity Model of Passenger and Mobility Fleet Origin–Destination Patterns with Partially Observed Service Data:. *Transportation Research Board 2021*, Vol. 2675, No. 6, 2021, pp. 235–253.
7. Van Acker, V., S. Sandoval, and M. Cools, Value-Based Approach to Assess the Impact of Lifestyles on Mode Shares. *Transportation Research Record*, Vol. 2675, No. 3, 2021, pp. 313–325.
8. Shafiei, S., A.-S. Mihăiţă, H. Nguyen, and C. Cai, Integrating data-driven and simulation models to predict traffic state affected by road incidents. *https://doi.org/10.1080/19427867.2021.1916284*, 2021.
9. Shen, L., Z. Shao, Y. Yu, and X. Chen, Hybrid Approach Combining Modified Gravity Model and Deep Learning for Short-Term Forecasting of Metro Transit Passenger Flows:. *Transportation Research Board*, Vol. 2675, No. 1, 2020, pp. 25–38.
10. Ceder, A., *Public transit planning and operation : theory, modelling and practice*. Elsevier, 2007.
11. Evans, S. P., A relationship between the gravity model for trip distribution and the transportation problem in linear programming. *Transportation Research*, Vol. 7, No. 1, 1973, pp. 39–61.
12. Wilson, A. G., *Entropy in urban and regional modelling*. Taylor and Francis, 2013.
13. Sbai, A. and F. Ghadi, Impact of Aggregation and Deterrence Function Choice on the Parameters of Gravity Model. *Lecture Notes in Networks and Systems*, Vol. 37, 2018, pp. 54–66.
14. Feldman, O., J. Forero-Martinez, and D. Coombe, ALTERNATIVE GRAVITY MODELLING APPROACHES FOR TRIP MATRIX SYNTHESIS. *Transportation Research Board*, 2012.
15. Lu, Q. C., Modeling network resilience of rail transit under operational incidents. *Transportation Research Part A: Policy and Practice*, Vol. 117, 2018, pp. 227–237.
16. Abdelghany, A. F., H. S. Mahmassani, and A. F. Abdelghany, *A stochastic temporal-spatial microassignment and activity sequencing model for Temporal-Spatial Microassignment and Sequencing of Travel Demand with Activity-Trip Chains*. Journal of the Transportation Research Board, 2003.
17. Krishnakumari, P., H. van Lint, T. Djukic, and O. Cats, A data driven method for OD matrix estimation. *Transportation Research Part C: Emerging Technologies*, Vol. 113, 2019, pp. 38–56.

Under Review

18. Zhang, Y. and S. T. Ng, A hypothesis-driven framework for resilience analysis of public transport network under compound failure scenarios. *International Journal of Critical Infrastructure Protection*, Vol. 35, 2021, p. 100455.
19. Takhtfiroozeh, H., M. Golias, and S. Mishra, Topological-Based Measures with Flow Attributes to Identify Critical Links in a Transportation Network:. *Transportation Research Board*, Vol. 2675, No. 10, 2021, pp. 863–875.
20. Yun, J., J. Lee, J. Park, K. Chung, and J. Lee, How to Measure the Network Vulnerability of Cities to Wildfires: Cases in California, U.S.A.:. *Transportation Research Board*, 2022, p. 036119812210955.
21. Evans, A. W., The calibration of trip distribution models with exponential or similar cost functions. *Transportation Research*, Vol. 5, No. 1, 1971, pp. 15–38.
22. Wen, T., A.-S. Mihăiţă, H. Nguyen, C. Cai, and F. Chen, Integrated Incident Decision-Support using Traffic Simulation and Data-Driven Models. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2672, No. 42, 2018, pp. 247–256.
23. Mihăiţă, A. S., L. Dupont, and M. Camargo, Multi-objective traffic signal optimization using 3D mesoscopic simulation and evolutionary algorithms. *Simulation Modelling Practice and Theory*, Vol. 86, 2018, pp. 120–138.
24. Mihăiţă, A. S., M. B. Ortiz, M. Camargo, and C. Cai, Predicting Air Quality by Integrating a Mesoscopic Traffic Simulation Model and Simplified Air Pollutant Estimation Models. *International Journal of Intelligent Transportation Systems Research 2018 17:2*, Vol. 17, No. 2, 2018, pp. 125–141.
25. Kumar, P., A. Khani, and G. A. Davis, Transit Route Origin-Destination Matrix Estimation using Compressed Sensing. *Research Article Transportation Research Record*, Vol. 2673, No. 10, 2019, pp. 164–174.
26. Shafiei, S., A.-S. Mihaita, and C. Cai, Trip Table Estimation and Prediction for Dynamic Traffic Assignment Applications. *26 th ITS World Congress*, 2019, pp. 21–25.
27. Ou, Y., A. S. Mihaita, and F. Chen, Dynamic Train Demand Estimation and Passenger Assignment. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020*, 2020.
28. Ai, X., Node Importance Ranking of Complex Networks with Entropy Variation. *Entropy 2017, Vol. 19, Page 303*, Vol. 19, No. 7, 2017, p. 303.
29. Qi, X., Y. Ni, Y. Xu, Y. Tian, J. Wang, and J. Sun, Autonomous Vehicles' Car-Following Drivability Evaluation Based on Driving Behavior Spectrum Reference Model:. *Transportation Research Board*, Vol. 2675, No. 7, 2021, pp. 129–141.
30. Wei, J., Y. Cheng, K. Chen, M. Wang, C. Ma, and X. Hu, Nonlinear Model-Based Subway Station-Level Peak-Hour Ridership Estimation Approach in the Context of Peak Deviation. *Transportation Research Board*, Vol. 2676, No. 6, 2022, pp. 549–564.
31. Yang, X. H., G. Chen, S. Y. Chen, W. L. Wang, and L. Wang, Study on some bus transport networks in China with considering spatial characteristics. *Transportation Research Part A: Policy and Practice*, Vol. 69, 2014, pp. 1–10.
32. OpenData, *Open Data | TfNSW Open Data Hub and Developer Portal*, 2017.
33. Shannon, C. E., A Mathematical Theory of Communication. *Bell System Technical Journal*, Vol. 27, No. 3, 1948, pp. 379–423.

Under Review

34. McClean, S. I., Data Mining and Knowledge Discovery. *Encyclopedia of Physical Science and Technology*, 2003, pp. 229–246.
35. Nie, T., Z. Guo, K. Zhao, and Z. M. Lu, Using mapping entropy to identify node centrality in complex networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 453, 2016, pp. 290–297.
36. Lin, J. and Y. Ban, Complex Network Topology of Transportation Systems. *Transport Reviews*, Vol. 33, No. 6, 2013, pp. 658–685.
37. Australian Bureau of Statistics, *Australian Bureau of Statistics*, 2021.
38. TransitFeeds, *Greater Sydney GTFS - TransitFeeds*, 2017.

Under Review