

Deviated Fixed-route Microtransit: Design and Operations

Kayla Cummings, Alexandre Jacquillat

Sloan School of Management and Operations Research Center, MIT

Bernardo Martin-Iradi

DTU Management, Technical University of Denmark

Microtransit offers opportunities to enhance urban mobility by combining the reliability of public transit and the flexibility of ride-sharing. This paper optimizes the design and operations of a deviated fixed-route microtransit system that relies on *reference lines* but is allowed to deviate in response to passenger demand. We formulate a *Microtransit Network Design (MiND)* model via two-stage stochastic optimization. The model features a tight second-stage formulation thanks to a subpath-based representation of microtransit operations in a load-expanded network, which optimizes on-demand deviations between checkpoint stops. We develop a double-decomposition algorithm combining Benders decomposition and subpath-based column generation armed with a tailored label-setting algorithm. Using real-world data from Manhattan, results suggest that our method scales to large practical instances, with up to 10-100 candidate lines and hundreds of stations. Comparisons with transit and ride-sharing benchmarks suggest that microtransit can provide win-win outcomes toward efficient mobility (high demand coverage, low operating costs, high level of service), equitable mobility (broad geographic reach) and sustainable mobility (limited environmental footprint). We provide an open-source implementation in an online repository to enable replication.

Key words: Microtransit, stochastic optimization, Benders decomposition, column generation

1. Introduction

Major cities face critical challenges to meet mobility needs in the midst of rising congestion, greenhouse gas emissions and socioeconomic inequalities. Static transit infrastructure offers limited flexibility to respond to ever-changing mobility needs, resulting in a ridership decline (The Economist 2018) and transit deserts (Allen 2017). Simultaneously, ride-sharing provides flexible, on-demand mobility services, but low-occupancy vehicles still lead to high fares, congestion, and emissions (Statista 2023, Schaller 2018). This context identifies opportunities to leverage emerging *microtransit* services toward efficient, equitable, and sustainable mobility. Broadly defined by the US DoT (2016) as “privately owned and operated shared transportation system(s) that can offer fixed routes and schedules, as well as flexible routes and on-demand scheduling,” microtransit shepherds the digital capabilities and operating flexibility of ride-sharing into the realm of public transit. Yet, microtransit raises critical questions about how to combine transit and ride-sharing components into low-cost, high-quality services and how to develop corresponding routing capabilities, thus requiring dedicated analytics and optimization capabilities (McKinsey & Co. 2018).

In response, this paper proposes a two-stage stochastic optimization approach to support the design and operations of microtransit services. We focus on *deviated fixed-route microtransit*, which relies on transit lines to consolidate passenger demand into high-capacity vehicles—as in public transit—while allowing on-demand routing deviations to provide convenient mobility options in response to passenger requests—as in ride-sharing. This paper aims to develop a methodology that jointly optimizes the network of transit lines at the strategic level and routing deviations at the operational level, to establish its scalability to large instances arising in practice, and to assess the performance of deviated fixed-route microtransit in the urban mobility ecosystem.

1.1. Theoretical Motivation, Practical Experience, and Literature Review

Transit planning. Our first-stage problem designs a microtransit network to maximize demand coverage and passenger level of service. This relates to the vast literature on transit planning optimization (see Desaulniers and Hickman (2007) for a review, and Ortega et al. (2018), Wei et al. (2022), Sun et al. (2023) for recent contributions). These problems have often been solved with heuristics (see, e.g., Ceder and Wilson 1986, Walteros et al. 2015). Exact methods were initially confined to small instances with 10-25 stops (Wan and Lo 2003, Marín and Jaramillo 2009). Borndörfer et al. (2007) proposed a column generation algorithm for an incremental network design problem. Bertsimas et al. (2021) developed a column generation methodology for a comprehensive network design problem, which scales to instances with hundreds of stops and thousands of edges.

In our problem, on-demand deviations introduce linking constraints between first- and second-stage decisions, which complicates the use of column generation to dynamically generate candidate lines. Instead, we design a pre-processing procedure to define candidate lines—a common approach in the literature and in practice (Ceder and Wilson 1986, Steiner and Irnich 2020). Our methodology still employs column generation in the second stage to optimize on-demand deviations.

Ride-sharing. The convenience of on-demand mobility has enabled staggering growth in the ride-sharing market (McKinsey & Co. 2021). Yet, the reliance on small-occupancy vehicles leads to high fares along with contributions to congestion and greenhouse gas emissions. This paper seeks complementary ways of integrating on-demand mobility into higher-capacity microtransit vehicles.

Our second-stage problem serves discrete passenger requests with time windows and capacitated vehicles—a vehicle routing problem with a challenging integer optimization structure. To retain a tight second-stage formulation, we propose a network representation building upon the vehicle-shareability network for fleet sizing from Vazifeh et al. (2018), the vehicle-sharing network for single-occupancy ride-sharing from Bertsimas et al. (2019b), and the ride-pooling network from Zhang et al. (2023). Our paper contributes a tailored load-expanded subpath-based network that exploits the structure of deviated fixed-route microtransit to encode on-demand routing deviations.

Microtransit. Microtransit aspires to combine the economies of scale of public transit with the flexibility of ride-sharing. One possible model is to design a joint system combining fixed-route transit and ride-sharing, which Chopra et al. (2023) frame via dual sourcing. Another model is to provide on-demand door-to-door transportation with high-capacity vehicles, which Alonso-Mora et al. (2017) optimize via a request-trip-vehicle network. However, on-demand high-capacity operations may induce detours, delays and long travel times. Blanchard et al. (2023) showed that the optimal latency in a probabilistic traveling repairman problem grows with the size of the geographic area and the dispersion of customers, and also grows at a supra-linear rate of $\Theta(n\sqrt{n})$ where n is the number of customers. The convexity of this function reflects negative spatial externalities across customers induced by on-demand operations with high-capacity—in fact, negative spatiotemporal externalities given the temporal discrepancies among trip requests.

This theoretical result outlines two approaches to alleviate spatiotemporal externalities in on-demand mobility: restricting vehicle occupancy—as in ride-sharing—or operating in small or concentrated areas. In practice, on-demand microtransit operations with high-occupancy vehicles have been successful in small municipalities¹ and university campuses.² In larger regions, microtransit needs to restrict the scope of on-demand operations to alleviate spatiotemporal externalities. Two prominent strategies under this umbrella are zone-based and deviated fixed-route microtransit.

Zone-based microtransit restricts the geographic locale of on-demand operations to limit the detours and delays required to pick up or drop off passengers. For instance, MetroConnect operates in four areas of Miami, and Metro Micro operates in eight areas of Los Angeles.³ To serve longer trips, zone-based microtransit often acts as a first- and last-mile feeder into fixed-route transit.⁴ This has generated research on multimodal operations (Maheo et al. 2019, Steiner and Irnich 2020, Banerjee et al. 2021, Guan et al. 2023) and design (Cummings et al. 2023, Silva et al. 2022).

In practice, multimodal microtransit can expand the catchment area of transit systems (Ma et al. 2019, Stiglic et al. 2018). Yet, they introduce complexity to establish first- and last-mile zones that can achieve coverage at scale; for instance, DART’s GoLink service in Dallas partitions the service region into 32 zones.⁵ Fundamentally, such partitioning raises similar trade-offs as in door-to-door microtransit. On the one hand, small zones and low-occupancy vehicles lead to high operating costs and contributions to congestion and emissions. On the other hand, larger zones and higher-capacity vehicles may induce detours and delays in first- and last-mile operations.

¹ See, e.g., <https://city.ridewithvia.com/salem-skipper>, <https://city.ridewithvia.com/newmo-newton>

² See, e.g., <https://ridewithvia.com/news/northeastern-university-taps-via-to-power-new-on-demand-safety-shuttle>

³ <https://city.ridewithvia.com/go-connect-miami>, <https://micro.metro.net>

⁴ See examples in Dallas (<https://dart.org/guide/transit-and-use/golink>) and Atlanta (<https://www.itsmartaa.com>)

⁵ <https://www.dart.org/guide/transit-and-use/golink>

Deviated fixed-route microtransit, in contrast, consolidates demand into high-occupancy vehicles along transit routes while allowing on-demand deviations in response to passenger requests. This model leverages *virtual bus stops* to consolidate pickups and dropoffs in central locations.⁶ On-demand operations with virtual bus stops induce challenging routing problems (see Zhang et al. (2023) in ride-sharing and Bertsimas et al. (2019a) in public school bus systems). Viewed through this lens, deviated fixed-route microtransit leverages transit lines to act as a natural regularization of on-demand operations. In practice, it has been implemented in paratransit for passengers with limited mobility,⁷ in a pilot in Kansas City with ten 14-passenger shuttles (Westervelt et al. 2018), and in seven counties in Pennsylvania;⁸ It is also considered as a policy alternative to existing transit offerings, especially in low-density regions.⁹ This experience suggests a possible pathway toward scalable microtransit using transit lines, on-demand deviations, and virtual bus stops.

Yet, deviated fixed-route microtransit has been subject to limited research. Quadrifoglio et al. (2007, 2008) optimized on-demand deviations with a single vehicle. Quadrifoglio et al. (2006) and Zhao and Dessouky (2008) quantified trade-offs between frequencies, deviations, and service levels. Galarza Montenegro et al. (2022) optimized on-demand operations in a related system in which transit vehicles can skip stops. Liu et al. (2021) formulated a mixed-integer linear optimization model to optimize on-demand deviations with autonomous vehicles. All these methods focus on the operational dynamics alone, and scale to small instances with 1-5 vehicles and 10-50 stops.

