Dynamic capacity planning for demand-responsive multimodal transit

Bernardo Martin-Iradi

Francesco Corman

Nikolas Geroliminis

STRC conference paper 2024 **April 30, 2024**

STRC | **24th Swiss Transport Research Conference**
Monte Verità / Ascona, May 15-17, 2024

Dynamic capacity planning for demand-responsive multimodal transit

Bernardo Martin-Iradi
IVT
ETH Zurich
bernardo.martin-iradi@ivt.baug.ethz.ch,

Francesco Corman
IVT
ETH Zurich
francesco.corman@ivt.baug.ethz.ch

Nikolas Geroliminis
Urban Transport Systems Laboratory
EPFL, Lausanne
nikolas.geroliminis@epfl.ch

April 30, 2024

Abstract

Demand-responsive multimodal transit offers opportunities to complement existing public transport systems and provide an overall better service level to passengers while at the same time making better use of the resources. This study optimizes the capacity of such system by strategically sizing the required fleet and allocating it to the operating services. We formulate a two-stage stochastic optimization model that plans the transit system and the required fleet in the first stage, and optimizes the demand-responsive operations in the second stage. We develop a decomposition-based method that exploits the network-based formulation of the second stage, allowing us to solve practical instances. Preliminary results from a case study in the city of Zurich show that designing a public transport system together with demand-responsive mobility systems can benefit both transport operators and passengers. By allocating the system capacity more efficiently, operators reduce operational costs while maintaining or improving the travel experience for passengers.

Keywords

Multimodal transit; Public transport; On-demand mobility; Stochastic optimization

Suggested Citation

Contents

List of Tables

List of Figures

1 Introduction

This work studies a mixed transport system where public transport (i.e., transit) and on-demand vehicles are optimized in a synchronized and integrated manner. This type of system, also known as On-Demand Multimodal Transit System (ODMTS), has been actively studied in recent years (Basciftci and Van Hentenryck, 2020; Bertsimas *et al.*, 2020; Dalmeijer and Van Hentenryck, 2020; Steiner and Irnich, 2020; Calabrò *et al.*, 2023). The rise of digitalization capabilities, the need to address new and changing commuting patterns, and the higher customer expectations underscore the potential value of such systems. Pilot projects have proven the viability of ODMTS in practice (Van Hentenryck *et al.*, 2023), but modeling and planning these systems at scale remains a challenge, and most of the studies consider different assumptions to be able to solve such models (Banerjee *et al.*, 2021; Sumalee *et al.*, 2011). Among others, the large-scale models focus on a deterministic scenario and consider homogeneous fleets within modal systems (Lienkamp and Schiffer, 2023). One of the areas for improvement from ODMTS is dimensioning the system capacity accurately. Running half-empty vehicles or overcrowded ones hinder the offered service level to passengers. Adjusting the frequency of the transit lines is one way of adapting the system capacity to the demand. However, this may not be enough, and it is also susceptible to unexpected variations in demand. We believe that planning and sizing a heterogeneous fleet can be a more efficient approach to optimize the system capacity and offer a better level of service. This also fits with the needs of many transit operators to renew their fleets and the model we present can be seen as an opportunity to make such decisions effectively.

The public transport system is seen as the core part of an ODMTS where demand-responsive vehicles act as a complement to transit rather than a replacement. Sizing the system's fleet is a strategical decision, whereas operating on-demand services is an operational one as they do not rely on planning in advance. This brings us to formulate the system as a two-stage model formulation in which we make strategic decisions in the first stage (i.e., system fleet sizing and scheduling) and operational decisions in the second stage (i.e., planning of on-demand vehicles and passenger routing) to respond to demand uncertainty.

We address two main research questions in this study:

1. Can a model for ODMTS with heterogeneous fleet planning decisions and stochasticity in demand lead to better passenger service levels and lower transport operations costs?

