Updating origin-destination matrices and link probabilities in public transportation networks

M. Victoria Chávez-Hernández[1], Yasmín Á. Ríos-Solís[2*],
L. Héctor Juárez Valencia[3], Roger Z. Ríos-Mercado[4]

[1]Departamento de Matemáticas, Instituto Tecnológico Autónomo de México, Río Hondo No. 1, Álvaro Obregón, 01080, Mexico City, Mexico.
[2*]Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Monterrey, 64849, Nuevo León, Mexico.
[3]Departamento de Matemáticas, Universidad Autónoma Metropolitana Unidad Iztapalapa, Av. San Rafael Atlixco 186, Mexico City, 09340, Mexico City, Mexico.
[1]Graduate Program in Electrical Engineering, Universidad Autónoma de Nuevo León, Av. Universidad s/n, San Nicolás de los Garza, 66455, NL, Mexico.

*Corresponding author(s). E-mail(s): yasmin.riossolis@tec.mx;

Abstract

To update a public transportation origin-destination (OD) matrix, the link choice probabilities by which a user transits along the transit network are usually calculated beforehand. In this work, we reformulate the problem of updating OD matrices and simultaneously update the link proportions as an integer linear programming model based on partial knowledge of the transit segment flow along the network. We propose measuring the difference between the reference and the estimated OD matrices with linear demand deficits and excesses and simultaneously having slight deviations from the link probabilities to adjust to the observed flows in the network. In this manner, our integer linear programming model is more efficient in solving problems and is more accurate than quadratic or bilevel programming models. To validate our approach, we build an instance generator based on graphs that exhibit a property known as a "small-world phenomenon" and mimic real transit networks. We experimentally show the efficiency of our model by comparing it with an augmented Lagrangian approach solved by a dual ascent and multipliers method. Additionally, we compared our methodology with other instances in the literature.

Keywords: origin-destination matrix, integer linear programming, transit counts, transit assignment problem

1 Introduction

The dynamics of the city in terms of population and mobility are among the most critical challenges we face nowadays (Cascetta, 2009). Public transport systems often fail to offer a high-quality service, which may imply long travel times (Ceder, 2015). These failures are often due to the lack of information about the daily trips at each period of the day, the specific trip purpose (work, school, hospital, entertainment), and how people move in the public transportation system. These movements can be represented in a two-dimensional array known as the Origin-Destination matrix (OD matrix).

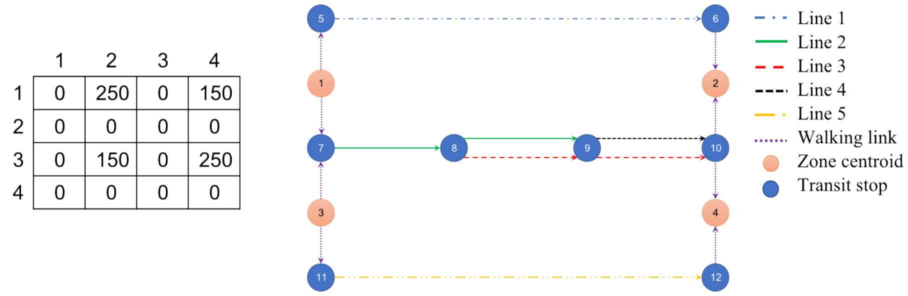
OD matrices are relevant at each stage of the transit network planning process, usually divided into two main stages (Ibarra-Rojas et al, 2015). The first stage is tactical planning, where accurate OD matrices are needed for the transit line design and the generation of useful timetables (departure times of the trips). This stage focuses on offering high-quality service to the customers, including line frequency, waiting times, and short transfers (Ibarra-Rojas et al, 2014). The second stage is operational planning, where the vehicle and crew scheduling problems seek to minimize the transport system operating costs (Ge et al, 2022). Updated OD matrices guide decision-makers in establishing service frequencies or designing new transit lines to match trip demand.

Another example is when a driver does not show up or when there are accidents or any other contingency case. The more information about the OD matrices, the faster the network can be restored (Boyer et al, 2018). Forecast OD matrices allow us to test the current system under more demanding scenarios and adapt the infrastructure for future demand.

OD matrices are usually obtained from home-based surveys every ten years (Bera and Rao, 2011). They are expensive and time-consuming to process (six months to one year in Mexico). Thus, once the new OD matrix is available after processing the surveys, it may already be obsolete. Therefore, we propose a methodology to rapidly update OD matrices in public transportation.

Lately, technology has allowed us to obtain more information about the trips made by users, and that information could be used efficiently to update OD matrices. In this work, we use flow observations made at some transit segments obtained by fare-box, automated fare collection systems, automatic passenger counter systems, geographical positioning on cellular phones, or even surveillance videos to update an obsolete OD matrix. In this sense, the link probabilities correspond to the likelihood that a traveler uses a particular transit segment of the network, such as a specific bus route or subway line. The traveler's preferences, the available routes to make the trip, and the costs influence these probabilities. They are often estimated using path choice models describing how passengers choose their travel route. Moreover, while we estimate public transit OD matrices, we also consider that small perturbations of the link probabilities could have arisen to verify our punctual flow observations of the network. We name the inverse problem of updating an OD matrix and its link probabilities the ODA problem.

Let us illustrate and exhibit the importance of the ODA problem. Figure 1 shows an example of an (obsolete) OD matrix and its corresponding transit network presented in Wu and Lam (2006). Notice that the diagonal entries of the OD matrix are 0 and that it is not symmetric (in the morning, people go downtown and few to the peripheries). Entry (1,2) of the OD matrix means that 250 users enter the network at zone 1 and exit it at node 2 during a certain period of the day. The transit network has five lines (blue, green, red, black, and yellow), four-zone centroids (where trips originate and end), eight transit stops, and eight walking links that connect centroids to transit stops.



	1	2	3	4
1	0	250	0	150
2	0	0	0	0
3	0	150	0	250
4	0	0	0	0

Fig. 1 Reference (obsolete) OD matrix (left-hand side) of the transit network (right-hand side) with five lines and eight transit stops.

Table 1 specifies the headway of each line: the difference in minutes between two different vehicles at the depot, and it also indicates the travel time in minutes for each line-node-node transit segment (l, i, j).

Table 1 The network's five lines (blue, green, red, black, and yellow) are on the right side of Figure 1, with each segment's headway and travel time.

lines	headway	transit segment	travel time
	(min)	(l, i, j)	min.
1-blue	12	(1,5,6)	25
2-green		(2,7,8)	10
	6	(2,8,9)	10
3-red	6	(3,8,9)	10
		(3,9,10)	10
4-black	6	(4,9,10)	10
5-yellow	12	(5,11,12)	25

Figure 2 shows only the links a user may take from origin 1 to destination 4. Suppose the limits of the capacity of the vehicle and the network are not considered. In that case, users may not consider taking the transit segment (3,8,9) because, though the travel time by using line 3 or line 2 is the same, the waiting time is only determined by the frequency of service of line 3 in contrast with the case where the user change

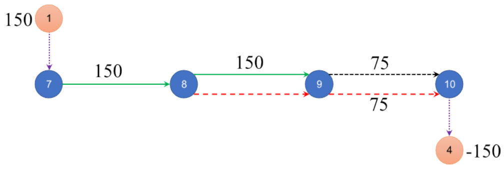


Fig. 2 Transit segments a user may take to go from origin 1 to destination 4 and the flow volume at each segment.

from line 2 to line 3 or line 4 at node 9 where the waiting time can be reduced by the combined frequencies of lines 3 and 4, it can be validated with data from Table 1. Values on the segments of Figure 2 are the number of passengers traveling from 1 to 4 according to the obsolete OD matrix and the link proportions in the third column of Table 2. Node value -150 means that node 4 attracts 150 trips from node 1. With the obsolete link proportions, all passengers use the walking link to go from node 1 to node 7 and then board the green line from node 7 to node 9; next, half of the passengers use the black line (4,9,10), and the other half use the red line (3,9,10) to arrive at node 10. None use the (3,8,9) segment. Finally, at node 10, all passengers take the walking link to arrive at node 4.

The first and second columns of Table 2 indicate the lines and segments of the network depicted in Figure 2. The third column is the link proportions that mimic how a user moves in a network based on a linear transit assignment procedure described by Spiess and Florian (1989). In this work, we consider that the obsolete link proportions may be updated to represent some slight changes in the behavior of the users (for instance, users may prefer to board line 3 at node 8 because if they wait to board at node 9 there may be more users waiting and not all of them might board the vehicle that arrives first to the node due to vehicles' capacity), as shown in the fourth column of the table. It is worth mentioning that although in this work, the Spiess and Florian assignment model is being considered, the methodology proposed in our work can be used considering other assignment models, such as the stochastic user equilibrium assignment used by Wu and Lam (2006) or even a shortest path model such as the one used by Cervantes-Sanmiguel et al (2023). The methodology we propose is to slightly modify the link proportions obtained from any assignment model to better represent the actual observed flows in the network.

Table 2 Transit segment probabilities associated to Figure 2.

Lines	Transit segments	Obsolete probabilities	Updated probabilities
2 green	(2,7,8)	1.0	1.0
	(2,8,9)	1.0	1.0
3 red	(3,8,9)	0.0	0.0
	(3,9,10)	0.5	0.6
4 black	(4,9,10)	0.5	0.4

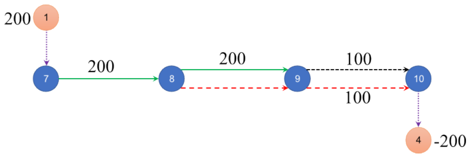


Fig. 3 The segment flows are generated with the obsolete probabilities for an updated amount of 200 users from 1 to 4.