1.2. Contributions and Outline

This paper develops a scalable methodology to jointly optimize, for the first time to our knowledge, the design and operations of a deviated fixed-route microtransit system under demand uncertainty, including network design (which reference lines to operate), service scheduling (frequency and timetable), and on-demand operations (how to serve on-demand passenger requests).

Our first contribution is to formulate a *Microtransit Network Design (MiND)* model via two-stage stochastic optimization (Section 2). The first-stage problem selects *reference trips*, each encapsulating a *reference line* and a service frequency. The second-stage problem reflects on-demand routing deviations to serve passenger requests. The model features a multi-objective structure to minimize planning costs, maximize ridership and maximize a passenger level of service metric encapsulating walking, waiting, in-vehicle travel, and arrival delays. For simplicity and conciseness, we focus primarily on a MiND-VRP problem, corresponding to a vehicle routing setting in which all passengers

⁶ <https://ridewithvia.com/resources/multimedia/less-is-more>, <https://ridewithvia.com/case-study/jersey-city>

⁷ <https://www.nationalrtap.org/Toolkits/ADA-Toolkit/Service-Type-Requirements/Route-Deviation-Requirements>

⁸ <http://www.rideata.com/all-routes/fixed-routes-with-deviation/about-ata-fixed-routes-w-deviation>

⁹ <https://www.mass.gov/doc/flexible-transit-service-final-report/download>

have the same origin or the same destination, motivated by use cases such as airport shuttles. This setting captures the core elements of network design and on-demand operations. In EC.1.1, we extend the methodology and main results to a MiND-DAR problem, corresponding to a more complex dial-a-ride setting in which passengers request transportation from origin to destination; we also introduce pointers throughout the paper to the corresponding changes.

The MiND features an adaptive optimization structure with two challenging discrete optimization problems: network design and capacitated vehicle routing with time windows. To retain a tight recourse formulation, we propose a subpath-based representation of second-stage microtransit operations in a load-expanded network, in which each node encodes a checkpoint on the reference line and a vehicle load, and each arc characterizes on-demand operations between checkpoints. Load expansion accommodates vehicle capacities without big- M constraints, leading to a continuous recourse function approximation. We show that our subpath-based variables enable a more effective formulation than a segment-based benchmark with variables connecting consecutive stops (by integrating time windows in the definition of subpaths without involving a time-load-expanded network) and than a path-based benchmark with variables connecting the start to the end of each transit line (by drastically quelling the rate of exponential growth in the number of variables).

Our second contribution is a scalable double-decomposition algorithm combining Benders decomposition and subpath-based column generation to solve large-scale MiND instances (Section 3). The Benders decomposition scheme iterates between a first-stage network design problem and second-stage routing problems, exploiting the nested block-diagonal structure to decompose on-demand operations for each reference trip in each scenario. The column generation scheme adds subpath-based variables iteratively in the Benders subproblem. We develop exact and heuristic label-setting algorithms to generate subpaths of negative reduced cost while keeping track of vehicle load and level of service. As compared to typical combinations of Benders decomposition and (path-based) column generation, our modeling and algorithmic approach induces a double-decomposition structure: the column generation pricing problem adds subpaths between checkpoints, the Benders subproblem combines them to optimize the operating performance of each reference trip in each scenario, and the Benders master problem selects reference trips to optimize the overall network.

Our third contribution is to demonstrate the scalability of our methodology to large instances arising in practice (Section 4). We develop a real-world setup in Manhattan using data from the NYC Taxi & Limousine Commission (2021). Results show the benefits of our subpath-based formulation as compared to the segment- and path-based benchmarks and, most importantly, the combined benefits of Benders decomposition, subpath-based column generation, and our label-setting algorithm toward solving large-scale and otherwise-intractable instances. In particular, our MiND-VRP algorithm scales to instances of the size of the full Manhattan network with up to

100 candidate lines, hundreds of stations, thousands of passenger requests, and 5–20 demand scenarios; and our MiND-DAR algorithm scales to instances of the size of Midtown Manhattan with 10 candidate lines. Our methodology can handle much larger instances than previous approaches to microtransit operations, while also adding a first-stage design layer under demand uncertainty. Ultimately, our results show the practical benefits of our integrated stochastic optimization methodology, with a value of the stochastic solution of 5–7% against a deterministic benchmark.

Our final contribution is to derive evidence that deviated fixed-route microtransit can provide win-win outcomes toward efficient, equitable and sustainable mobility (Section 5). As compared to ride-sharing, microtransit consolidates demand into high-capacity vehicles along reference lines. As compared to fixed-route transit, it achieves higher demand coverage and comparable levels of service by leveraging on-demand routing flexibility. In turn, the optimized microtransit network has a broader catchment area than its fixed-route counterpart, thus enhancing accessibility in otherwise-unserved regions. Finally, demand consolidation and high coverage result in a significant decrease in distance traveled per passenger, which yields environmental benefits and can enable more affordable on-demand mobility options. Altogether, deviated fixed-route microtransit can contribute to efficient (high demand coverage, low operating costs per passenger, high service levels), equitable (broad geographic reach), and sustainable mobility (limited environmental footprint). Since Manhattan represents a high-density region with good transit alternatives, these results can be seen as conservative estimates of the impact of microtransit in other, lower-density areas with fewer transit alternatives. These results have inspired ongoing collaborations with transit operators toward the pilot deployment of deviated fixed-route microtransit, based on the methodology from this paper.

2. Microtransit Network Design (MiND) Model

The MiND optimizes the design and operations of deviated fixed-route microtransit via two-stage stochastic optimization. The first-stage design phase defines reference lines and service schedules (Section 2.1). The second-stage operational phase defines on-demand deviations in response to passenger requests, using a subpath-based representation in a load-expanded network (Section 2.2). The problem combines network design, frequency planning, and vehicle routing decisions. We formulate the MiND-VRP in Section 2.3 and compare the subpath-based formulation to segment- and path-based benchmarks in Section 2.4. The extension to the MiND-DAR is presented in EC.1.1.

2.1. First-stage Problem: Network Design and Frequency Planning

The first-stage problem defines *reference trips*, each characterized by a reference line and a departure time. Each reference line is defined as an ordered set of checkpoints, and each reference trip determines the scheduled time at each checkpoint. Vehicles are required to visit some checkpoints at the scheduled times, but will also be allowed to visit other locations in-between (Section 2.2).

Operations occur over a roadway network. Let \mathcal{N} denote the set of stations, including all candidate checkpoints and possible stopping locations. Stations may have physical infrastructure (e.g., benches), or may represent roadway intersections with sufficient curbside space. We represent demand as a set of *passenger requests* \mathcal{P} , which we assume scales linearly with the number of stations $|\mathcal{N}|$. In the MiND-VRP, each request $p \in \mathcal{P}$ is characterized by an origin $o(p) \in \mathcal{N}$ and a requested drop-off time t_p^{req} (see EC.1.1 for the MiND-DAR extension). To capture demand uncertainty, we define a set of scenarios \mathcal{S} , with D_{ps} passengers from request $p \in \mathcal{P}$ in scenario $s \in \mathcal{S}$.

Network design. We pre-process candidate reference lines in a set \mathcal{L} . Let h_ℓ denote the cost to operate one trip of line $\ell \in \mathcal{L}$. Let \mathcal{T}_ℓ store time periods when a vehicle can depart from the first checkpoint in line $\ell \in \mathcal{L}$. We introduce the following decision variables to define reference trips:

$$x_{\ell t} = \begin{cases} 1 & \text{reference trip } (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell \text{ is selected,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{I}_\ell \subseteq \mathcal{N}$ index the checkpoints in reference line ℓ , of cardinality $I_\ell = |\mathcal{I}_\ell|$. Let $\mathcal{I}_\ell^{(i)}$ refer to the i^{th} checkpoint in the line, for $i \in \{1, \dots, I_\ell\}$. In the MiND-VRP, all reference lines share the same final checkpoint $\mathcal{I}_\ell^{(I_\ell)}$. For a given reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, the vehicle is scheduled in checkpoint $i \in \mathcal{I}_\ell$ at time $T_{\ell t}(i)$. Let $tt_\ell = T_{\ell t}(\mathcal{I}_\ell^{(I_\ell)}) - T_{\ell t}(\mathcal{I}_\ell^1)$ be the total travel time on line ℓ .

We impose a fleet budget constraint by limiting the number of active trips at any time t :

$$\sum_{\ell \in \mathcal{L}} \sum_{t' \in \mathcal{T}_\ell : t' \leq t \leq t' + tt_\ell} x_{\ell t} \leq F, \quad \forall t \in \bigcup_{\ell \in \mathcal{L}} \mathcal{T}_\ell \quad (1)$$

Internal passenger assignments. Let $\mathcal{M}_p \subseteq \mathcal{L} \times \mathcal{T}_\ell$ denote the subset of reference trips that can serve request $p \in \mathcal{P}$. Adopting a dropoff-oriented view, we define \mathcal{M}_p as the set of trips that can drop off the passengers from request p within a tolerance α of their requested drop-off time:

$$\mathcal{M}_p = \left\{ (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell : \left| T_{\ell t}(\mathcal{I}_\ell^{(I_\ell)}) - t_p^{\text{req}} \right| \leq \alpha \right\}, \quad \forall p \in \mathcal{P}$$

We define assignment variables to identify a candidate reference trip for each passenger:

$$z_{\ell p s t} = \begin{cases} 1 & \text{if passenger type } p \in \mathcal{P} \text{ is internally assigned to trip } (\ell, t) \in \mathcal{M}_p \text{ in scenario } s \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

We impose packing constraints so that each passenger is assigned to at most one reference trip:

$$\sum_{(\ell, t) \in \mathcal{M}_p} z_{\ell p s t} \leq 1, \quad \forall p \in \mathcal{P}, \forall s \in \mathcal{S} \quad (2)$$

Note that the assignment variables do not define decisions that are executed in practice, since passenger service is optimized at the operational level. First-stage passenger assignments merely link first-stage and second-stage decisions to strengthen the second-stage formulation. The $z_{\ell p s t}$ variables are scenario-dependent; yet, we still refer to them as first-stage variables because they are associated with network design—they will be treated as first-stage variables in our algorithm.