2. What are the benefits this model can bring and how can it operate in practice?

The first question is about how can we leverage the uncertainty realization and additional operational flexibility to plan the system capacity more efficiently. The second question is about quantifying the value of integrating on-demand mobility systems into public transportation line and timetable planning. This value addresses all the actors involved: (i) operational costs for operators, (ii) passenger level of service for customers, and (iii) environmental and congestion impact for the society.

This study aims at making the following contributions:

1. New model: *Demand-responsive multimodal transit with heterogenous fleet and stochastic demand*: A two-stage stochatic optimization model.
2. Efficient exact algorithm: A solution method based on double (i.e., Benders and Dantzig-Wolfe) decomposition.
3. High-quality solutions on large real-life instances.
4. Practical impact: Benefits of integrated multimodal planning versus independent planning of public transport and demand-responsive services.

2 Problem description

We present a two-stage stochastic problem formulation that defines the transit schedule and its required fleet in the first stage, and plans the on-demand vehicles and passenger routing at the trip level in the second stage.

2.1 First stage

The first stage problem covers strategical decisions including, (i) which schedule and (ii) with which fleet should we operate the transit system, and (iii) which transit lines and/or on-demand vehicle serve each passenger origin-destination. It minimizes both transit operational costs and the expected travel costs of passengers.

We consider a set of public transport lines $\mathcal{L}$ to operate (e.g., bus or tram) formed by a set of stations S_ℓ, that can be operated using vehicles of types $\mathcal{B}$. We denote $\mathcal{F}_{\ell b}$ to the set of frequencies that line $\ell \in \mathcal{L}$ could operate using bus type $b \in \mathcal{B}$. Moreover, we predefine a set of possible schedules $\mathcal{P}_{\ell f}$ for such frequency $f \in \mathcal{F}_{\ell tb}$ as well as the number of vehicles required $N_{\ell fb}$. Each schedule is formed by the set of services $\mathcal{H}_p$ and each bus type $b \in \mathcal{B}$ has capacity C_b^B. For the passengers, we denote OD to the entire sets of possible origin-destination pairs, and let D_d be the expected demand flow on origin-destination $d \in OD$. Passengers can travel using the transit systems but also cover part or the whole trip using on-demand transport.

Based on this setup, we define $c_{d\ell}^L$ and c_{dk}^V as the travel cost of serving the passenger flow $d \in OD$ using line $\ell \in \mathcal{L}$ and served partly by an on-demand vehicle, respectively. This can be seen as the fixed cost related to sizing the system fleet. Finally, to account for operational costs we define $c_{\ell fbp}^Q$ as the cost of operating line $\ell \in \mathcal{L}$ running at frequency $f \in \mathcal{F}_{\ell b}$ using vehicles of type $b \in \mathcal{B}$ using schedule $p \in \mathcal{P}_{\ell tf}$. Each of these costs has an associated variable in the model, namely in order, $z_{d\ell}$ (binary variable that denotes if OD pair $d \in OD$ is served by line $l \in \mathcal{L}$), v_d (integer variable that defines how many on-demand partial legs are needed to serve OD pair $d \in OD$), and $q_{\ell fbp}$ (a binary variable that denotes if line $\ell \in \mathcal{L}$ runs with frequency $f \in \mathcal{F}_{\ell tb}$ using vehicles of type $b \in \mathcal{B}$ using schedule $p \in \mathcal{P}_{\ell f}$).

The first-stage minimizes in the objective function Eq. (1) the *a priori* costs of serving origin-destinations by a transit system and on-demand vehicles and the operational costs of the chosen transit schedule. We find the schedule of each transit line in Eq. (2), we size the system to the expected demand in Eq. (3), and Eqs. (4)- (6) define the nature of variables.