Let us suppose that the number of trips originated at node 1 whose destination is node 4, is no longer 150 but increased to 200, and the infrastructure of the network has not changed. Using the obsolete probabilities of Table 2, we obtain Figure 3 with the updated segment flows. Nevertheless, this distribution of passengers along the network may not match the observed flow at some segments. For instance, let us suppose that 80 users have been observed at segment (4,9,10) and 120 users at segment (3,9,10). Thus, the link proportions and the number of users of the OD matrix must also be updated. Notice that the updated link proportions, presented in the fourth column of Table 2, must be close to the reference (obsolete) ones, and the new OD matrix must be similar to the reference one to preserve the dynamics of the city (Cascetta, 1984).

Therefore, the ODA problem aims to update all the OD matrix entries simultaneously and the new transit segment proportions by matching the observed flow at some transit segments of the network. We formulate a mixed-integer linear program

(MILP) to solve the ODA problem to avoid a quadratic objective function or a bilevel programming model, as most of the literature does (Wu and Lam, 2006; Mahmoodjanlou et al, 2019). Moreover, by estimating the OD matrix and the variation of the link proportions, we are dealing with some congestion effects without explicit them in our methodology. In addition to being one of the few linear models in the related literature, our model optimally solves instances larger than those reported with bilevel programming methods. Indeed, our MILP is enhanced by a family of valid inequalities and establishes bounds on the variables to provide a tighter integer linear programming formulation. Moreover, continuous variables count the number of users in all other approaches. Thus, the solutions obtained are not integer numbers, and it is necessary to round the number of users in the OD matrix entries to implement them in practice. In discrete combinatorial optimization, relaxed solutions often lead to large differences from the optimal integer ones (Schrijver, 1998; Wolsey, 1998). This work considers integer variables, even if the ODA problem becomes more challenging to solve, to stay closer to a realistic scenario and to have more practical and implementable solutions.

This paper is organized as follows. We first present the literature review in Section 2. The ODA problem is formally defined in Section 3. Then, its MILP model is presented in Section 4. To test our methodology, we propose a random instance generator to obtain close to real public transportation networks. Experimental results validate our methodology. In Section 5, we present a comparison of our method with a penalty-based quadratic model from the literature, showing the benefits of our approach in terms of accuracy and time. Moreover, we compare our approach with the instance presented by Wu and Lam (2006), which was solved with a bilevel programming approach. Then, Finally, in Section 6, we present the conclusions of this work.

2 Literature review

The OD matrix estimation approaches usually combine two stages of the four-stage sequential procedure. For instance, Fisk and Boyce (1983) propose a model that combines trip distribution and traffic assignment, while Fisk (1989) combines the entropy maximization method with traffic assignment. Also, Yang et al (1992) extended these results to congested networks where the link proportions are not constant. Most models in the literature update OD matrices in road networks (Cascetta, 1984), but only a few, including us, update OD matrices for public transportation networks.

Some approaches have used bilevel programming models to address the OD matrix estimation problem from segment counts. The upper level represents the OD estimation process, and the lower level represents a network equilibrium assignment (Yang et al, 1992; Florian and Chen, 1995; Wu and Lam, 2006; Frederix et al, 2013). In Liu and Fricker (1996), the authors propose a two-stage iterative method to estimate an OD matrix and the variation in link choices among trip makers. Still, inconsistencies arise when congestion effects are considered. In our work, we can slightly modify the link proportions to adapt to these effects.

Wu and Lam (2006) formulate a bilevel program with a stochastic user equilibrium assignment for congested transit networks; they simultaneously determine transit line frequencies and the network flow pattern in congested transit networks using a heuristic solution algorithm adapted for solving the OD estimation problem. In Yang et al (2001), the authors use the link flows obtained from the stochastic user equilibrium traffic assignment and estimated OD flows in the cost function. They use a sum of the squares of errors as the objective function and propose a successive quadratic algorithm to solve the model. Our work avoids a quadratic objective function, making the solution method based on branch and bound more efficient.

Based on the user equilibrium principle, some models succeeded in incorporating congestion effects into the estimation process, but the perception of travel costs does not vary among travelers (García-Ródenas and Verastegui-Rayo, 2013). A more realistic approach allows for the difference in cost perceptions and different link choice behaviors among travelers using a stochastic user equilibrium assignment (Lo and Chan, 2003; Mahmoodjanlou et al, 2019).

In Chávez-Hernández et al (2019), the authors consider a penalized quadratic model to update OD matrices from observed transit flow volumes. They present an augmented Lagrangian model that aims to minimize the difference between a reference matrix and the estimated one and between the observed segment flows and those obtained after a linear transit assignment of the estimated OD matrix. To solve the problem, the Karush-Kuhn-Tucker optimality conditions were formulated and solved with a dual ascent technique. In Section 5, we compare our numerical results with theirs. However, quadratic models are constructed to deal with large-scale networks. To overcome the difficulty of dealing with large networks with explicit management of route choice probabilities, Walpen et al (2020) have used heuristic methods where those probabilities are not used explicitly to solve the problem, such that the bilevel program is iteratively solved.

The growing interest in using big data from mobile phones in transport research has been important (Landmark et al, 2021; Cantelmo and Viti, 2020; Caceres et al, 2020); however, they have the limitation of having population biases in addition to their difficult identifiability (Liao et al, 2022). In He et al (2023), the authors use a deep learning approach with a multi-fused residual network. In contrast, López-Ospina et al (2022) use a maximum entropy optimization model to forecast travel demand, but they do not consider the assignment probabilities. In our research, the type of information used is the passenger flow in some transit network segments; according to Castillo et al (2013), this information can be sufficient to obtain a unique solution under reasonably strong assumptions on the assignment process.

Ge et al (2022) integrated vehicle and crew scheduling by revisiting an earlier formulation incorporating days-off patterns delay propagation. Zúñiga et al (2021) use available historical data and combine it with online information regarding the entry and exit of each particular user to make predictions and updates for the OD matrices.

Most of the mentioned approaches formulate the problem as a quadratic optimization problem. They include observed data, such as the flow of people at some segments of the transit network, and a reference matrix obtained from surveys or projections based on economic growth. There are relatively few authors who propose a linear model. For example, Ashok and Ben-Akiva (2002) formulate a model to estimate a dynamic OD matrix by defining a state vector regarding the departure rates from each origin to each destination. Pitombeira-Neto et al (2018) propose a linear model to estimate a dynamic OD matrix to represent the stochastic evolution of OD flows over time. They propose a Markov chain Monte Carlo algorithm to approximate the mean OD flows and the link choice model parameters.

3 The ODA problem

In this section, we formally present the ODA problem. We are considering updating an obsolete OD matrix at a specific period of the day. Our methodology is based on an optimization network flow model that avoids the most often used quadratic models for this problem (Chávez-Hernández et al, 2019). Instead, we count the excess or deficit of trips at each OD pair, as shown in Section 4.

Let us consider a public transit network with a set of lines $\mathcal{L}$. The public transit system is represented by a directed multigraph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the set of nodes (bus or subway stops), and $\mathcal{A}$ is the multiset of transit segments (directed links) of the lines in $\mathcal{L}$.

Segment or link $a \in \mathcal{A}$ is a triplet (l, i, j) indicating the line $l \in \mathcal{L}$ and the nodes i and j linked by line l, with both nodes in $\mathcal{N}$. Notice that in link (l, i, j), line l first passes through i and then through j. All the nodes (or centroids) in $\mathcal{N}$ are origins and destinations, thus $\mathcal{PQ} = \{(p, q) \in \mathcal{N} \times \mathcal{N}$ and $p \neq q\}$.

The reference OD matrix, denoted by $\widehat{\mathbf{g}} = \{\widehat{g}_{pq}\}$, corresponds to the obsolete number of users entering the transportation network at node p, whose final destination is node q, for all $(p, q) \in \mathcal{PQ}$. The objective of the ODA problem is to determine the estimated OD matrix denoted by $\mathbf{g} = \{g_{pq}\}$, which is close to matrix $\widehat{\mathbf{g}}$ and verifies measured observation of the flow volumes at some transit segments of the network, for $(p, q) \in \mathcal{PQ}$. While the updated OD matrix $\mathbf{g} = \{g_{pq}\}$ corresponds to the variables in our methodology, the reference matrix $\widehat{\mathbf{g}} = \{\widehat{g}_{pq}\}$ values are data known beforehand, for $(p, q) \in \mathcal{PQ}$.

When traveling on public transportation, the route passengers follow is determined by the transit lines they board. Sometimes, when traveling along a particular section of their route, passengers can choose from a set of transit lines with equivalent travel times. If, for instance, passengers board the first available transit line, their waiting time can be reduced, which will ultimately improve their total travel time. As a result, people waiting at a transportation hub to board a line from a set of available lines will be distributed based on the frequency of service. From now on, we will refer to the transit line segments that passengers can use for their entire journey as a "strategy". Following the first principle of Wardrop (1952), users do not consider a transit segment if the total travel time increases. Let $S_{pq} \subseteq \mathcal{A}$ be the subset of transit segments (strategy) travelers may take from p to q, with $(p, q) \in \mathcal{PQ}$. Thus, for each pair of nodes, there could be several routes that the user may take. After solving a transit assignment problem (see Figure 4), we can determine the proportion of trips traveling by each transit segment $a = (l, i, j) \in S_{pq}$, π_{pq}^{a}, from p to q, $(p, q) \in \mathcal{PQ}$. This proportion is $\pi_{pq}^{a'} = 0$ for segment $a' \notin S_{pq}$. The obsolete link proportions are used as references and considered parameters in this study, but we allow a slight deviation from them.