Vehicle load. We assume that vehicles are homogeneous within each reference line $\ell \in \mathcal{L}$, with capacity C_ℓ . Since the $z_{\ell p s t}$ variables do not necessarily result in actual passenger assignments, we impose a *target load factor* $\kappa \in (0, 1)$ to induce high vehicle utilization. We also allow first-stage assignments to exceed vehicle capacities by a factor κ to create operating flexibility, but the second-stage passenger service decisions will strictly comply with vehicle capacities.

$$\sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} D_{ps} z_{\ell p s t} \geq (1 - \kappa) C_\ell x_{\ell t} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall s \in \mathcal{S} \quad (3)$$

$$\sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} D_{ps} z_{\ell p s t} \leq (1 + \kappa) C_\ell x_{\ell t} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall s \in \mathcal{S} \quad (4)$$

2.2. Second-stage Problem: On-demand Deviations

To adhere to the reference schedule, second-stage deviations must stay within a distance Δ of the reference line and must respect scheduled arrival times at the checkpoints. The reference schedule includes buffers between checkpoints to allow for deviations. Moreover, vehicles may skip up to $K \in \{0, 1\}$ checkpoints in a row: $K = 0$ induces closer adherence to the reference trip, whereas $K = 1$ provides more flexibility. We denote by Γ_ℓ the checkpoint pairs separated by up to K checkpoints.

The second-stage problem involves capacitated vehicle routing with time windows for each reference trip and in each scenario. To avoid a formulation with a loose discrete recourse function, we formulate the second-stage problem in a load-expanded network with new subpath-based variables to characterize on-demand deviations, leveraging the structure of microtransit operations.

Subpaths. For reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and demand scenario $s \in \mathcal{S}$, a subpath $r \in \mathcal{R}_{\ell s t}$ is identified by its starting checkpoint $u_r \in \mathcal{I}_\ell$, its ending checkpoint $v_r \in \mathcal{I}_\ell$, and the passenger requests $\mathcal{P}_r \subseteq \mathcal{P}$ served in between. The set $\mathcal{R}_{\ell s t}$ includes all subpaths such that the distance to the reference line never exceeds Δ ; the load satisfies $\sum_{p \in \mathcal{P}_r} D_{ps} \leq C_\ell$; the travel time does not exceed $T_{\ell t}(v_r) - T_{\ell t}(u_r)$; and up to K checkpoints are skipped. The second-stage problem selects a sequence of subpaths that (i) starts at the origin of the reference line and ends at its destination while maintaining flow balance; and (ii) serves up to C_ℓ passengers.

Load-expanded subpath network. We represent routing operations in a load-expanded subpath network, illustrated in Figure 1. Each node tracks the checkpoint and the vehicle load, and each arc encapsulates a subpath between checkpoints. In this network, flow balance constraints capture physical flow balance and vehicle capacity constraints. As compared to a physical network, our load-expanded network representation involves more variables, but enables a tighter second-stage formulation by capturing vehicle capacities without involving big-M constraints.

Let $\mathcal{C}_\ell = \{0, 1, \dots, C_\ell\}$ store valid vehicle loads. For each reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$, we denote the load-expanded network by $(\mathcal{V}_{\ell s t}, \mathcal{A}_{\ell s t})$. A dummy sink node $v_{\ell s t}$ represents

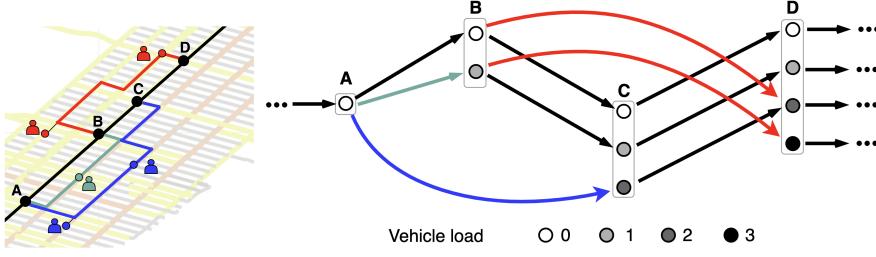


Figure 1 *Left: Physical network with three candidate deviations. Right: Load-expanded subpath network.*

the end of a trip. Each other node $n \in \mathcal{V}_{\ell st}$ corresponds to a tuple (k_n, c_n) consisting of checkpoint $k_n \in \mathcal{I}_\ell$ and load $c_n \in \mathcal{C}_\ell$. Each arc $a \in \mathcal{A}_{\ell st}$ connects nodes $\text{start}(a) \in \mathcal{V}_{\ell st}$ and $\text{end}(a) \in \mathcal{V}_{\ell st}$. We partition the arc set $\mathcal{A}_{\ell st} = \bigcup_{r \in \mathcal{R}_{\ell st}} \mathcal{A}_r \cup \mathcal{A}_{\ell st}^v$ into traveling arcs (\mathcal{A}_r) and terminating arcs ($\mathcal{A}_{\ell st}^v$). Each traveling arc $a \in \mathcal{A}_r$ captures the endpoints $u_r, v_r \in \mathcal{I}_\ell$ and the incremental load $\sum_{p \in \mathcal{P}_r} D_{ps}$ of subpath $r \in \mathcal{R}_{\ell st}$; vice versa, $r(a) \in \mathcal{R}_{\ell st}$ denotes the subpath corresponding to arc $a \in \bigcup_{r \in \mathcal{R}_{\ell st}} \mathcal{A}_r$. Each terminating arc $a \in \mathcal{A}_{\ell st}^v$ connects the line's destination to the dummy sink node. Specifically:

$$\mathcal{A}_r = \left\{ (n, m) \in \mathcal{V}_{\ell st} \times \mathcal{V}_{\ell st} : k_n = u_r, k_m = v_r, c_m - c_n = \sum_{p \in \mathcal{P}_r} D_{ps} \right\}, \quad \forall r \in \mathcal{R}_{\ell st}, \quad (5)$$

$$\mathcal{A}_{\ell st}^v = \{(n, m) \in \mathcal{V}_{\ell st} \times \mathcal{V}_{\ell st} : k_n = \mathcal{I}_\ell^{(I_\ell)}, m = v_{\ell st}\}. \quad (6)$$

Our second-stage decisions select subpaths in the load-expanded networks via the following variables. These define on-demand deviations and pickups for each reference trip and each scenario.

$$y_a = \begin{cases} 1 & \text{if arc } a \in \mathcal{A}_{\ell st} \text{ is selected, for } (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

Passenger service. Any passenger $p \in \mathcal{P}$ assigned to reference trip $(\ell, t) \in \mathcal{M}_p$ in the first stage would start walking at time $t_{\ell pt}^0$ to the closest checkpoint—as indicated on a mapping application, for instance. To guarantee a convenient service and ensure consistency between planned and realized trips, we restrict on-demand deviations so that passengers do not start before their planned departure times $t_{\ell pt}^0$, do not wait more than a limit Ψ , and do not walk more than a limit Ω . Let $\psi_{o,d}$ and $\omega_{o,d}$ be the walking time and the walking distance from o to d , respectively. A pickup in location $i \in \mathcal{N}$ at time t is only acceptable if $t_{\ell pt}^0 + \psi_{o(p),i} \leq t \leq t_{\ell pt}^0 + \psi_{o(p),i} + \Psi$ and if $\omega_{o(p),i} \leq \Omega$.

Beyond these restrictions, we propose a multi-objective second-stage formulation to maximize demand coverage and passenger level of service. Coverage is formalized via a large reward M incurred for every successful pick-up. Level of service is formalized as a four-dimensional objective reflecting the *generalized cost of travel* (Ceder and Wilson 1986, Desaulniers and Hickman 2007):

1. τ_{rp}^{walk} : walking time from passenger p 's origin to the pickup location via subpath $r \in \mathcal{R}_{\ell st}$;
2. τ_{rp}^{wait} : waiting time of passenger p prior to pickup via subpath $r \in \mathcal{R}_{\ell st}$;
3. $\frac{\tau_{rp}^{travel}}{\tau_p^{dir}}$: relative detour, defined as the in-vehicle travel time of passenger p via subpath $r \in \mathcal{R}_{\ell st}$ normalized with respect to the direct trip time (e.g., a taxi trip); and

4. $\frac{\tau_{\ell tp}^{late}}{\tau_p^{dir}}, \frac{\tau_{\ell tp}^{early}}{\tau_p^{dir}}$: relative delay and earliness of passenger p at the destination via trip $(\ell, t) \in \mathcal{M}_p$.