The first-stage model looks like this:

$$\min \sum_{d \in OD} (c_d^W w_d + c_d^V v_d + \sum_{\ell \in \mathcal{L}_d} c_{d\ell}^Z z_{d\ell}) + \sum_{\ell \in \mathcal{L}} \sum_{f \in \mathcal{F}_{\ell b}} \sum_{b \in \mathcal{B}} \sum_{p \in \mathcal{P}_{\ell f}} c_{\ell fbp}^Q q_{\ell fbp} + \mathbb{E}[Y(\boldsymbol{z}, \boldsymbol{q}, \boldsymbol{v}), \mathcal{S}] \tag{1}$$

s.t.

$$\sum_{b \in \mathcal{B}} \sum_{f \in \mathcal{F}_{\ell tb}} \sum_{p \in \mathcal{P}_{\ell tf}} q_{\ell ftbp} = 1 \quad \forall \ell \in \mathcal{L}, t \in \mathcal{T} \tag{2}$$

$$\sum_{d \in OD_\ell} D_d z_{d\ell} \leq \sum_{b \in \mathcal{B}} \sum_{f \in \mathcal{F}_{\ell b}} \sum_{p \in \mathcal{P}_{\ell f}} C_b^B N_{\ell fk} q_{\ell fbp} \quad \forall \ell \in \mathcal{L} \tag{3}$$

$$v_d \in \mathbb{Z}^+ \quad \forall d \in OD \tag{4}$$

$$z_{d\ell} \in \{0,1\} \qquad \forall d \in OD, \ell \in \mathcal{L}_d \tag{5}$$

$$q_{\ell fbp} \in \{0,1\} \qquad \forall \ell \in \mathcal{L}, b \in \mathcal{B}, f \in \mathcal{F}_{\ell b}, p \in \mathcal{P}_{\ell f} \tag{6}$$

2.2 Second stage

The second stage tackles the operations of the transit and on-demand services and the routing of all the passengers. This stage is divided into a set of scenarios $\mathcal{S}$, each scenario $s \in \mathcal{S}$ corresponding to a different realization of the passenger demand with probability π_s. We define the discrete set of passenger requests $\mathcal{R}_s$ in scenario $s \in \mathcal{S}$. Each passenger request $r \in \mathcal{R}_s$ is characterized by an origin o_r, destination d_r, request time t_r and number of passengers n_r.

To capture each passenger's trip, we use a graph $\mathcal{G}$ with the set of nodes $\mathcal{V}$ and arcs $\mathcal{A}$. These node and arc sets can be divided into the following subsets.

The node set $\mathcal{V}$ is divided into the following subsets:

- *Artificial source and sink nodes*: These nodes represent the origin and destination of a passenger's path. We denote $\hat{o}_{rs}$ and $\hat{d}_{rs}$ to the origin and destination node for passenger $r \in \mathcal{R}_s$ in scenario $s \in \mathcal{S}$.
- *Location nodes*: These nodes represent a time and space coordinate, where the space is characterized by a location in the network (e.g., passenger origin/destination, or transit stop), and the time is characterized by a time instant in the planning horizon. These nodes are used to characterized the arrival or departure time of passengers to the different locations in the network.
- *Transit nodes*: These nodes represent the scheduled stops of each transit trip. This set can be divided into *arrival* and $\mathcal{V}^{BD}$ nodes. For each trip $p \in \mathcal{P}_{\ell tf}$ of line $\ell \in \mathcal{L}$,

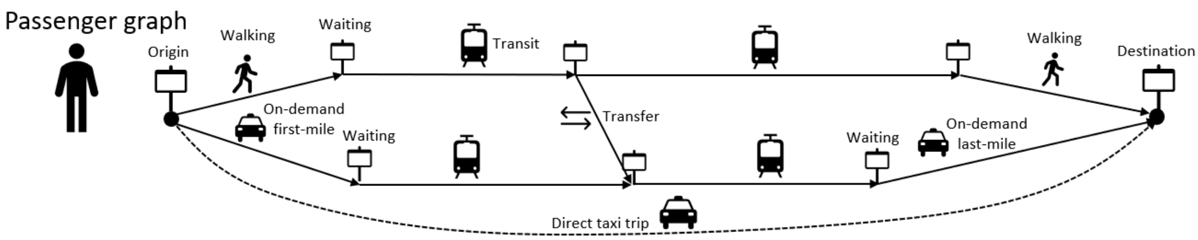
stopping location $n \in S_\ell$, and bus type $b \in \mathcal{B}$, there is one node for arrival time, and one node for the departure time.
- *Boarding nodes*: These nodes represent the change of transport mode when boarding a transit trip or an on-demand first/last mile trip.