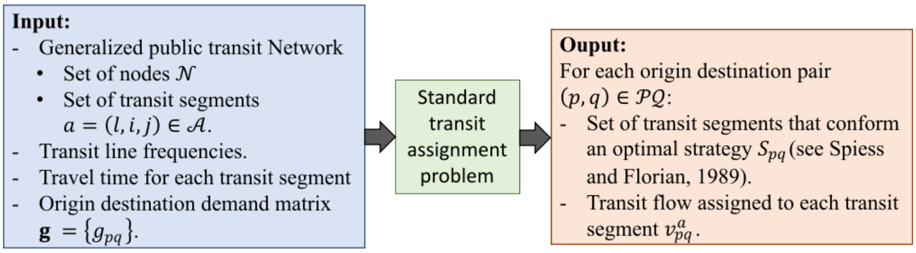

Input:
- Generalized public transit Network
• Set of nodes $\mathcal{N}$
• Set of transit segments $a = (l, i, j) \in \mathcal{A}$.
- Transit line frequencies.
- Travel time for each transit segment
- Origin destination demand matrix $\mathbf{g} = \{g_{pq}\}$.
Standard transit assignment problem
Ouput:
For each origin destination pair $(p, q) \in \mathcal{PQ}$:
- Set of transit segments that conform an optimal strategy S_{pq} (see Spiess and Florian, 1989).
- Transit flow assigned to each transit segment v_{pq}^{a}.

Fig. 4 Transit assignment diagram.

We rely on observed flow volumes ϕ^a of travelers at some transit segments $a \in \Phi$ to update the OD matrix. Note that $\Phi \subset \mathcal{A}$. The ODA problem can now be formally stated: find the OD matrix $\mathbf{g}$ that minimizes the difference between this matrix and the reference OD matrix $\widehat{\mathbf{g}}$ such that the flow volumes $\phi^{a'}$ in the observed links $a' \in \Phi$ are verified, with a maximal deviation of ε from the link probabilities π_{pq}^{a} for all $a \in A$, $(p, q) \in \mathcal{PQ}$.

4 Mixed-integer linear programming model for the ODA problem

To formulate a mixed-integer linear programming model for the ODA problem, we must determine the OD matrix variable $\mathbf{g} = \{g_{pq}\}$ with $(p, q) \in \mathcal{PQ}$. These variables are the estimated values of the OD matrix: the estimated users from the origins to the destinations.

In this work, we use the minimization of the absolute distance between $\mathbf{g}$ and the reference OD matrix $\widehat{\mathbf{g}}$ to allow the new demand to reproduce the observed flow volumes at specific transit segments in Φ as an objective function. We use two sets of variables to control the difference between the reference OD matrix and the estimated one to obtain a linear objective function. The excess integer variables E_{pq} with $(p, q) \in \mathcal{PQ}$ indicate more users from p to q. Thus, $\widehat{g}_{pq} < g_{pq}$ and in this manner, this excess is defined as $E_{pq} = \max\{g_{pq} - \widehat{g}_{pq}, 0\}$. Similarly, we introduce deficit variables D_{pq} with $(p, q) \in \mathcal{PQ}$ for the case where there are fewer users from p to q, that is, $\widehat{g}_{pq} \geq g_{pq}$.

Therefore, $D_{pq} = \max\{\widehat{g}_{pq} - g_{pq}, 0\}$. Note that when $D_{pq} > 0$ then $E_{pq} = 0$, and vice versa.

The objective function of the MILP for the ODA problem is to obtain an estimated OD matrix $\mathbf{g}$ as close as possible to the reference one $\hat{\mathbf{g}}$ by minimizing the total sum of the excess and the deficits of the estimated OD matrix $\mathbf{g}$:

$$\min \sum_{(p,q)\in\mathcal{PQ}} \alpha D_{pq} + \beta E_{pq}. \tag{1}$$

where linear parameters α and β allow the decision maker to prefer the user's excess or deficits. For example, for a city whose population has been growing over the years, we may expect that there will be more users in most of the OD matrix entries, thus $\beta < \alpha$. Similarly, a rural zone may be experiencing a population decrease, thus $\beta > \alpha$.

To linearly express the deficits and the excess of the estimated OD matrix, we need the following equations for each $(p, q) \in \mathcal{PQ}$:

$$D_{pq} \geq \widehat{g}_{pq} - g_{pq}, \tag{2}$$

$$E_{pq} \geq g_{pq} - \widehat{g}_{pq}. \tag{3}$$

The ODA problem updates the OD matrix and determines the volume of people traveling throughout each link $a \in S_{pq} \subset A$ that connects p with q, $(p, q) \in \mathcal{PQ}$. Hence, we introduce integer variables v^a_{pq} to indicate the actual number of people going from p to q using segment $a = (l, i, j) \in S_{pq}$. As mentioned before, there are some transit segments $a' \in \Phi \subset \mathcal{A}$ where the flow of passengers ϕ^a is observed and counted. These observations are our most important tool for updating the OD matrix. We do not know the passengers' origin or destination using this segment. We only have that the sum of all volumes should be equal to the observations:

$$\phi^{a'} = \sum_{(p,q)\in\mathcal{PQ}|a'\in S_{pq}} v_{pq}^{a'}, \quad \text{for } a' \in \Phi. \tag{4}$$

After solving a transit assignment problem (Spiess and Florian, 1989), the proportion of users that take each transit segment for each origin-destination pair $\pi_{pq}^a = v_{pq}^a/g_{pq}$ is computed. The passenger flow traveling from p to q that uses link $a \in \mathcal{A}$ is obtained by multiplying the total number of trips g_{pq} by the proportion π_{pq}^a, that is, $\pi_{pq}^a g_{pq} = v_{pq}^a$. Nevertheless, by using this equation, the actual passenger flow, in some cases, might yield inconsistencies with the assignment probabilities. That implies that the link proportions may have suffered a small deviation. Indeed, the assignment problem may establish a proportion of 0.6 for some transit segments, but in reality, it may be 0.59. This difference may be due to aspects the modeler does not consider, such as a more realistic user behavior or changes in the network's operative factors like service frequencies or delays. Thus, these proportions need slight adjustments to reflect the actual volumes. Therefore, we compute them as follows for $a \in S_{pq}$ and every $(p,q) \in \mathcal{PQ}$:

$$\lfloor \max\{(\pi_{pq}^a - \varepsilon), 0\} \; g_{pq} \rfloor \leq v_{pq}^a, \tag{5}$$

$$\lceil \min\{(\pi_{pq}^a + \varepsilon), 1\} \; g_{pq} \rceil \geq v_{pq}^a, \tag{6}$$

with $\varepsilon \geq 0$. Interval $[\max\{(\pi_{pq}^a - \varepsilon), 0\}, \min\{(\pi_{pq}^a + \varepsilon), 1\}]$ represents the allowed deviation from the obsolete link proportions. The $\lfloor\cdot\rfloor$ denotes the floor function in constraints (5) and $\lceil\cdot\rceil$ the ceiling function in constraints (6). Also, notice that we are not enforcing equality since we have the floor operator and positive values of ε. After solving our ODA MILP model, we obtain the updated OD matrix and the updated link proportions that fit the observed flow in the network.

Then, we must handle the network flow constraints. The sum of the flow volumes at origin p in the set of origin nodes $\mathcal{P}$ must equal the number of users entering the transport network at this node, as constraints (7) state. Similarly, with constraints (8), all flow volumes arriving at destination q in the set of destination nodes $\mathcal{Q}$ equal the total number of users ending their journey at q. Flow conservation at every node is guaranteed by constraints (9): the flow entering node $k \in \mathcal{N} \setminus \{p, q\}$ must be equal to the flow leaving it.

$$\sum_{l \in \mathcal{L}} \sum_{\{i | (l,p,i) \in S_{pq}\}} v_{pq}^{(l,p,i)} = g_{pq}, \qquad (p,q) \in \mathcal{PQ}, \tag{7}$$

$$\sum_{l \in \mathcal{L}} \sum_{\{i | (l,i,q) \in S_{pq}\}} v_{pq}^{(l,i,q)} = g_{pq}, \qquad (p,q) \in \mathcal{PQ}, \tag{8}$$

$$\sum_{l \in \mathcal{L}} \sum_{\{i | (l,i,k) \in S_{pq}\}} v_{pq}^{(l,i,k)} = \sum_{l \in \mathcal{L}} \sum_{\{j | (l,k,j) \in S_{pq}\}} v_{pq}^{(l,k,j)}, \qquad k \in \mathcal{N} \setminus \{p,q\}, (p,q) \in \mathcal{PQ}. \tag{9}$$

Valid inequalities strengthen a MILP formulation since they do not cut any feasible integer solution but make the solution space polyhedron closer to the integer solution convex hull (Wolsey, 1998; Schrijver, 1998). Thus, we introduce valid inequalities (10) to our MILP to decrease the computational running time of the branch-and-bound algorithms without compromising the optimality of the solution since it bounds the volume of each arc by the total number of persons going from p to q:

$$v_{pq}^{a} \leq g^{pq}, \; a \in S_{pq}, (p,q) \in \mathcal{PQ}. \tag{10}$$

Inequalities (10) are valid by definition; preliminary results show a slight advantage of using them in terms of the difference between the real and estimated OD-matrix. Notice that by imposing a positive integer value on the volumes, we also ensure the integrality of the estimated OD values of the matrix and the excess and deficits. Thus,

D_{pq} and E_{pq} may be defined as real variables, but they will take integer values. This is formally stated by (11)-(13).

$$v_{pq}^{a} \in \mathbb{Z}^{+}, \qquad (p,q) \in \mathcal{PQ}, a \in \mathcal{A}, \tag{11}$$

$$\delta_1 \hat{g}_{pq} \leq g_{pq} \leq \delta_2 \hat{g}_{pq} \in \mathbb{R}^{+}, \qquad (p,q) \in \mathcal{PQ}. \tag{12}$$

$$D_{pq}, E_{pq} \in \mathbb{R}^{+}, \qquad (p,q) \in \mathcal{PQ}, \tag{13}$$

where δ_1 and δ_2 are constants known by the user to bound $\mathbf{g}$ and remain close to $\widehat{\mathbf{g}}$. For example, a census or some statistical information may estimate that a particular population has grown no more than 10%.