The reference line guarantees that the vehicle reaches the destination at a specified time, so this cost can be expressed at the trip level. We penalize delay twice as much as earliness.

We define non-negative hyperparameters λ, μ, σ , and δ to weigh the level of service cost components.

The arc costs in the load-expanded network are defined as follows for all $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$.

$$g_a = \begin{cases} \sum_{p \in \mathcal{P}_{r(a)}} D_{ps} \left(\lambda \tau_{r(a)p}^{walk} + \mu \tau_{r(a)p}^{wait} + \sigma \frac{\tau_{r(a)p}^{travel}}{\tau_p^{dir}} + \delta \frac{\tau_{\ell tp}^{late}}{\tau_p^{dir}} + \frac{\delta}{2} \frac{\tau_{\ell tp}^{early}}{\tau_p^{dir}} - M \right) & \forall a \in \bigcup_{r \in \mathcal{R}_{\ell st}} \mathcal{A}_r, \\ 0 & \forall a \in \mathcal{A}_{\ell st}^v. \end{cases} \quad (7)$$

2.3. Two-stage Stochastic Optimization Formulation (MiND-VRP)

The MiND-VRP minimizes planning costs, maximizes demand coverage, and maximizes level of service (Equation (8)). The constraints apply the fleet size, the target load factor, and packing constraints (Equations (1)–(4)); enforce flow balance over load-expanded networks (Constraint (9)); and link first-stage assignments with second-stage operations (Constraint (10)). Table EC.2 in EC.2.1 summarizes notation. The MiND-DAR is formulated similarly, with an extra consistency constraint to ensure that each served passenger is picked up *and* dropped off (EC.1.1).

$$\min \sum_{\ell \in \mathcal{L}} \sum_{t \in \mathcal{T}_\ell} \left(h_\ell x_{\ell t} + \sum_{s \in \mathcal{S}} \pi_s \sum_{a \in \mathcal{A}_{\ell st}} g_a y_a \right) \quad (8)$$

s.t. First-stage constraints: Equations (1)–(4)

$$\sum_{m:(n,m) \in \mathcal{A}_{\ell st}} y_{(n,m)} - \sum_{m:(m,n) \in \mathcal{A}_{\ell st}} y_{(m,n)} = \begin{cases} x_{\ell t} & \text{if } n = u_{\ell st} \\ -x_{\ell t} & \text{if } n = v_{\ell st} \\ 0 & \text{otherwise} \end{cases} \quad \forall \ell \in \mathcal{L}, t \in \mathcal{T}_\ell, s \in \mathcal{S}, n \in \mathcal{V}_{\ell st} \quad (9)$$

$$\sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}} y_a \leq z_{\ell p st} \quad \forall s \in \mathcal{S}, p \in \mathcal{P}, (\ell, t) \in \mathcal{M}_p \quad (10)$$

$$\mathbf{x}, \mathbf{y}, \mathbf{z} \text{ binary} \quad (11)$$

2.4. Comparison to Segment-based and Path-based Benchmarks

We compare our subpath-based formulation to the following benchmarks, detailed in EC.2:

- A segment-based model optimizes over arc-based variables connecting stations in a time-load-expanded network to enforce vehicle capacities and time windows. To appropriately reflect vehicle and passenger operations, the second-stage time discretization needs to be much more granular than the one governing first stage frequency planning (sets \mathcal{T}_ℓ). Moreover, the model is further complicated by two multi-commodity flow structures with additional linking constraints: flows from checkpoint to checkpoint (so the vehicle does not skip more than K checkpoints in a row) and flows from station to station (to maintain continuity in time and space).
- A path-based model optimizes over path-based variables, each characterizing a sequence of subpaths that starts at the line's origin, ends at its destination, and serves at most C_ℓ passengers. The model minimizes cost subject to set partitioning and passenger service constraints.

Proposition 1 shows that the three formulations are equivalent, as long as time discretization can be sufficiently granular in the segment-based benchmark (we formalize this condition in EC.2.4). The segment-based benchmark induces a weaker relaxation due to the double flow structure with linking constraints. The subpath-based formulation achieves an equally strong relaxation as the path-based benchmark in the MiND-VRP thanks to the flow balance structure on the load-expanded network. Most importantly, Proposition 2 shows the size benefits of the subpath-based formulation. The subpath-based model captures timing requirements without time discretization, whereas the segment-based benchmark scales linearly with T_S (which needs to be very large to avoid discretization errors). The subpath-based model also drastically quells the rate of exponential growth in the number of variables as compared to the path-based benchmark; specifically, the number of subpaths scales exponentially with the number of stations between checkpoints whereas the number of paths scales exponentially with the total number of stations along the reference line.

PROPOSITION 1. *The path-based and subpath-based formulations are equivalent and define identical linear relaxations. If all subpath travel times are strictly less than the elapsed time between the scheduled arrival times at the checkpoints, there exists a time discretization such that the segment-based formulation is also equivalent but its linear relaxation is at most as strong.*

PROPOSITION 2. *Consider the second-stage problem for reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ in scenario $s \in \mathcal{S}$. Let Ξ be the maximum number of stops between any pair of checkpoints in Γ_ℓ . The segment-based formulation has $\mathcal{O}(T_S \cdot C_\ell^2 \cdot I_\ell \cdot \Xi^2)$ variables and $\mathcal{O}(|\mathcal{P}| + T_S \cdot C_\ell \cdot |\mathcal{N}| + T_S \cdot C_\ell^2 \cdot I_\ell \cdot \Xi^2)$ constraints. The subpath-based formulation has $\mathcal{O}(I_\ell \cdot C_\ell \cdot 2^\Xi)$ variables and $\mathcal{O}(|\mathcal{P}| + C_\ell \cdot I_\ell)$ constraints. The path-based formulation had $\mathcal{O}(2^{\Xi \cdot I_\ell})$ variables and $\mathcal{O}(|\mathcal{P}|)$ constraints.*

3. Double-Decomposition Algorithm

The MiND-VRP exhibits a two-stage optimization structure with a tight recourse function and exponentially many second-stage variables. We propose a solution algorithm based on Benders decomposition to exploit the nested block-angular structure (Section 3.1), and subpath-based column generation in the Benders subproblem (Section 3.2). The algorithm relies on a tailored label-setting algorithm to generate subpaths in a time-expanded network (Section 3.3). We formalize the algorithm and establish its exactness in Section 3.4. Again, we extend it to the MiND-DAR in EC.1.2.

Combinations of column generation and Benders decomposition fall into three categories: (i) simultaneous column-and-row generation (Muter et al. 2013); (ii) path-based column generation in the Benders master problem (Zeigham and Soumis 2019); and (iii) path-based column generation in the Benders subproblem (Karsten et al. 2018). Our algorithm relates to this third category by adding variables iteratively to the Benders subproblem. But rather than adding path-based

variables, our modeling and algorithmic approach induces an extra decomposition layer at the subpath level, giving rise to a novel double-decomposition structure: the column generation pricing problem adds subpaths between checkpoints, the Benders subproblem combines them into a full microtransit trip, and the Benders master problem selects network-wide reference trips accordingly.

We refer to the model's partial relaxation with first-stage binary variables and second-stage continuous variables as MiND-VRP'. This relaxation is close to the full problem due to the reliance on flow balance constraints in load-expanded networks (Equation (9)). Still, they can induce fractional solutions due to passenger service constraints (Equation (10)). Upon convergence, we solve a final second-stage model to obtain feasible integer solution and an optimality gap.

3.1. Benders Reformulation

We propose a multi-cut Benders decomposition of MiND-VRP' into a Benders master problem (BMP) and Benders subproblems (BSP). This approach exploits the nested block-angular structure of the formulation to decompose the second-stage problem in each scenario and for each reference trip—due to the fact that on-demand deviations are independent across reference trips. Note that MiND-VRP' has relatively complete recourse, because a feasible second-stage solution can always be constructed by following the reference trip. By the Minkowski-Weyl theorem, the dual second-stage polyhedron can therefore be characterized as a convex combination of its extreme points.

Let $\theta_{\ell st}$ denote the second-stage objective, for each reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$. It is given by the following Benders subproblem, for each first-stage decision (\mathbf{x}, \mathbf{z}) :

$$\text{BSP}(\mathbf{x}, \mathbf{z}) \quad \theta_{\ell st} = \min_{\mathbf{y} \geq \mathbf{0}} \sum_{a \in \mathcal{A}_{\ell st}} g_a y_a \quad \text{s.t. Equations (9)-(10)} \quad (12)$$