The arc set $\mathcal{A}$ is also divided into different arc subsets. Following a similar approach to the traffic assignment using diachronic graphs from Gentile (2016), we differentiate the following types of arcs:

- *Source and sink arcs*: These arcs connect the artificial source (resp. sink) node of each request $r \in \mathcal{R}_s$ with the location node of its origin location o_r (resp. destination location d_r) and request time t_r (resp. arrival time).
- *Walk arcs*: These arcs represent passenger walking between their origin (respct. destination) and their nearest transit stops. These arcs connect two location nodes within walking distance and the arrival time is directly given by the starting walking time. The cost of the arc is computed as $\omega^{WK}\Delta^{WK}$ where ω^{WK} is the walking coefficient and Δ^{WK} is the walking time.
- *Driving arcs*: These arcs represent passenger using an on-demand vehicle as part of their trip (first-mile, last-mile, or door-to-door) and are modeled in the same way as the walking arcs. The cost of the arc is computed as $\omega^{DV}\Delta^{DV}$ where ω^{DV} is the driving coefficient and Δ^{DV} is the driving time. This coefficient is also used as a proxy to measure the operational (variable) costs of utilizing an on-demand vehicle. Remember that these arcs are associated with the first-stage variable v_d, except the door-to-door driving arc, that has a higher penalty cost, as we want to incentivize passenger to use the transit system as part of their trip.
- *Transit Boarding arcs*: These arcs represent waiting at a certain location and boarding a transit service. These arcs connect a location node set with the corresponding boarding node. The cost of the arc is computed as ω^{TB} where ω^{TB} is the on/off-boarding-time coefficient.
- *Transit running arcs*: These arcs represent the running time between stations of the transit services. Each arc connects the departure with the arrival transit node of two consecutive stations. The cost of the arc is computed as $\omega^{TR}\Delta^{TR}$ where ω^{TR} is the transit run-time coefficient and Δ^{TR} is the running time.
- *Transit dwell arcs* : These arcs represent the dwell time at stations of the transit services. Each arc connects the arrival with the departure transit node at the station. The cost of the arc is computed as $g_a = \omega^{TD}\Delta^{TD}$ where ω^{TD} is transit dwell-time coefficient and Δ^{TD} is the dwell time.
- *Transit transfer arcs*: These arcs represent transfers between line services at stations.

Each arc connects two transit nodes, where stops are the same, but the trips and vehicle types are different. The arrival time from the incoming service and the departure time of the outgoing service need to satisfy the minimum transfer time at the station. The cost of the arc is computed as $g_a = \omega^{TT}\Delta^{TT}$ where ω^{TT} is the transfer-time coefficient and Δ^{TT} is the transfer time.

Figure 1: Example of a passenger graph.