To summarize, we denote as the ODA-MILP(ε) the MILP model of the ODA problem with a parameter ε such that it minimizes objective function (1) subject to constraints (2)-(13). Our methodology consists of starting with $\varepsilon = 0$ and then increasing it by 0.02 units until a feasible solution for the ODA-MILP(ε) is reached. In this manner, we obtain an estimated OD matrix and the actual flow volumes or the actual link proportions.

5 Experimental results

In this section, we generate a set of random instances that mimic real transport networks to validate our methodology, the ODA-MILP(ε) model. Section 5.1 describes these randomly generated matrices $\bar{\mathbf{g}}$ and how we perturb them to obtain the reference ones. Section 5.2 compares the ODA-MILP(ε) with the augmented Lagrangian method introduced in Chávez-Hernández et al (2019). Finally, in Section 5.3, we compare our methodology with the bilevel programming approach presented in Wu and Lam (2006) for the example transit network shown in Figure 1, and finally, we apply our proposed methodology to the benchmark instance based on the Swiss network with 15 nodes and 21 edges and the network design proposed by Cervantes-Sanmiguel et al (2023).

The general scheme of the comparison process we use in this study to validate the ODA-MILP(ε) model is depicted in Figure 5. We start with the real matrix $\bar{\mathbf{g}}$; Section 5.1 explains how to generate it. This matrix is usually unknown, but we consider that we are in an ideal case where we know it to validate our approach. Then, we perturb the real matrix to obtain the reference or obsolete matrix $\hat{\mathbf{g}}$. Finally, we obtain $\mathbf{g}$ using the ODA-MILP(ε), which estimates the real matrix. Two questions must be validated. First, we must assess how close the reference OD matrix $\hat{\mathbf{g}}$ is to the estimated one $\mathbf{g}$. That would verify the mathematical model's correctness and ensure the previous knowledge of the population dynamics. Second, we must assess how close the real OD matrix $\bar{\mathbf{g}}$ is to the estimated one $\mathbf{g}$. This is the most challenging question.

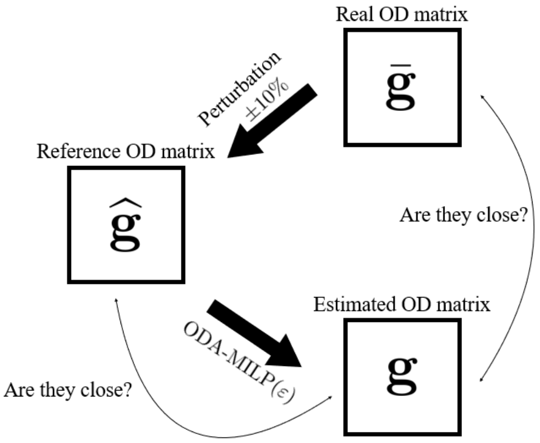


Fig. 5 Comparison process to validate the ODA-MILP(ε) model.

For the ODA-MILP(ε), the excess and deficit parameters of the objective function (1) are set to $\alpha = 1$ and $\beta = 1$ for all instances. Thus, no previous knowledge about the dynamics of centroids is known in advance. The parameters in equations (12) that bound the estimated OD matrix values are set to $\delta_1 = 0.9$ and $\delta_2 = 1.1$.

The ODA-MILP(ε) was coded in Python 3.7 and solved with a branch-and-bound implementation by the linear solver Gurobi 8.1 with the default algorithmic parameters. All experiments were executed in a computer with an Intel(R) Core(TM) i7 processor and 12 GB RAM.

5.1 Randomly generated instances

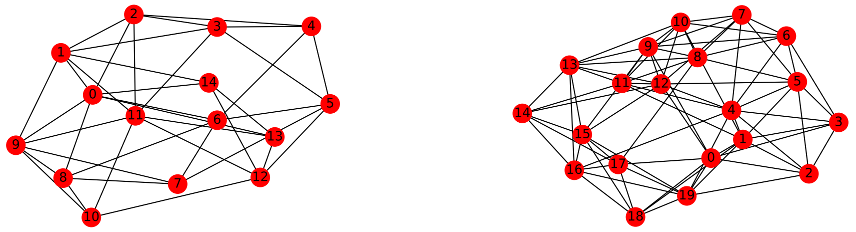
The generation of public transportation instances that mimic real networks is an active research area. Public transportation networks have special properties, such as growing evolutionarily, being embedded into two-dimensional space, having small-world properties, and having hierarchical organization. Based on von Ferber et al (2007); Chatterjee et al (2016); Sienkiewicz and Hołyst (2005), we generate a set of random instances containing the matrices corresponding to the real OD matrices $\bar{\mathbf{g}}$. All our instances and results can be found online at: https://doi.org/10.6084/m9.figshare.13838819.

Each instance representing a public transit system is composed of the real OD matrix $\bar{\mathbf{g}}$, the reference OD matrix $\hat{\mathbf{g}}$, its associated directed multigraph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the set of nodes, $\mathcal{A}$ is the multiset of directed links between the lines in $\mathcal{L}$, and the link proportions per segment in $\mathcal{A}$. The following methodology to generate the set of instances is used in this study.

1. The exact OD matrices $\bar{\mathbf{g}}$ are random integers between $[0, 500|\mathcal{N}|]$ for each pair $(p, q) \in \mathcal{PQ}$. The diagonal entries are all zeros.
2. A Newman-Watts-Strogatz small-world graph (Newman and Watts, 1999) is generated (with the Python package NetworkX (Hagberg et al, 2008), which can be used as a generator for small-world graphs and calculating shortest paths) by creating a ring over $|\mathcal{N}|$ nodes. Each node in the ring is connected with its $k = \lceil 0.3|\mathcal{N}| \rceil$ nearest neighbors (or $k - 1$ neighbors if k is odd), and the probability of adding new arcs is set to 0.3. The resulting Newman-Watts-Strogatz small-world graph has undirected edges, as the two graphs with 15 and 20 nodes shown in Figure 6.

3. Now that we have a graph that resembles a public transportation network, we form the transit lines (corresponding to buses, underground, or any other transit mode) $|\mathcal{L}|$. For each OD pair of nodes $(p, q) \in \mathcal{PQ}$, we compute all the non-intersecting routes between them and select the $|\mathcal{L}|$ ones with the shortest number of segments. If there are fewer than $|\mathcal{L}|$ non-intersecting routes, we choose them all. In this manner, each route is associated with a line l and an edge (i, j) belonging to the non-intersecting shortest routes between (p, q). Notice that the lines may visit all the nodes.
4. We establish the same frequency for all the lines. The link proportions π^a_{pq} for each $a \in S_{pq}$ are then computed along the (p, q) OD pair routes by solving the standard transit assignment problem of Spiess and Florian.
5. To generate the reference matrices $\hat{\mathbf{g}}$, 15% of the OD pairs of the exact OD matrix $\bar{\mathbf{g}}$ are randomly selected and uniformly perturbed by ±10%. The OD pairs not selected have the same value in the real matrix $\bar{\mathbf{g}}$ and the reference one $\hat{\mathbf{g}}$. These instances are named *Instances-ED*.
6. We compute the segment flows v^a_{pq}, $a \in \mathcal{A}, (p, q) \in \mathcal{PQ}$ using the link probabilities, for each matrix $\bar{\mathbf{g}}$ and $\hat{\mathbf{g}}$. All the segment flows in set *Instances-ED* have been observed. Set *Instances-ED*$^{1/2}$ comprises the same instances, but only half of the transit segments are observed this time.
7. Another set of instances is generated to test that the link probabilities are modified. This time, the real matrix and the reference one are equal, so the reference demand entries are not perturbed. Nevertheless, 15% of the segment flows at the network segments are perturbed by ±10%, but all are still observed. The resulting instances, named *Instances-ε*, aim to show that the assignment probabilities may differ from the initial ones and must be modified with the demand OD matrix.

In this manner, we have generated 201 instances with the number of nodes in the transit network between [4,20] and transit lines between [1,5]: 67 instances in the



(a) 15 nodes.

(b) 20 nodes.

Fig. 6 Newman-Watts-Strogatz small-world graphs with 15 and 20 nodes.