Let φ_i and γ_p respectively denote the dual variables corresponding to Equations (9) and (10), respectively. The dual Benders subproblem is then formulated as follows:

$$\max \quad x_{\ell t} \cdot (\varphi_{\bar{u}_{\ell st}} - \varphi_{\bar{v}_{\ell st}}) - \sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} z_{\ell p st} \cdot \gamma_p \quad (13)$$

$$\text{s.t.} \quad \varphi_n - \varphi_m - \sum_{p \in \mathcal{P}_a} \gamma_p \leq g_a \quad \forall a = (n, m) \in \mathcal{A}_{\ell st} \quad (14)$$

$$\varphi_i \in \mathbb{R} \quad \forall i \in \mathcal{V}_{\ell st} \quad (15)$$

$$\gamma_p \geq 0 \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (16)$$

Let $\Lambda_{\ell st}$ store the extreme points of the dual second-stage polyhedron, each corresponding to a second-stage solution $(\boldsymbol{\varphi}, \boldsymbol{\gamma})$ for reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$. Let $\boldsymbol{\Lambda} = (\Lambda_{\ell st})_{(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}}$ store all extreme points. The MiND-VRP' reformulation optimizes network design and passenger assignments subject to a piece-wise linear recourse approximation:

$$\text{BMP}(\boldsymbol{\Lambda}) \quad \min \quad \sum_{\ell \in \mathcal{L}} \sum_{t \in \mathcal{T}_\ell} \left(h_\ell x_{\ell t} + \sum_{s \in \mathcal{S}} \pi_s \theta_{\ell st} \right) \quad (17)$$

$$\text{s.t. Equations (1)–(4)} \quad (18)$$

$$\theta_{\ell st} \geq x_{\ell t} \cdot (\varphi_{\bar{u}_{\ell st}} - \varphi_{\bar{v}_{\ell st}}) - \sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} z_{\ell p st} \cdot \gamma_p, \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}, (\boldsymbol{\varphi}, \boldsymbol{\gamma}) \in \Lambda_{\ell st} \quad (19)$$

$$\mathbf{x}, \mathbf{z} \text{ binary} \quad (20)$$

To circumvent the exponential number of extreme points in the dual second-stage polyhedron, the Benders master problem solves a relaxation $\text{BMP}(\overline{\boldsymbol{\Lambda}})$ containing a subset of constraints $\overline{\boldsymbol{\Lambda}} \subseteq \boldsymbol{\Lambda}$. By design, the BMP yields a lower bound of MiND-VRP' and the combination of the BMP and BSP yield an upper bound. If the gap lies within a given tolerance, the algorithm stops; otherwise, we retrieve the optimal dual solution $(\bar{\boldsymbol{\varphi}}, \bar{\boldsymbol{\gamma}})$ of $\text{BSP}(\mathbf{x}, \mathbf{z})$, and add an optimality cut to the BMP:

$$\theta_{\ell st} \geq x_{\ell t} \cdot (\bar{\varphi}_{\bar{u}_{\ell st}} - \bar{\varphi}_{\bar{v}_{\ell st}}) - \sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} z_{\ell p st} \cdot \bar{\gamma}_p \quad (21)$$

Benders decomposition iterates between the BMP and BSP until convergence to an optimal solution of MiND-VRP'. To strengthen the recourse approximation, we developed a two-phase implementation by first applying Benders decomposition to the full relaxation with continuous first- and second-stage decisions. However, Benders decomposition remains hindered by the large number of subpath-based variables in the subproblem, motivating our column generation procedure.

3.2. Subpath-based Column Generation for Benders Subproblem

Recall that the arc set $\mathcal{A}_{\ell st}$ grows exponentially with the number of candidate stops between checkpoints. The number of second-stage variables is especially large when microtransit vehicles can skip checkpoints ($K = 1$). Our column generation procedure decomposes the BSP into a restricted Benders subproblem (RBSP) and a pricing problem (PP) to generate subpaths iteratively.

Restricted Benders subproblem. The RBSP simply solves the Benders subproblem with a subset of subpath-based arcs by $\mathcal{A}'_{\ell st} \subseteq \mathcal{A}_{\ell st}$. It is formulated as follows:

$$\text{RBSP}(\mathcal{A}'_{\ell st}, \mathbf{x}, \mathbf{z}) \quad \min_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{a \in \mathcal{A}'_{\ell st}} g_a y_a \quad (22)$$

$$\text{s.t.} \quad \sum_{m:(n,m) \in \mathcal{A}'_{\ell st}} y_{(n,m)} - \sum_{m:(m,n) \in \mathcal{A}'_{\ell st}} y_{(m,n)} = \begin{cases} x_{\ell t} & \text{if } n = \bar{u}_{\ell st} \\ -x_{\ell t} & \text{if } n = \bar{v}_{\ell st} \\ 0 & \text{otherwise} \end{cases} \quad \forall n \in \mathcal{V}_{\ell st} \quad (23)$$

$$\sum_{a \in \mathcal{A}'_{\ell st} : p \in \mathcal{P}_{r(a)}} y_a \leq z_{\ell p st} \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (24)$$

Subpath characterization. Consider a subpath $r \in \mathcal{R}_{\ell st}$ starting from checkpoint $u_r \in \mathcal{I}_\ell$ at time $T_{\ell t}(u)$ and ending in checkpoint $v_r \in \mathcal{I}_\ell$ at time $T_{\ell t}(v)$ (we denote u_r and v_r by u and v for simplicity). Let $\mathcal{N}_{uv} \subseteq \mathcal{N}$ denote the set of stations between checkpoints u and v that lie within

the allowable deviation Δ from the reference line. Let $\mathcal{E}_{uv} \subset \mathcal{N}_{uv} \times \mathcal{N}_{uv}$ store the directed roadways connecting them. Each subpath will be characterized by a sequence of segments in \mathcal{E}_{uv} and a set of passenger pickups satisfying capacity, time window, deviation, and stop-skipping requirements.

To capture timing requirements, we characterize subpaths in a time-expanded network $(\mathcal{U}_{\ell st}^{uv}, \mathcal{H}_{\ell st}^{uv})$. Let $\mathcal{T}_{\ell t}^{uv}$ be a set of discretized time intervals between the departure time $T_{\ell t}(u)$ from checkpoint u and the arrival time $T_{\ell t}(v)$ at checkpoint v . As in the segment-based formulation (Section 2.4), the sets $\mathcal{T}_{\ell t}^{uv}$ need to be more granular than the first-stage sets \mathcal{T}_ℓ (30 seconds vs. 15 minutes, in our experiments). Yet, the sets $\mathcal{T}_{\ell t}^{uv}$ remain manageable due to the restriction between times $T_{\ell t}(u)$ and $T_{\ell t}(v)$. Each node $m \in \mathcal{U}_{\ell st}^{uv}$ is represented by a tuple $(k_m, t_m) \in \mathcal{N}_{uv} \times \mathcal{T}_{\ell t}^{uv}$; $(u, T_{\ell t}(u)) \in \mathcal{U}_{\ell st}^{uv}$ is the source node and $(v, T_{\ell t}(v)) \in \mathcal{U}_{\ell st}^{uv}$ is the sink node. The arc set $\mathcal{H}_{\ell st}^{uv}$ comprises (i) traveling arcs connecting any node pair $(i, t) \rightarrow (j, t + tt_{ij})$ where $(i, j) \in \mathcal{E}_{uv}$ defines a road segment, tt_{ij} defines the corresponding travel time, with $t \in \mathcal{T}_{\ell t}^{uv}$ and $t + tt_{ij} \in \mathcal{T}_{\ell t}^{uv}$; and (ii) idling arcs connecting any node pair $(i, t) \rightarrow (i, t + 1)$ where $i \in \mathcal{N}_{uv}$ defines a station and $t, t + 1 \in \mathcal{T}_{\ell t}^{uv}$.

Each node $m \in \mathcal{U}_{\ell st}^{uv}$ also defines passengers' waiting, walking and travel times, as well as arrival delays and earliness, which we store in parameters τ_{mp}^{walk} , τ_{mp}^{wait} , $\tau_{mp}^{\text{travel}}$, τ_{mp}^{late} , and τ_{mp}^{early} . We denote by $\mathcal{P}_m \subset \mathcal{P}$ the set of passengers that can be picked up at node $m \in \mathcal{U}_{\ell st}^{uv}$ given the walking and waiting restrictions defined in Section 2.2. Table EC.2 in EC.2.1 summarizes notation.

Pricing problem. Consider two nodes in the load-expanded network $(u, c_1), (v, c_2) \in \mathcal{V}_{\ell st}$. The pricing problem seeks a subpath that starts in checkpoint $u \in \mathcal{N}$ at time $T_{\ell t}(u)$ with vehicle load c_1 , and ends in checkpoint $v \in \mathcal{N}$ at time $T_{\ell t}(v)$ with load $c_2 \geq c_1$. We define the following variables:

$$f_{mq} = \begin{cases} 1 & \text{if arc } (m, q) \in \mathcal{H}_{\ell st}^{uv} \text{ is traversed in the time-expanded road segment network,} \\ 0 & \text{otherwise.} \end{cases}$$

$$w_{mp} = \begin{cases} 1 & \text{if passenger } p \in \mathcal{P}_m \text{ is picked up in node } m \in \mathcal{U}_{\ell st}^{uv}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\xi_m = \text{vehicle load in node } m \in \mathcal{U}_{\ell st}^{uv}$$

Let \hat{g}_a denote the reduced cost of arc-based variable $a = ((u, c_1), (v, c_2)) \in \mathcal{A}_{\ell st}$. From Equation (14), the reduced cost can be separated into a routing component and a load component. The routing component comprises (i) the level-of-service penalty for passengers receiving a service, and (ii) the value of serving a passenger, captured by the actual value M adjusted with the dual price γ_p . The load component reflects the dual cost $\varphi_{(v, c_2)} - \varphi_{(u, c_1)}$ of increasing the vehicle load.

$$\hat{g}_a = \underbrace{\sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} d_{mp} w_{mp}}_{\text{routing component}} + \underbrace{\varphi_{(v, c_2)} - \varphi_{(u, c_1)}}_{\text{load component}} \quad (25)$$

$$\text{with } d_{mp} = D_{ps} \left(\frac{\delta \tau_{mp}^{\text{late}} + \frac{\delta}{2} \tau_{mp}^{\text{early}} + \sigma \tau_{mp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \lambda \tau_{mp}^{\text{walk}} + \mu \tau_{mp}^{\text{wait}} - M \right) + \gamma_p.$$