A simplified example of such network for a passenger is shown in Figure 1. The model has one set of binary variables to find the passenger trips. For each scenario $s \in \mathcal{S}$, let y_{ras} be 1 if arc $a \in \mathcal{A}$ is traversed by passenger request $r \in \mathcal{R}_s$. Associated with these variables, we have the corresponding costs $\hat{c}_{ras}$. These costs capture the generalized cost of travel, but also the variable costs of operating the fleet of on-demand vehicles as mentioned in each of the arc subsets. To ease the notation, for each passenger request $r \in \mathcal{R}_s$, we define $\mathcal{A}_r$ the set of arcs that the passenger can traverse. We denote $\mathcal{A}_r^+(i), \mathcal{A}_r^-(i)$ to the set of outgoing arcs from and incoming arcs to node $i \in \mathcal{V}$ respectively. We also define $\mathcal{A}(\ell)$ as the arcs related to transit line $\ell \in \mathcal{L}$, $\mathcal{A}^{DV}$ as the driving arcs, $\mathcal{A}(\bar{q}_{\ell fbp})$ as the arcs related to the services defined by variable $q_{\ell fbp}$, and $\mathcal{A}^+_{\ell fbph}(n)$ as the transit running arcs from station n corresponding to service h of schedule $p \in \mathcal{P}_{\ell f}$ from line $\ell \in \mathcal{L}$, frequency $f \in \mathcal{F}_{\ell tb}$ and vehicle of type $b \in \mathcal{B}$. The second-stage model is as follows:

$$\mathrm{Y}(\bar{z}, \bar{q}, \bar{v}, \mathcal{S}) = \min \sum_{s\in\mathcal{S}} \pi_s \Big(\sum_{r\in\mathcal{R}_s} \sum_{a\in\mathcal{A}_r} \hat{c}_{ras} y_{ras}\Big) \tag{7}$$

s.t.

$$\sum_{a\in\mathcal{A}_r^+(\hat{o}_{rs})} y_{ras} = 1 \qquad \forall r \in \mathcal{R}_s, s \in \mathcal{S} \tag{8}$$

$$\sum_{a\in\mathcal{A}_r^-(\hat{d}_{rs})} y_{ras} = 1 \qquad \forall r \in \mathcal{R}_s, s \in \mathcal{S} \tag{9}$$

$$\sum_{a \in \mathcal{A}_r^+(i)} y_{ras} - \sum_{a \in \mathcal{A}_r^-(i)} y_{ras} = 0 \qquad \forall i \in \mathcal{V} \backslash \{\hat{o}_{rs}, \hat{d}_{rs}\}, r \in \mathcal{R}_s, s \in \mathcal{S} \tag{10}$$

$$\sum_{a \in \mathcal{A}(\ell)} y_{ras} \leq \bar{z}_{(o_r,d_r)\ell} \qquad \forall r \in \mathcal{R}_s, \ell \in \mathcal{L}, , s \in \mathcal{S} \tag{11}$$

$$\sum_{r \in \mathcal{R}_s | (o_r,d_r)=d} \sum_{a \in \mathcal{A}^{DV} \backslash \{(\hat{o}_{rs}, \hat{o}_{ds})\}} y_{ras} \leq \bar{v}_d \qquad \forall s \in \mathcal{S}, d \in OD \tag{12}$$

$$\sum_{a \in \mathcal{A}(\bar{q}_{\ell fbp})} y_{ras} \leq \bar{q}_{\ell fbp} \qquad \forall r \in \mathcal{R}_s, \ell \in \mathcal{L}, b \in \mathcal{B}, f \in \mathcal{F}_{\ell b}, p \in \mathcal{P}_{\ell f}, s \in \mathcal{S} \tag{13}$$

$$\sum_{a \in \mathcal{A}^+_{\ell fbph}(n)} \sum_{r \in \mathcal{R}_s} n_r y_{ras} \leq C_b^B \bar{q}_{\ell fbp} \qquad \forall b \in \mathcal{B}, \ell \in \mathcal{L}, n \in S_\ell, f \in \mathcal{F}_{\ell b}, p \in \mathcal{P}_{\ell f}, h \in \mathcal{H}_p, s \in \mathcal{S} \tag{14}$$

$$y_{ras} \in \{0, 1\} \qquad \forall r \in \mathcal{R}_s, a \in \mathcal{A}_r, s \in \mathcal{S} \tag{15}$$

The objective function Eq. (7) minimizes the operational costs of the passenger door-to-door trips and the on-demand vehicles across scenarios. Eqs. (8)- (10) define a path for each passenger in the time-space graph $\mathcal{G}$. Eq. (11) only allows the passenger to use a transit line if its origin-destination has been assigned to it, and Eq. (12) only allow passengers to use the assigned fleet of on-demand vehicles. We can only choose transit arcs of the activated schedules in Eq. (13). Eq. (14) ensures capacity restrictions at the transit trip level and Eq. (15) define the binary nature of the variables. The second stage is always feasible as we assume that passengers can always be served by a door-to-door service.