Instances-ED set, 67 instances in the *Instances-ED*$^{1/2}$ set, and 67 instances in the *Instances-*ε one. Although our instances might seem oversimplified, it is important to notice that their use is to evaluate the size of the instances that our model can solve with an exact method.

5.2 Experimental results for the ODA problem

We compare the ODA-MILP(ε) performance with the augmented Lagrangian methodology of Chávez-Hernández et al (2019), which is based on an iterative dual ascent technique and the Lagrangian multipliers method. Their approach yields solutions with low CPU time when applied to large-scale networks.

To evaluate the performance of our model, we use the root mean square error, RMSE, to interpret deviations in the same scale as the variables. For instance, the RMSE between the exact matrix $\bar{\mathbf{g}}$ and the one obtained by our model $\mathbf{g}$ is calculated as follows:

$$\mathrm{RMSE}(\bar{\mathbf{g}}, \mathbf{g}) = \sqrt{\frac{1}{n} \sum_{(p,q)\in\mathcal{PQ}} (\bar{g}_{pq} - g_{pq})^2},$$

where n is the number of OD pairs.

Table 3 shows the comparison results of the ODA-MILP(ε) methodology and the Augmented Lagrangian algorithm of Chávez-Hernández et al (2019) for the *Instances-ED* set. The ODA-MILP(ε) is parameterized with $\varepsilon = 0$, sufficient for these instances to find a feasible and optimal solution. Later, this parameter will be flexible for the set *Instances-ε*. The first and second columns of Table 3 correspond to the number of nodes $|\mathcal{N}|$ and lines $|\mathcal{L}|$ in the transit system. The root mean squared error between the reference matrix $\bar{\mathbf{g}}$ and the estimated one $\mathbf{g}$ is in the columns "RMSE($\bar{\mathbf{g}}, \mathbf{g}$)". Columns with "RMSE($\hat{\mathbf{g}}, \mathbf{g}$)" are the root mean squared error between the exact demand $\hat{\mathbf{g}}$ and the estimated one $\mathbf{g}$. The columns labeled "RMSE($\bar{v}, v$)" correspond to the root mean squared error between the observed and the estimated segment flow volumes. Finally, column "time" is the CPU time in seconds to solve the instance with each methodology. Each line in this table averages all the instances with the same nodes and lines. For example, the first line represents the average of the instances with four to nine nodes and with a single line. The last line is the average of all instances.

Table 3 ODA-MILP(ε) methodology with $\varepsilon = 0$ and Augmented Lagrangian algorithm of Chávez-Hernández et al (2019) for the *Instances-ED* set.

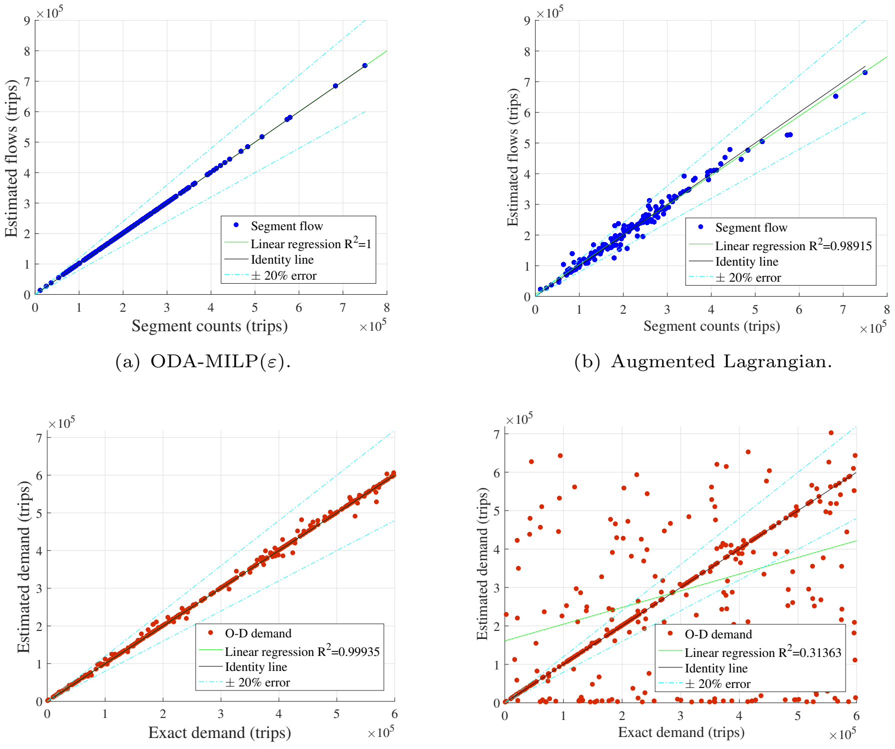
		ODA-MILP(ε)				Augmented Lagrangian algorithm			
$\|\mathcal{N}\|$	$\|\mathcal{L}\|$	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{v}, v$)	time	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{v}, v$)	time
4-9	1	182.36	134.66	0.00	0.01	3327.32	3325.85	2782.14	0.01
	2	418.32	318.34	0.00	0.01	7053.29	7042.63	1073.07	0.00
	3	408.50	303.20	0.00	0.01	11904.13	11895.27	2789.14	0.05
10-15	1	972.13	706.25	0.00	0.02	23758.67	23754.84	9960.47	0.11
	2	1731.19	1223.90	0.00	0.02	38827.60	38769.26	10119.73	0.37
	3	2437.99	1404.72	0.00	0.03	55952.64	56148.93	8939.14	0.75
	4	3674.35	2510.42	0.00	0.04	104932.99	104909.87	7688.41	2.34
16-20	1	1399.07	746.53	0.00	0.03	33624.09	33592.08	19351.70	1.73
	2	2697.02	1919.76	0.00	0.05	74077.07	74214.09	11373.41	4.18
	3	3939.14	2303.91	0.00	0.06	109488.89	109556.34	14999.98	4.35
	4	5203.55	3212.61	0.00	0.08	129385.39	129667.00	18705.03	5.14
	5	6816.88	5045.99	0.00	0.08	181655.45	181777.15	15671.14	9.67
	Av.	2490.04	1652.52	0.00	0.04	64498.96	64554.44	10287.78	2.39

As we can observe from Table 3, the best results are for the ODA-MILP(ε) method in terms of root mean square error and time. Indeed, contrary to the augmented Lagrangian, the estimated matrices obtained with the ODA-MILP(ε) method are closer to the real ones than the reference ones are. The difference of the flow volumes equals zero for the ODA-MILP(ε) method since the model tries to reproduce exactly this behavior with equations (4). Notice that the augmented Lagrangian method does not reproduce closely the observed segment flows. The larger the instances, the larger the root square mean errors for both methods. Remarkably, the execution time for the ODA-MILP(ε) method is better than the augmented Lagrangian algorithm, which is intended for large instances. Although the resolution time for the ODA-MILP(ε) is less than one second, the model construction is time-consuming. Indeed, the number of variables is $\mathcal{O}(|\mathcal{N}|^3 + 3|\mathcal{N}|^2)$ while the number of restrictions is $\mathcal{O}(3|\mathcal{N}|^3 + |L||\mathcal{N}|^2 + 5|\mathcal{N}|^2)$. A research line is then about the data structures, preprocessing algorithms, and dominant solution properties to increase the size of the instances. In terms of percentages, considering the largest RMSE obtained in the networks with 16-20 and 5 lines, we improved the quality of the estimated average number of trips by 97% (we compared the ODA-MILP(ε) and the Augmented Lagrangian corresponding RMSE($\bar{\mathbf{g}}, \mathbf{g}$).)

Figure 7 compares the ODA-MILP(ε) methodology with the Augmented Lagrangian algorithm of Chávez-Hernández et al (2019). Figures 7(a) and 7(b) show the scatter plots of the segment flows. In contrast, Figures 7(c) and 7(d) show the scatter plots for the estimated trip OD demand matrix concerning the real one, for an instance with 20 nodes and five lines of the *Instances-ED* set.

Figure 7 shows that the ODA-MILP(ε) recovers almost entirely both the real OD matrices and the observed segment flow volumes. This is not the case for the Augmented Lagrangian method, where there is a relatively small scattering in the transit segment flows but a large one in demand OD matrices.

Most of the time, not all the flows of every transit segment can be observed. Thus, we compare the ODA-MILP(ε) performance when only half of the transit network links have been observed. The results of the *Instances-ED*$^{1/2}$ set are displayed in Table 4. It has the same structure as Table 3.



(a) ODA-MILP(ε).

(b) Augmented Lagrangian.

(c) ODA-MILP(ε).

(d) Augmented Lagrangian.

Fig. 7 The ODA-MILP(ε) compared to the Augmented Lagrangian algorithm of Chávez-Hernández et al (2019) presented with scatter plots of the segment flows, (a) and (b), and of the OD matrices, (c) and (d), for an instance with 20 nodes and five lines of the *Instances-ED* set.