The pricing problem seeks a subpath with minimum reduced cost (Equation (26)). Constraints (27)–(29) define the load at each node, starting from load c_1 and ending with load c_2 . Constraints (30) and (31) ensures that passenger pickups occur only in visited nodes, and at most once. Constraints (32) apply flow balance in the time-expanded network.

$$\text{PP}_{\ell st}^{u,v,c_1,c_2} \quad \min \quad \sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} d_{mp} w_{mp} + \varphi_{(v,c_2)} - \varphi_{(u,c_1)} \quad (26)$$

$$\text{s.t.} \quad \xi_{(u,T_{\ell t}(u))} = c_{(u,c_1)}, \quad \xi_{(v,T_{\ell t}(v))} = c_{(v,c_2)} \quad (27)$$

$$\xi_q - \xi_m \leq \sum_{p \in \mathcal{P}_m} D_{mp} w_{mp} + C_{\ell}(1 - f_{mq}), \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{uv} \quad (28)$$

$$\xi_q - \xi_m \geq \sum_{p \in \mathcal{P}_m} D_{mp} w_{mp} - C_{\ell}(1 - f_{mq}), \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{uv} \quad (29)$$

$$w_{mp} \leq \sum_{q:(m,q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, \quad \forall p \in \mathcal{P}_m \quad (30)$$

$$\sum_{m \in \mathcal{U}_{\ell st}^{uv}: p \in \mathcal{P}_m} w_{mp} \leq 1 \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (31)$$

$$\sum_{q:(m,q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} - \sum_{q:(q,m) \in \mathcal{H}_{\ell st}^{uv}} f_{qm} = \begin{cases} 1 & \text{if } m = (u, T_{\ell t}(u)), \\ -1 & \text{if } m = (v, T_{\ell t}(v)), \\ 0 & \text{otherwise.} \end{cases} \quad \forall m \in \mathcal{U}_{\ell st}^{uv} \quad (32)$$

$$\boldsymbol{f}, \boldsymbol{w} \text{ binary, } \boldsymbol{\xi} \text{ non-negative integer} \quad (33)$$

Column generation iterates between the RBSP and the PP. If all reduced costs are non-negative, the column generation algorithm terminates and the Benders decomposition algorithm proceeds. Otherwise, any subpath-based arc $a \in \mathcal{A}_{\ell st}$ with negative reduced cost gets added to the load-expanded network, by augmenting $\mathcal{A}'_{\ell st} \leftarrow \mathcal{A}'_{\ell st} \cup \{a\}$ and defining the level of service g_a as:

$$g_a = \sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} D_{ps} \left(\frac{\delta \tau_{mp}^{\text{late}} + \frac{\delta}{2} \tau_{mp}^{\text{early}} + \sigma \tau_{mp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \lambda \tau_{mp}^{\text{walk}} + \mu \tau_{mp}^{\text{wait}} - M \right) w_{mp} \quad (34)$$

Note that the PP searches over all subpaths, including those corresponding to non-selected reference lines (Equation (32)) and non-assigned passengers (Equation (31)). Such subpaths will necessarily be primal infeasible in the RBSP with the incumbent BMP solution (Equations (23) and (24)). However, the corresponding RBSP constraints take the form “ $0 \leq 0$ ” and cannot be assumed to have zero duals. The more general PP formulation is essential to certify that all subpath-based variables satisfy Equation (14) upon convergence of column generation, hence to ensure the validity of Benders decomposition. In other words, the dual BSP polyhedron is independent on the incumbent first-stage variables \boldsymbol{x} and \boldsymbol{z} , and so is the PP. This observation is formalized in Remark 1, which we prove when establishing the exactness of the algorithm in Proposition 5.

REMARK 1. The right-hand side of Equation (32) (resp, Equation (31)) must be 1 rather than $x_{\ell t}$ (resp., $z_{\ell pst}$) to certify optimality of the RBSP solution and guarantee the algorithm’s exactness.

The pricing problem is defined for each reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, each scenario $s \in \mathcal{S}$, and for node pair $((u, c_1), (v, c_2)) \in \mathcal{V}_{\ell st} \times \mathcal{V}_{\ell st}$ in the load-expanded network. Rather than solving a pricing problem for each pair of checkpoints and each load pair, we can reduce the number of pricing problems by exploiting the decomposition of the reduced cost into a routing component and a load component. Specifically, we first maximize the load component for each load differential $\varepsilon \in \mathcal{C}_\ell$:

$$\Delta\varphi_{\ell st}^{u,v,\varepsilon} = \max \left\{ \varphi_{start(a)} - \varphi_{end(a)} : a \in \mathcal{A}_{\ell st}, k_{start(a)} = u, k_{end(a)} = v, c_{end(a)} - c_{start(a)} = \varepsilon \right\}$$

We then seek a subpath that serves ε passengers and minimizes the routing component of the reduced cost. It is formulated as follows (without the big-M constraints (28)–(29)):

$$Z_{\ell st}^{u,v,\varepsilon} = \min \sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} d_{mp} w_{mp}; \text{ s.t. } \sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp} = \varepsilon; \text{ Equations (30)–(33)}$$

Proposition 3 shows that we can solve one pricing problem for each *load differential* and every pair of checkpoints. This result reduces the number of pricing problem by a factor $\mathcal{O}(\max_{\ell \in \mathcal{L}} C_\ell)$, while retaining the finite convergence and exactness of the column generation scheme.

PROPOSITION 3. $Z_{\ell st}^{u,v,\varepsilon} - \Delta\varphi_{\ell st}^{u,v,\varepsilon}$ is the minimum reduced cost across all arc-based variables between checkpoints u and v with load differential ε , for all $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, $s \in \mathcal{S}$.

3.3. Label-setting Algorithm

Exact algorithm. The pricing problem exhibits a resource-constrained shortest path structure. We design a label-setting algorithm by exploiting the directed and acyclic structure of $(\mathcal{U}_{\ell st}^{uv}, \mathcal{H}_{\ell st}^{uv})$, with a two-dimensional state space to store the set of passenger pickups and the reduced cost.

State definition. Let $(m^\sigma, \mathbb{P}^\sigma)$ denote a state, where m^σ tracks the “current” node, and \mathbb{P}^σ tracks the set of served passengers $p \in \mathcal{P}$ each with pickup node ρ_p . We track the reduced cost $G(m^\sigma, \mathbb{P}^\sigma)$.

Initial state: $(m^0 = m, \mathbb{P}^0 = \emptyset)$, where m is such that $k_m = u$ and $t_m = T_{\ell t}(u)$; $G(m^0, \mathbb{P}^0) = 0$.

State transitions. For each arc $(m, q) \in \mathcal{H}_{\ell st}^{uv}$ and each passenger combination $\mathbb{P}_m \subseteq \mathcal{P}_m$, the state is updated to $(q, \mathbb{P}^\sigma \cup \mathbb{P}_m)$. For each new passenger $p \in \mathbb{P}_m \setminus \{\mathbb{P}^\sigma\}$, the pickup point is set to $\rho_p = m$. For existing passengers $p \in \mathbb{P}_m \cap \mathbb{P}^\sigma$, we update the pickup node to be $\rho_p = m$ if $d_{mp} < d_{\rho_p, p}$. This transition is admissible if the vehicle has enough capacity, i.e., if $\sum_{p \in \mathbb{P}^\sigma \cup \mathbb{P}_m} D_{ps} \leq C_\ell$.

Reward function. $G(m^\sigma, \mathbb{P}^\sigma) = \sum_{p \in \mathbb{P}^\sigma} d_{\rho_p, p}$ tracks the reduced cost of a subpath up to state σ .

Dominance rule. σ^1 dominates σ^2 if $m^{\sigma^1} = m^{\sigma^2}$, $\mathbb{P}^{\sigma^1} = \mathbb{P}^{\sigma^2}$, and $G(m^{\sigma^1}, \mathbb{P}^{\sigma^1}) \leq G(m^{\sigma^2}, \mathbb{P}^{\sigma^2})$.

Upon termination, we extract all non-dominated states l such that $m^\sigma = m : k_m = v$ and $t_m = T_{\ell t}(v)$. We then add to the RBSP all arcs $a \in \mathcal{A}_{\ell st} \setminus \{\mathcal{A}'_{\ell st}\}$ such that $k_{start(a)} = u$, $k_{end(a)} = v$, $c_{end(a)} - c_{start(a)} = \sum_{p \in \mathbb{P}^\sigma} D_{ps}$, with reduced cost $\hat{g}_a = G(m^\sigma, \mathbb{P}^\sigma) - \varphi_{start(a)} + \varphi_{end(a)}$, as long as $\hat{g}_a < 0$.