3 Solution method

To solve large-scale instances of the model, we exploit the decomposability of problem, and in particular, of the second-stage graph. First, we can apply Benders decomposition to the entire two-stage problem. The first-stage problem becomes the Benders Master Problem (BMP) and the second-stage problem becomes the Benders Sub-problem (BSP). Leveraging this decomposition, we can reformulate the second-stage problem as a set partitioning formulation where variables refer to paths instead of arcs in the time-space graph. Each path corresponds to a passenger trip from origin to destination.

Given the complexity of the second stage, enumerating all path-based variables is not tractable. Therefore, we opt to generate them dynamically using column generation. Due to the reformulation, we only convexify Eqs. (8)- (10), meaning that the pricing problem (PP) is a shortest path problem in a directed and acyclic graph (DAG). Therefore, we can use efficient label-setting algorithms to solve them. We acknowledge that, when the PP is a pure shortest path problem, column generation does not provide a benefit in terms of root node relaxation (i.e., solving the root node with column generation and solving the relaxed version of the original formulation provide equal bounds), but we foresee to solve the root node faster.

Once the column generation does not provide any path with negative reduced cost and the Benders decomposition procedures converge, if the solution is still fractional, we solve the integer version of the BSP to guarantee an integer feasible solution. The BSP formulation is tight in itself, and therefore, we expect the integrality gap to also be tight. An overview of the solution method is depicted in Figure 2.

4 Preliminary results

We compare our two-stage stochastic model for ODMTS, with two different benchmarks. On one side, to assess the value of the stochastic optimization, we compare our model with a deterministic equivalent, and on the other side, to evaluate the value of integrating multiple modes of transport, we compare ODMTS with a system in which transit and on-demand services are planned separately.

Figure 2: Algorithm overview.

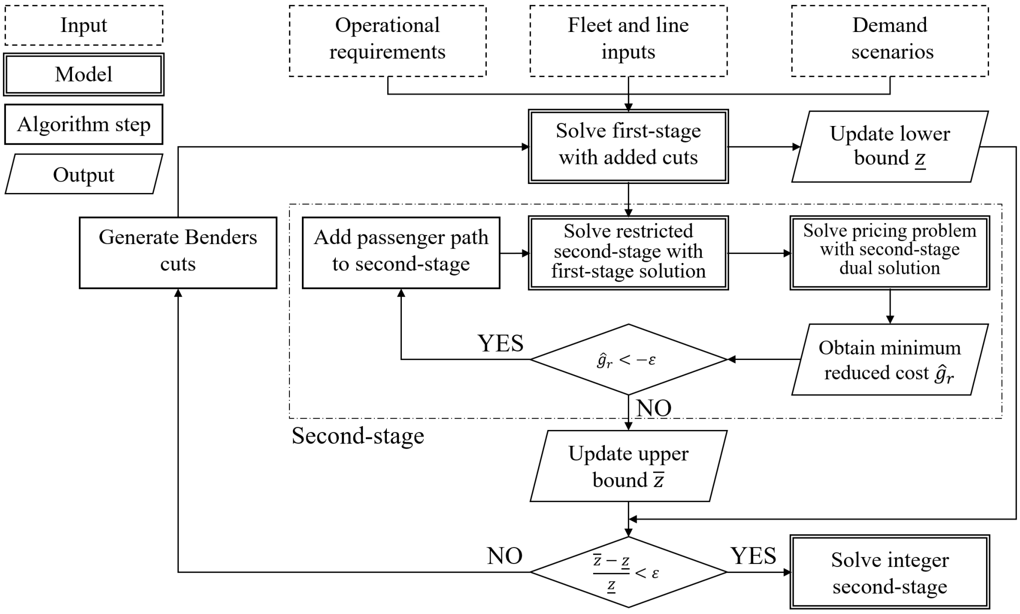


Table 1: Model benchmark comparison.