Table 4 shows that the ODA-MILP(ε) cannot exactly reproduce the observed segment flows since the RMSE($\bar{v}, v$) are no longer zero as for the *Instances-ED* set. By comparing the largest RMSE with the ODA-MILP(ε) and the Augmented Lagrangian, we are improving the average error in the estimated trips by 95.5%. The differences between the estimated OD matrix and the real or the reference ones are larger. This behavior is expected since it has less information about the network structure and more degrees of freedom. Although the computational time is still short, it takes slightly longer than considering observations at all the segments. In the case of the augmented Lagrangian, we can see that the RMSE($\bar{v}, v$), the RMSE($\bar{\mathbf{g}}, \mathbf{g}$), and the computational

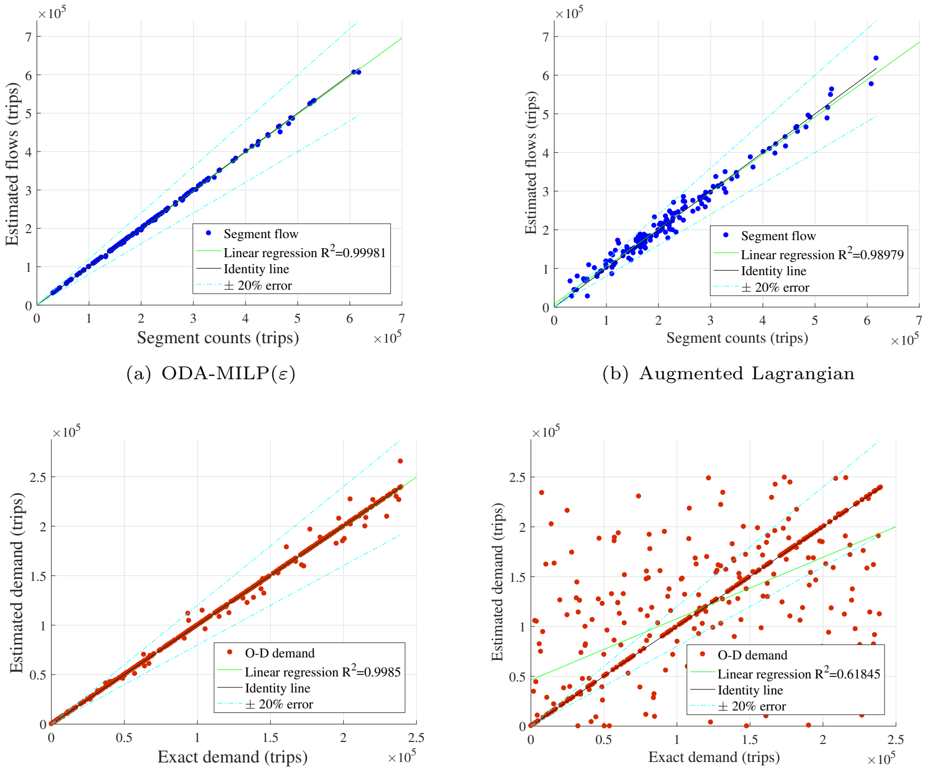
Table 4 ODA-MILP(ε) methodology and the Augmented Lagrangian algorithm of Chávez-Hernández et al (2019) for the *Instances-ED*$^{1/2}$ set.

		ODA-MILP(ε)				Augmented Lagrangian algorithm			
$\|\mathcal{N}\|$	$\|\mathcal{L}\|$	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{v}, v$)	time	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{v}, v$)	time
4-9	1	518.72	546.77	417.96	0.05	3626.40	3626.40	498.12	0.00
	2	404.27	483.43	252.36	0.15	6543.11	6543.11	1054.17	0.00
	3	336.24	464.14	191.27	0.15	11968.09	11968.09	1285.14	0.00
10-15	1	3086.89	3393.05	1941.77	0.84	17620.94	17620.94	1082.77	0.01
	2	2131.2	2533.54	1411.99	2.57	32150.13	32150.13	11373.13	0.00
	3	3318.08	3552.34	1101.28	4.19	56788.61	56788.61	6179.43	0.02
	4	4381.59	4979.35	1287.87	4.72	94491.07	94491.07	7521.56	0.15
16-20	1	3570.02	3771.38	1498.87	3.11	22171.57	22171.57	10761.14	0.02
	2	3225.94	3462.36	1738.42	6.78	70228.30	70228.30	10263.14	0.26
	3	4538.72	4807.14	1631.43	11.67	121422.94	121422.94	8709.94	0.63
	4	5421.6	6399.04	1688.64	13.9 4	133037.24	133037.24	17501.44	0.54
	5	6959.75	8255.14	1713.79	20.94	185924.72	185924.72	17467.71	1.27
	Av.	3157.75	3553.97	1239.64	5.76	62997.76	62997.76	7808.14	0.24

time decrease concerning the values obtained with the *Instances-ED* set. The difference in the performance of both methodologies can be explained due to the different hypotheses made for each model. For example, for the augmented Lagrangian method, the variables are assumed to be continuous, and the observed flows can differ from those calculated with the new demand matrix. Furthermore, it is assumed that the segment flows are obtained by solving a linear assignment problem (the proportions of the arcs do not change regardless of the demand matrix). In the case of the model proposed in this work, it is assumed that the variables are discrete, and it is forced to reproduce precisely the flows observed in the segments from the new matrix obtained and the adjustment in the arc proportions. The dual ascent and Lagrange multipliers methods are accurate and efficient, especially in large-scale problems with continuous variables. These assumptions are no longer valid in small networks with integer values, and gradient-based iterative algorithms are no longer a good option. The feasible solution regions (convex hulls) of the discrete and continuous cases for the same instance have a discrepancy. Thus, the discrete optimum is underestimated or overestimated by the computational solution.

Furthermore, in the augmented Lagrangian and multiplier methodology, it is considered that the link flows can be obtained from the product of a matrix (containing the proportions of trips in each link for each OD pair without considering the capacity limits of the vehicles) times a vector (containing the travel demand between each OD pair) and the Karush-Kunh-Tucker optimality conditions are formulated assuming that the entries of the proportion matrix are constant. In cases where vehicle capacity limits play an essential role in the distribution of trips across the network, the link proportions are expected to change with the demand. This can be seen when a passenger adds more segments to the strategy (considers more options) due to the congestion that some transit lines present, so segments that previously had no flow now carry a small proportion. Thus, assuming that the arc proportions obey an assignment problem without capacity limits would lead us to maintain the same structure of the probability matrix, which may not represent the reality of the problem. The inconsistency between link probabilities and the assignment model is more significant for the quadratic continuous model. In this case, a total variation diminishing model and non-smooth regularization might be better for dealing with high gradients that may arise in the number of users.

Figure 8 is similar to Figure 7 but for one instance of the *Instances-ED*$^{1/2}$ set with 20 nodes and 2 lines. The scatter plots for the segment flows are depicted in Figures 8(a) and 8(b), while Figures 8(c) and 8(e) show the scatter plots of the estimated demand. The differences are slight, although we do not obtain a perfect fit between the observed volumes and those calculated with the ODA-MILP(ε). The adjustment in both segment flows and demand of the ODA-MILP is imperfect. Still, the dispersion between the reference values and the estimates is smaller than that obtained with the Augmented Lagrangian. Furthermore, for the augmented Lagrangian, we can see that although the estimated volumes remain relatively close to the observed ones, in general, the estimated demand is far from the exact solution.



(a) ODA-MILP(ε)

(b) Augmented Lagrangian

(c) ODA-MILP(ε)

(d) Augmented Lagrangian

Fig. 8 Scatter plots of the volumes, (a) and (b), and of the OD matrices, (c) and (d), for an instance with 20 nodes and two lines of the *Instances-ED*$^{1/2}$ set.

Our previous experimental results yield an $\varepsilon = 0$, meaning that the link probabilities have not changed. Frequently, users may follow different routes because of network changes due to variations in transit line service frequencies or longer travel times due to street congestion. An example of the network of Figure 1 is presented in Table 2. In our model, this phenomenon is modeled by constraints (5) and (6), where the new proportion of trips traveling on each transit segment $a \in \mathcal{A}$ may slightly change concerning the proportion π_{pq}^{a} traveling from p to q, $(p, q) \in \mathcal{PQ}$ obtained previously from a transit assignment, which can be the transit assignment of Spiess and Florian, the all or nothing assignment or any other.

Therefore, we now test the *Instances-ε* set where we parametrically change the values of ε until the problem is feasible. These results are presented in Table 5. As for the previous tables, the first and second columns correspond to the number of nodes $|\mathcal{N}|$ and the number of lines $|\mathcal{L}|$ in the transit system. The third column is the ε parameter value needed to obtain a feasible estimated OD matrix. In the ODA-MILP(ε) method, we start with $\varepsilon = 0$ and then iteratively increase it by 0.02 until we obtain a feasible solution. The last column shows the time in seconds needed by the ODA-MILP(ε) method. The values shown at each line of this table represent averages. Notice that we do not report the RMSE values since our instances were constructed to force the link proportions to change, such that the real and the reference matrix are close. Moreover, the augmented Lagrangian method cannot deal with these instances since it does not modify the link probabilities. Indeed, it does not converge to any solution. For example, suppose that the probability of a segment $a \in \mathcal{A}$ from the OD pair (p, q) is $\pi^a_{pq} = 0.5$. Suppose the ODA-MILP(ε) yields a $\varepsilon = 0.05$. In that case, the updated probability is now in the interval [0.475,0.525], and it can be computed once we have the estimated OD matrix and the corresponding segment flows.

Table 5 Values of ε to obtain a feasible solution and time in seconds for the ODA-MILP(ε) methodology for the *Instance-ε* set.