By design, the dominance rule yields the subpath of minimum reduced cost for each passenger combination—hence, for each load differential. Thus, we apply the label-setting algorithm for each

pair of checkpoints $u, v \in \mathcal{I}_\ell$, but do not duplicate it for each load differential. The number of checkpoint pairs grows linearly with $|\mathcal{I}_\ell|$ because subpaths can skip up to $K \in \{0, 1\}$ checkpoint. Combined with Proposition 3, we obtain the following reduction on the number of pricing problems:

PROPOSITION 4. *The label-setting algorithm generates $\mathcal{O}(2^\Xi |\mathcal{V}_{\ell st}|)$ variables at a time by only solving $\mathcal{O}(|\mathcal{I}_\ell|)$ pricing problems, for each reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$.*

Heuristic acceleration. The label-setting algorithm can lead to a weak dominance rule with extensive enumeration of subpaths serving slightly different passenger combinations. In fact, subpaths are relatively short, so the pricing problem rarely rejects a passenger with a negative reduced cost ($d_{mp} < 0$) to free up capacity for a subsequent passenger. Moreover, it can be undesirable in practice to reject a passenger at a stop visited by the vehicle. We therefore propose a heuristic acceleration such that, in each node $m \in \mathcal{U}_{\ell st}^{uv}$, all candidate passengers $p \in \mathcal{P}_m$ with negative reduced cost contribution ($d_{mp} < 0$) are served, as long as the vehicle does not operate at capacity. This heuristic yields an upper-bounding approximation of the pricing problem, i.e., it generates solutions with a negative reduced cost but can potentially miss other subpaths with negative reduced cost, especially with high demand concentration. In that case, we can switch back to the full label-setting algorithm in final iterations to derive a certificate of optimality. In our experiments, the heuristic results in significant speedups without strongly compromising solution quality.

3.4. Solution Algorithm

Our solution algorithm, summarized in Figure 2, involves two interconnected decomposition structures. An outer loop solves the MiND-VRP' via Benders decomposition: the BMP generates a first-stage solution and a lower bound, and the BSP generates a second-stage solution and an upper bound. At each outer iteration, the algorithm certifies the optimality of the MiND-VRP' solution, or otherwise generates an optimality cut in the BMP. Then, an inner loop solves the BSP via subpath-based column generation: the RBSP generates a BSP solution, and the PP identifies the subpath-based variables with minimal reduced cost for each pair of checkpoints and each load differential—using the label-setting algorithm. At each inner iteration, the algorithm generates new variables or certifies optimality of the BSP solution. Proposition 5 establishes the exactness of the algorithm, as long as the time discretization is sufficiently granular in the pricing problem.

PROPOSITION 5. *If all subpath travel times are strictly less than the elapsed time between the scheduled arrival times at the checkpoints, there exists a time discretization for the pricing problem such that the algorithm returns an optimal solution to MiND-VRP' in a finite number of iterations.*

Figure 3 illustrates the double-decomposition algorithm. The BMP solves a network design problem with all reference lines and all scenarios (Figure 3a). The BSP decomposes the second-stage routing operations across reference lines and scenarios (Figure 3b). Column generation further

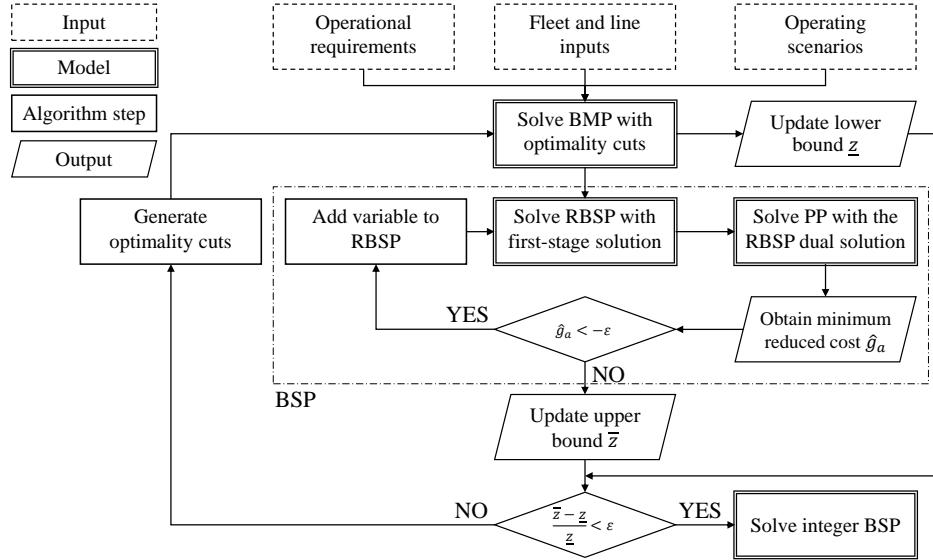


Figure 2 Overview of the algorithm, combining Benders decomposition and subpath-based column generation.

decomposes the BSP between checkpoints, exploiting the subpath-based second-stage formulation (Figure 3c). As a result, the pricing problem adds subpaths between checkpoints; the restricted Benders subproblem combines them to optimize the operating performance of each reference trip in each scenario; and the Benders master problem selects first-stage reference trips accordingly.

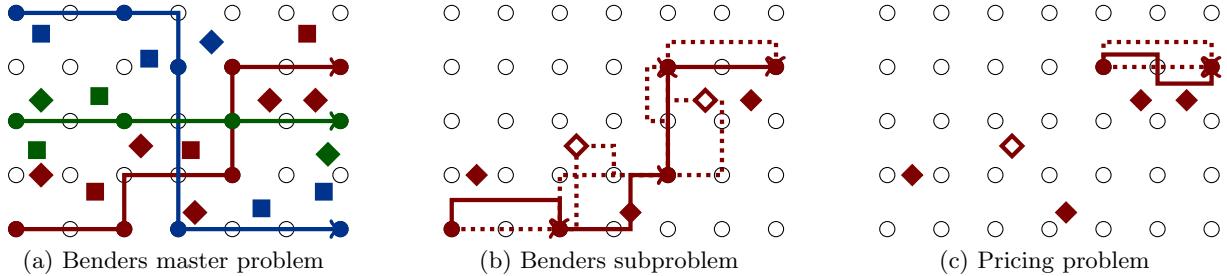


Figure 3 Double-decomposition algorithm. *Left:: BMP with three reference lines (blue, red, green); passenger requests in two scenarios (squares, diamonds) with their first-stage assignments (colors).* *Middle: BSP for one reference trip and one scenario; full diamonds encode passengers served; solid lines characterize selected subpaths in RBSP.* *Right: PP to generate new subpath between checkpoints (solid line).*

To avoid fractional second-stage variables and retrieve a feasible MiND-VRP solution, we solve the BMP one last time with integrality constraints. The optimal solution to the partial relaxation MiND-VRP' still provides a valid lower bound, hence an optimality gap. As we shall see experimentally, the optimality gap is very small due to the tight second-stage formulation.

In the MiND-DAR, the label-setting algorithm features a three-dimensional state space to track pickups and dropoffs, which weakens domination and increases computational requirements. Nonetheless, our methodology will scale to practical instances of the MiND-VRP and MiND-DAR.

4. Computational Assessment of Methodology

We develop a real-world experimental setup in Manhattan. We use demand data from the NYC Taxi & Limousine Commission (2021) during the morning rush (6–9 am). We define a road network and travel times using data from Google Maps, OpenStreetMap, and Uber (2020). Parameter values are reported in EC.4.1. We design candidate reference lines using breadth-first search trees over a comprehensive routing graph, which we cluster and filter for service quality (see EC.4.2).

We consider a MiND-VRP setting corresponding to a shuttle service from Manhattan to LaGuardia Airport with vehicles of capacity 10 to 20 passengers. We vary the number of candidate reference lines (5 to 100), the planning horizon (1 to 3 hours), whether on-demand deviations can skip a checkpoint ($K = 0$ vs. $K = 1$), and the number of scenarios (5 to 20). We use a 15-minute discretization to schedule transit vehicles in the first stage (sets \mathcal{T}_ℓ), and a 30-second discretization to route vehicles in the second stage (sets $\mathcal{T}_{\ell t}^{uv}$). Our problem includes up to 1,900 passenger requests, 640 stations, and 100 candidate reference lines (Figure EC.2 in EC.4), resulting in over 1 million first-stage variables, 25,000 Benders subproblems, and 200,000 pricing problems. We also develop a real-world experimental setup in Midtown Manhattan for the MiND-DAR in EC.1.3.

All models are solved with Gurobi v8.1 using the JuMP package in Julia (Dunning et al. 2017). We impose a three-hour time limit for optimization. All instances and code are available online.¹⁰

4.1. Benefits of Subpath Modeling and Double-decomposition Algorithm

Table 1 compares the three formulations in terms of solution quality (normalized to the best-found solution), computational time (preprocessing plus solution times), and number of second-stage variables. For a fair comparison, all models are solved with off-the-shelf methods, using our label-setting algorithm to enumerate subpaths and paths; and we define up to 1 million paths per subproblem in the path-based formulation. Note that the segment-based formulation does not scale to even small problems, requiring 30 million variables in the smallest instance due to the granular time discretization. The path-based formulation scales to medium instances but its performance quickly deteriorates due to the exponential number of path-based variables. In comparison, the subpath-based formulation requires orders of magnitude fewer variables, terminates much faster, and returns a superior solution with 10 candidate lines. These results underscore the benefits of the subpath-based formulation in terms of problem size, solution time, and solution quality. Yet, no formulation scales to larger instances, which motivates our double-decomposition algorithm.

Table 2 compares Benders decomposition with full subpath enumeration (“Benders”) and our double-decomposition algorithm with the exact and heuristic label-setting algorithms (“DD-E” and “DD-H”). Benders decomposition yields some speedups against direct implementation of the

¹⁰ <https://github.com/martiniradi/DeviatedFixedRouteMicrotransit>

Table 1 Comparison of path-based, subpath-based and segment-based MiND-VRP formulations.