ODMTS vs.	Transit fleet costs	On-demand service distance	Transit users	Average passenger delay	Number of direct taxi trips
Deterministic equivalent	+0%	-25.9%	x5.8	+1.2%	-28.1%
Non-integrated system	+0%	-11.8%	x1.3	-1.7%	-18.6

The results in Table 1 show a comparison of the out-of-sample solutions in a relatively small test case (three lines, tens of stops, and hundreds of passengers) in the city of Zurich, Switzerland. Thanks to the additional planning flexibility, our model is able to drastically improve the transit ridership while maintaining a similar level of service and without the need to increase the system capacity. This translates in a significant reduction of the operating costs (i.e., distance) of on-demand services, which can also translate in reduced consgestion and pollution.

For the results conducted so far, the model could be solved directly using commercial solvers, but the decomposition-based method presented is expected to scale better and solve real-life city-scale instances efficiently. We also plan to conduct detailed assessments on the value of fleet heterogeneity and the potential of ride-pooling on first/last mile on-demand services.

In overall, results suggest that ODMTS can become a relevant solution in the current mobility ecosystem and provide benefits to users, operators and society as a whole.

5 References

Banerjee, S., C. Hssaine, Q. Luo and S. Samaranayake (2021) Plan your system and price for free: Fast algorithms for multimodal transit operations, *Available at SSRN 3972423*.

Basciftci, B. and P. Van Hentenryck (2020) Bilevel optimization for on-demand multimodal transit systems, *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, 52–68, ISSN 16113349, 03029743.

Bertsimas, D., Y. S. Ng and J. Yan (2020) Joint frequency-setting and pricing optimization on multimodal transit networks at scale, *Transportation Science*, **54** (3) 839–853, ISSN 15265447, 00411655.

Calabrò, G., A. Araldo, S. Oh, R. Seshadri, G. Inturri and M. Ben-Akiva (2023) Adaptive transit design: Optimizing fixed and demand responsive multi-modal transportation via continuous approximation, *Transportation Research Part A: Policy and Practice*, **171**, 103643, ISSN 0965-8564.

Dalmeijer, K. and P. Van Hentenryck (2020) Transfer-expanded graphs for on-demand multimodal transit systems, *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, 167–175, ISSN 16113349, 03029743.

Gentile, G. N. (2016) *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems, Springer Tracts on Transportation and Traffic*, Springer International Publishing, ISBN 3319250809, 3319250825, 9783319250809, 9783319250823.

Lienkamp, B. and M. Schiffer (2023) Column generation for solving large scale multi-commodity flow problems for passenger transportation, *European Journal of Operational Research*, ISSN 18726860, 03772217.

Steiner, K. and S. Irnich (2020) Strategic planning for integrated mobility-on-demand and urban public bus networks, *Transportation Science*, **54** (6) 1616–1639, ISSN 15265447, 00411655.

Sumalee, A., K. Uchida and W. H. Lam (2011) Stochastic multi-modal transport network under demand uncertainties and adverse weather condition, *Transportation Research Part C: Emerging Technologies*, **19** (2) 338–350, ISSN 0968-090X. Emerging theories in traffic and transportation and methods for transportation planning and operations.

Van Hentenryck, P., C. Riley, A. Trasatti, H. Guan, T. Santanam, J. A. Huertas, K. Dalmeijer, K. Watkins, J. Drake and S. Baskin (2023) Marta reach: Piloting an on-demand multimodal transit system in atlanta.