$\|\mathcal{N}\|$	$\|\mathcal{L}\|$	ε	time
4-9	1	0.13	0.00
	2	0.04	0.00
	3	0.09	0.00
10-15	1	0.05	0.01
	2	0.04	0.01
	3	0.04	0.02
	4	0.03	0.03
16-20	1	0.08	0.01
	2	0.03	0.03
	3	0.02	0.10
	4	0.03	0.06
	5	0.08	0.15
	Av.	0.05	0.04

Table 5 shows that the ODA-MILP(ε) methodology can adjust the ε parameter to consider that the trip distribution over the transit network is made differently than before. With this consideration, we can obtain an OD matrix that coincides with the real one. Moreover, the computational time does not increase, and the link probability variation is slight.

5.3 Instances from the literature

This section compares our methodology with the bilevel programming approach presented in Wu and Lam (2006). Also, we present some experiments on the transit network designed by Cervantes-Sanmiguel et al (2023) based on Mandl's Swiss network.

5.3.1 Network of Wu and Lam (2006)

Let us consider again the example transit network shown in Figure 1. We applied the linear transit assignment of Spiess and Florian (1989) with constant travel times to distribute a demand of 200 trips for each of the OD pairs (1,2), (1,4), (3,2), and (3,4). This yields the flow pattern shown in Figure 9, from which we compute the link proportions that we will refer to as obsolete proportions. They may differ from the observed ones in a more realistic scenario.

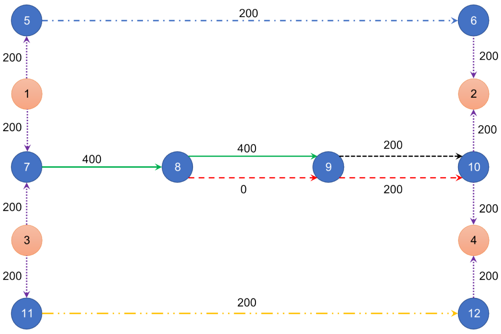


Fig. 9 Segment flow resulting from a linear transit assignment.

In the instance considered in Wu and Lam (2006), after performing a stochastic user equilibrium assignment, they get the flow over some route sections, which they consider as the observed flows. We deduced the flows for some transit network segments using the route sections they defined. Figure 10 shows the flows we computed by rounding off the mentioned results to obtain integer observed flows.

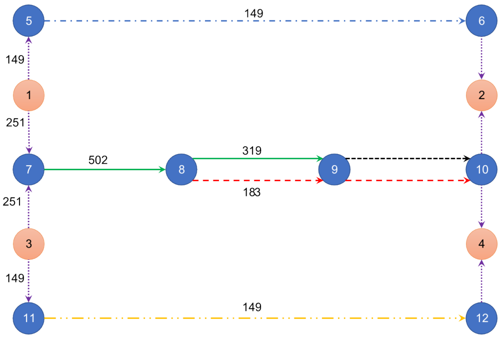


Fig. 10 Observed link flow resulting from a stochastic user equilibrium transit assignment.

By using the above information, we can compute some link proportions. Let us note that the number of users from node 1 to node 2 is 200; however, only 149 users are observed in the link $(1, 5, 6)$; this means that the other 51 users left node 1 using line 2, therefore $\pi_{1,2}^{(1,5,6)} = 149/200$ and $\pi_{1,2}^{(2,7,8)} = 51/200$. Similarly, it can be deduced that $\pi_{3,4}^{(5,11,12)} = 149/200$ and $\pi_{3,4}^{(2,7,8)} = 51/200$. Also, for the OD pairs (1,4) and (3,2), we have that the only option to leave the origin node is by boarding line 2, therefore $\pi_{1,4}^{(2,7,8)} = \pi_{3,2}^{(2,7,8)} = 1.0$. These results were considered as observed link proportions and are compared with those from solving the ODA-MILP problem with $\varepsilon = 0.3$ (estimated proportions) and shown in Table 6. In this case, the value of ε was computed iteratively, starting with 0.0 and increasing by 0.1, until our MILP model found a feasible solution. As we can see, the new proportions changed by around 25% concerning the ones obtained from the linear assignment and are less than 5% close to the real ones.

Table 6 ODA-MILP(ε) methodology with $\varepsilon = 0.3$ for the instance in Wu and Lam.

(p, q)	(l, i, j)	Obsolete proportion	Observed proportion	Estimated proportion
(1,2)	(1,5,6)	1.000	0.745	0.703
(1,2)	(2,7,8)	0.000	0.255	0.297
(1,4)	(2,7,8)	1.000	1.000	1.000
(3,2)	(2,7,8)	1.000	1.000	1.000
(3,4)	(2,7,8)	0.000	0.255	0.297
(3,4)	(5,11,12)	1.000	0.745	0.703

Table 7 Resulting estimated demand.

(p, q)	Obsolete demand	Exact demand	Estimated demand	Resulting demand in Wu and Lam
(1,2)	250	200	212	209.5
(1,4)	150	200	225	191.1
(3,2)	150	200	151	191.1
(3,4)	200	200	212	209.5

Table 8 RMSE for both methodologies comparing the deviation between the OD demand matrices and the deviation of the observed flows.

Model	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{v}, v$)
Wu and Lam	40.80	9.21	9.76
ODA-MILP(ε)	46.14	28.78	0.00

In this way, our model exactly reproduces the observed link flows of Figure 10. Regarding the estimated demand, the results are shown in Table 7.

With this information, the deviation of the estimated OD matrix with both methodologies against the obsolete and exact matrix can be computed. Table 8 shows the root mean square error of the estimated demand concerning the obsolete demand (RMSE($\hat{\mathbf{g}}, \mathbf{g}$)) and the exact demand (RMSE($\bar{\mathbf{g}}, \mathbf{g}$)), and the deviation of the estimated segment/route-section flow concerning the observed ones. Recall that in Wu and Lam (2006), they consider the flow over route sections as observations, while in this work, we consider the flow in some transit segments; however, in our results, the observed flows are reproduced exactly, so the deviation is null.

Although with our model, both deviations in demand are more significant than those obtained in Wu and Lam (2006), our model is competitive because it exactly reproduces the observed segment flows; this is because, in our model, we are forced to reproduce precisely the observed data so their accuracy strongly influences our estimates.

5.3.2 Transit network of Cervantes-Sanmiguel et al (2023)

Mandl's network is a well-known benchmark Swiss network (Mandl, 1980) from which Cervantes-Sanmiguel et al (2023) designed a transit network that minimizes the trade-off travel times and monetary costs for passengers. This transit network consists of 15 nodes and 5 transit lines that yield 64 transit segments, as shown in Figure 11 with the respective travel times and headways in minutes.

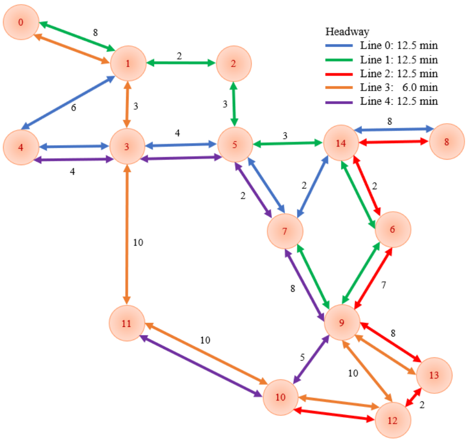


Fig. 11 Cervantes-Sanmiguel et al (2023) transit network.

For this set of experiments, we computed the shortest path between each OD pair. In those cases where the shortest path is not unique, we distributed the proportion of trips equally. For instance, let us observe that to go from node 6 to node 3, a user may board line 1, alight at node 5, and then continue the trip by boarding either line 1 or line 4; in this case, we set $\pi_{6,3}^{(1,6,14)} = \pi_{6,3}^{(1,14,5)} = 1$ and $\pi_{6,3}^{(0,5,3)} = \pi_{6,3}^{(4,5,3)} = 0.5$.

Table 9 Deviation results for the Cervantes-Sanmiguel et al transit network.

Transit segments (%)	RMSE($\hat{\mathbf{g}}, \mathbf{g}$)	RMSE($\bar{\mathbf{g}}, \mathbf{g}$)	time
80	1.35	1.58	0.018
70	0.95	1.31	0.020
60	0.94	1.34	0.032
50	1.06	1.53	0.056
40	0.87	1.42	0.027
30	1.12	1.51	0.037
20	0.63	1.59	0.047
10	0.53	1.56	0.056

As for the exact demand, we used the one reported in Cervantes-Sanmiguel et al (2023). Then, we generated the observed segment flows and the reference OD matrix as described in Section 5.1.

First, we experimented assuming that observations in all transit segments were available and tested different parameter ε values in our model as before; however, no feasible solution was found. Then, we considered only 90% of the segments with observations and repeated the procedure; again, no solution was found. We continued reducing the percentage of segments with observations by ten until we reached 10% of the segments with observations. In all cases, the optimal solution is obtained for $\varepsilon = 0.5$. Table 9 reports the RMSE between the reference matrix and the estimated matrix (column 2), the RMSE between the exact matrix and the estimated matrix (column 3), and the CPU time in seconds (column 4). Observe that the CPU time increases as the number of segments with observations decreases; this behavior of the model can also be observed in Tables 3 and 4. There is no clear tendency for the deviation measures between the reference, exact, and estimated demand. But we can observe that while the shortest/largest deviation (0.53/1.35) between the reference demand and the estimated one is obtained with 10%/80% of the transit segments with observations, the shortest/largest deviation (1.31/1.59) between the exact demand and the estimated one is obtained with 70%/20% of the transit segments with observations.

6 Conclusions

We solve the inverse problem of estimating the actual OD matrix based on a reference one and some flow observations at the transit segments. Indeed, OD matrices are relevant for different purposes; for instance, they are relevant for bus line design and the generation of useful timetables for adding new trips when drivers do not show up or when there are accidents, and the network should be rapidly restored. Moreover, updating OD matrices allows us to test the current system under more demanding scenarios and adapt it to future demand infrastructure.

An integer linear programming model was presented to estimate the OD matrix, simultaneously fitting the segment probabilities from a reference OD matrix and observed transit segment flows. We compare the performance of the proposed model with the augmented Lagrangian model previously introduced by Chávez-Hernández et al (2019). The results have shown that the ODA-MILP(ε) offers high-quality solutions for all the tested instances. Compared to the methodologies in the literature, the scatter plots of the demand and the segment flows are considerably lower than those obtained with other approaches. Moreover, the execution times are shorter with the ODA-MILP(ε). Also, we programmed one of the few instance generators that mimic transit networks to test the methodology presented in this paper.

Although our model considers only slight changes in the demand matrix and the link proportions, most authors only consider a change in the demand. During the COVID-19 pandemic, in most cities, there was a mobility reduction that can be seen as a decrease in the number of trips represented on an OD matrix; also, to avoid contracting the disease, people try to reduce their contact time with others, and the transit services modify the frequency of service, these changes have a direct impact in the segment proportions. The ODA-MILP(ε) can model the phenomenon, and more experiments should be carried out in scenarios where both the demand and the probabilities change. Therefore, it is an issue to handle scenarios with more substantial changes. Besides, our approach could be improved by indicating the assignment probability difference for each OD pair and each transit segment.

Our results for relatively small networks take a computational cost of less than one minute. Nevertheless, reading data and building the model before starting the branch-and-bound solver is the most time-consuming task. Therefore, this suggests the following areas of opportunity for further research. First, designing and developing specific data structures such as arrays or linked lists to improve RAM use in reading the instances. Second, an adequate preprocessing procedure must be developed to improve the CPU time to construct the model for the Gurobi solver. Third, developing dominant solution properties to improve the efficiency and scalability of handling large instances. This work considers integer variables for the OD matrix estimation, but a challenging area is handling the underlying data errors without rounding the values of the variables.

Finally, these results were obtained from instances generated as described in Section 5.1. This generator can be modified so that the link proportions represent an equilibrium assignment (Spiess and Florian, 1989) for cases without congestion and consider heuristic models to represent cases with congestion and capacity limits in transport vehicles.

Declarations

The authors want to acknowledge the Mexican Council for Science and Technology(CONACyT) for supporting M. Victoria Chávez-Hernández and Roger Z. Ríos-Mercado under its Frontier Science Program, grant FC2016/1948, and for supporting Yasmín A. Ríos Solís, grant 710289. M. Victoria Chávez-Hernández did part of the research at the Universidad Autónoma de Nuevo León. M. Victoria Chávez-Hernández is pleased to acknowledge the financial support of Instituto Tecnológico

Autónomo de México (ITAM), the support of Asociación Mexicana de Cultura A.C., the National System of Researchers (SNI).

References

Ashok K, Ben-Akiva ME (2002) Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation Science 36(2):184–198

Bera S, Rao K (2011) Estimation of origin-destination matrix from traffic counts: the state of the art. European Transport Trasporti Europei 49:3–23

Boyer V, Ibarra-Rojas OJ, Ríos-Solís YÁ (2018) Vehicle and crew scheduling for flexible bus transportation systems. Transportation Research Part B: Methodological 112:216–229

Caceres N, Romero L, Benitez FG (2020) Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys. Transportmetrica A: Transport Science 16(3):574–601

Cantelmo G, Viti F (2020) A big data demand estimation model for urban congested networks. Transport and Telecommunication 21(4):245–254

Cascetta E (1984) Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. Transportation Research Part B: Methodological 18(4–5):289–299

Cascetta E (2009) Transportation Systems Analysis: Models and Applications, Springer Optimization and Its Applications, vol 29, 2nd edn. Springer, New York

Castillo E, Jiménez P, Menéndez JM, et al (2013) A Bayesian method for estimating traffic flows based on plate scanning. Transportation 40(1):173–201

Ceder A (2015) Public Transit Planning and Operation: Modeling, Practice and Behavior. CRC Press, Boca Raton

Cervantes-Sanmiguel K, Chavez-Hernandez M, Ibarra-Rojas O (2023) Analyzing the trade-off between minimizing travel times and reducing monetary costs for users in the transit network design. Transportation Research Part B: Methodological 173:142–161

Chatterjee A, Manohar M, Ramadurai G (2016) Statistical analysis of bus networks in India. PLoS ONE 11(12):e0168478

Chávez-Hernández MV, Juárez Valencia LH, Ríos-Solís YA (2019) Penalization and augmented Lagrangian for O-D demand matrix estimation from transit segment counts. Transportmetrica A: Transport Science 15(2):915–943

von Ferber C, Holovatch T, Holovatch Y, et al (2007) Network harness: Metropolis public transport. Physica A: Statistical Mechanics and its Applications 380:585–591

Fisk CS (1989) Trip matrix estimation from link traffic counts: The congested network case. Transportation Research Part B: Methodological 23(5):331–336

Fisk CS, Boyce DE (1983) A note on trip matrix estimation from link traffic count data. Transportation Research Part B: Methodological 17(3):245–250

Florian M, Chen Y (1995) A coordinate descent method for the bi-level O/D matrix adjustment problem. International Transactions on Operations Research 2(2):165–179

Frederix R, Viti F, Tampère CMJ (2013) Dynamic origin-destination estimation in congested networks: theoretical findings and implications in practice. Transportmetrica A: Transport Science 9(6):494–513.

García-Ródenas R, Verastegui-Rayo D (2013) Adjustment of the link travel-time functions in traffic equilibrium assignment models. Transportmetrica A: Transport Science 9(9):798–824

Ge L, Kliewer N, Nourmohammadzadeh A, et al (2022) Revisiting the richness of integrated vehicle and crew scheduling. Public Transport Forthcoming, DOI: 10.1007/s12469-022-00292-6

Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proceedings of the 7th Python in Science Conference (SciPy 2008), Pasadena, pp 11—15

He Y, Zhao Y, Tsui KL (2023) Short-term forecasting of origin-destination matrix in transit system via a deep learning approach. Transportmetrica A: Transport Science 19(2):2033348

Ibarra-Rojas OJ, Giesen R, Ríos-Solís YA (2014) An integrated approach for timetabling and vehicle scheduling problems to analyze the trade-off between level of service and operating costs of transit networks. Transportation Research Part B: Methodological 70:35–46

Ibarra-Rojas OJ, Delgado F, Giesen R, et al (2015) Planning, operation, and control of bus transport systems: A literature review. Transportation Research Part B: Methodological 77:38–75

Landmark AD, Arnesen P, Södersten CJ, et al (2021) Mobile phone data in transportation research: Methods for benchmarking against other data sources. Transportation 48(5):2883–2905

Liao Y, Yeh S, Gil J (2022) Feasibility of estimating travel demand using geolocations of social media data. Transportation 49:137–161

Liu S, Fricker JD (1996) Estimation of a trip table and the θ parameter in a stochastic network. Transportation Research Part A: Policy and Practice 30(4):287–305

Lo HP, Chan CP (2003) Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts. Transportation Research Part A: Policy and Practice 37(9):771–788

López-Ospina H, Cortés CE, Pérez J, et al (2022) A maximum entropy optimization model for origin-destination trip matrix estimation with fuzzy entropic parameters. Transportmetrica A: Transport Science 18(3):963–1000

Mahmoodjanlou A, Hazelton ML, Parry K (2019) Apples versus oranges? comparing deterministic and stochastic day-to-day traffic assignment models. Transportmetrica B: Transport Dynamics 7(1):1426–1443

Mandl CE (1980) Evaluation and optimization of urban public transportation networks. European Journal of Operational Research 5(6):396–404

Newman MEJ, Watts DJ (1999) Renormalization group analysis of the small-world network model. Physics Letters A 263(4):341–346

Pitombeira-Neto AR, Loureiro CFG, Carvalho LE (2018) Bayesian inference on dynamic linear models of day-to-day origin-destination flows in transportation networks. Urban Science 2(4):1–17

Schrijver A (1998) Theory of Linear and Integer Programming. Wiley, New York

Sienkiewicz J, Hołyst JA (2005) Statistical analysis of 22 public transport networks in Poland. Physical Review E 72(4):046127

Spiess H, Florian M (1989) Optimal strategies: A new assignment model for transit networks. Transportation Research Part B: Methodological 23(2):83–102

Walpen J, Lotito PA, Mancinelli EM, et al (2020) The demand adjustment problem via inexact restoration method. Computational and Applied Mathematics 39(3):204

Wardrop JG (1952) Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1(3):325–362

Wolsey LA (1998) Integer Programming. Wiley, New York

Wu Z, Lam W (2006) Transit passenger origin-destination estimation in congested transit networks with elastic line frequencies. Annals of Operations Research 144:363–378

Yang H, Sasaki T, Iida Y, et al (1992) Estimation of origin-destination matrices from link traffic counts on congested networks. Transportation Research Part B: Methodological 26(6):417–434

Yang H, Meng Q, Bell MGH (2001) Simultaneous estimation of the origin-destination matrices and travel-cost coefficient for congested networks in a stochastic user equilibrium. Transportation Science 35(2):107–123

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