L	S	Horizon	K	Path-based			Subpath-based			Segment-based		
				Sol.	CPU (s)	Arcs	Sol.	CPU (s)	Arcs	Sol.	CPU (s)	Arcs
5	5	60	0	100	117s	3.1M	100	19s	34K	100	6,633s	30.0M
5	5	60	1	—	—	—	—	—	—	100	9,435s	30.0M
5	5	120	0	100	760s	8.6M	100	279s	94K	—	—	—
5	5	180	0	100	801s	9.6M	100	345s	130K	—	—	—
10	5	60	0	101.6	1,278s	29.1M	100	60s	882K	—	—	—

“—” indicates that the algorithm does not terminate due to memory limitations.

subpath-based formulation (Table 1), but full subpath enumeration remains intractable in realistic instances. In comparison, our double-decomposition algorithm achieves stronger scalability by leveraging column generation in the Benders subproblem, especially with label-setting acceleration.

Table 2 Algorithm comparison for the subpath-based MiND-VRP model.

L	S	Horizon	K = 0						K = 1							
			Benders			DD-E			DD-H			DD-E				
			Sol.	Gap	CPU(s)	Sol.	Gap	CPU(s)	Sol.	CPU(s)	Sol.	Gap	CPU(s)	Sol.	CPU(s)	
5	5	60	100	0.0	24	100	0.0	13	103.3	8	100	0.0	242	101.2	18	
		120	100	0.0	325	100	0.0	56	101.8	26	100	0.0	5,753	101.5	45	
		180	100	0.0	369	100	0.0	67	102	51	100	0.0	4,395	101	66	
	20	60	—	—	—	100	0.0	94	102.6	39	100	0.0	3,536	100.9	85	
		120	—	—	—	100	0.0	478	102.4	150	100	2.3	10,800	100.6	211	
		180	—	—	—	100	0.0	529	102.3	230	100	3.3	10,800	100.5	306	
	10	5	60	100	0.0	48	100	0.0	82	102	57	100	0.0	6,222	100.3	75
		120	—	—	—	100	0.0	256	100.8	121	102.5	6.5	10,800	100	187	
		180	—	—	—	100	0.0	407	101.1	200	104.4	10.7	10,800	100	280	
20	5	60	—	—	—	100	0.1	789	102	328	108.7	31.3	10,800	100	404	
		120	—	—	—	101.2	4.4	10,800	100	1,228	—	—	—	100	912	
		180	—	—	—	103	9.9	10,800	100	2,782	—	—	—	100	1,247	
	50	60	—	—	—	100	0.2	2,093	100.1	649	—	—	—	100	10,800	
		120	—	—	—	100.6	3.2	10,800	100	10,800	—	—	—	100	10,800	
		180	—	—	—	100	6.8	10,800	100.4	10,800	—	—	—	100	10,800	
	20	60	—	—	—	104.6	7.3	10,800	100	10,800	—	—	—	100	10,800	
		120	—	—	—	100.9	1.1	10,800	100	2,802	—	—	—	100	10,800	
		180	—	—	—	105.9	9.9	10,800	100	10,800	—	—	—	100	10,800	
	20	60	—	—	—	—	—	—	100	10,800	—	—	—	100	10,800	

Optimality gap: integer MiND-VRP solution vs. lower bound from Benders decomposition, in percentage terms. [values in bold indicate that the algorithm has converged; for others, the algorithm reached the time limit.]

Specifically, when all checkpoints must be visited ($K = 0$), the DD-E algorithm terminates with the full set of 100 candidate reference lines and a one- to two-hour horizon. When vehicles can skip checkpoints ($K = 1$), the longer subpaths result in exponentially larger second-stage problems. In that case, full subpath enumeration fails to find feasible solutions even in the smallest instances, whereas the DD-E algorithm can solve instances with up to 10 candidate reference lines and a three-hour horizon. These improvements are driven by the small number of variables needed

to guarantee convergence in column generation; for instance, in the largest instance solved by subpath enumeration, the DD-E algorithm converges with 93% fewer variables. Moreover, the DD-E algorithm yields a zero optimality gap whenever it reaches convergence, confirming the tightness of our subpath-based formulation. Still, the DD-E algorithm leaves an optimality gap in these instances, and fails to return feasible solutions in the largest instances.

In comparison, our heuristic label-setting acceleration scales to the largest instances of the problem. These benefits are stronger when vehicles can skip checkpoints because the stronger dominance criterion becomes more impactful with longer subpaths. In small instances, the DD-H algorithm terminates up to 3 times faster than DD-E, while returning solutions within 3% of the DD-E solution. In medium instances, DD-H actually returns higher-quality solutions in faster computational times. This is because DD-E leaves an optimality gap, whereas DD-H converges more effectively due to a much smaller number of subpaths (up to 52% and 78% fewer subpaths when $K = 0$ and $K = 1$, respectively). In other words, the benefits of acceleration can outweigh the slight loss of flexibility when choosing which passengers to pick up at each station. Ultimately, by combining Benders decomposition, column generation and label-setting acceleration, our algorithm can solve realistic instances with up to 100 candidate reference lines, hundreds of stations, 5 demand scenarios and a three-hour horizon (or 20 demand scenarios and a one-hour horizon).

Similarly, despite the higher complexity of the problem, the methodology can handle realistic MiND-DAR instances in Midtown Manhattan with up to 10 candidate lines (EC.1.3).

4.2. Benefits of Stochastic Optimization Methodology

Table 3 reports the Value of the Stochastic Solution (VSS) and the Expected Value of Perfect Information (EVPI) to quantify the relative improvement of the stochastic optimization solution (“Sol.”) against a deterministic model and its relative loss against a clairvoyant model.

Note the high VSS: 5-7% on average and up to 10%. The MiND-VRP reduces unmet demand by 6-7% on average, while reducing passengers’ walking time by 25-35% from the deterministic baseline based on average demand. In fact, our stochastic optimization approach bridges 40-50% of the gap on average between the deterministic baseline and the perfect-information benchmark. These results highlight the benefits of our two-stage stochastic optimization formulation (and our double-decomposition algorithm) to increase demand coverage while maintaining or even improving level of service, as compared to a deterministic model (which can be solved via off-the-shelf methods).

5. Practical Assessment of Deviated Fixed-route Microtransit

Finally, we conduct a comprehensive assessment of microtransit against fixed-route transit and ride-sharing. Fixed-route transit corresponds to a single-stage variant without second-stage deviations. Ride-sharing corresponds to an on-demand system with vehicle capacities of 1, 2, and 4.

Table 3 Value of Stochastic Solution (VSS) and Expected Value of Perfect Information (EVPI)

\mathcal{L}	$ \mathcal{S} $	Horizon	K	Heur.	Performance assessment			VSS breakdown						
					$ \frac{\text{VSS}}{\text{Sol.}} $	$ \frac{\text{EVPI}}{\text{Sol.}} $	$\frac{\text{VSS}}{(\text{VSS}+\text{EVPI})}$	Unmet demand (%)	Walking time (%)	Waiting time (%)	Earliness (%)	Delay (%)	Detour (%)	
10	5	60	0	\times	5.8	8.2	41.2	-6	-59.3	0.7	-2.1	13.2	1	
			1	\times	2.7	9.7	22	-2.7	36.3	-1.6	-9.9	3.9	0.4	
		120	0	\times	4	7.4	35.2	-4.3	-77	1.2	1.7	13	0.5	
			1	\checkmark	3	7.4	28.6	-2.8	-1.6	-6.8	3.3	-4.8	-0.8	
		180	0	\times	9.9	7.4	57.4	-11	-70.6	7.4	3.4	-1.5	-0.4	
			1	\checkmark	7	5.7	55.2	-7.5	-17.4	0.4	-5.8	4.1	0.5	
	20	60	0	\times	7.6	5.9	56.3	-8.1	-91.6	1.1	-5.6	15.7	1	
			1	\checkmark	9.8	6.1	61.5	-10.4	-8.1	-10.2	-11.4	23.9	-1.2	
		120	0	\checkmark	9.7	9.2	51.1	-10.8	-81.2	4.3	10.5	4.2	0.2	
			1	\checkmark	4.8	8	37.8	-5	49.7	7.4	-6.9	10.3	0	
		180	0	\checkmark	6.1	7.2	45.7	-6.6	-82.9	6.4	6.9	14.5	3.2	
			1	\checkmark	2.7	7.3	27.2	-2.8	-2.8	5.1	-0.6	14	1.4	
50	5	60	0	\times	5.8	2.5	70.3	-6.5	-6.1	0	-13	6.2	0.3	
		120	0	\checkmark	5.3	7	43.1	-5.7	-2	1.3	2.7	1.8	1	
		180	0	\times	4.3	10.7	30.5	-4.5	-16.7	-2.1	8.3	4.2	0.8	
100	5	60	0	\checkmark	8.8	2.3	79.4	-11.1	-56.2	-19.5	1.3	-1.5	-0.7	
		120	0	\checkmark	5	7.4	40.2	-5.6	-25.2	-7.6	0.4	0.1	0.6	
		180	0	\checkmark	2.4	15.5	13.4	-2.7	-3.7	-5.6	0.3	-3.2	1	
Average 5 scenarios					5.3	7.6	43.0	-5.9	-25.0	-2.7	-0.8	3.0	0.4	
Average 20 scenarios					6.8	7.3	46.6	-7.3	-36.2	2.4	-1.2	13.8	0.8	

"Heur.": solution from DD-H algorithm (\checkmark) vs DD-E algorithm (\times); "Sol.": stochastic optimization solution. Unmet demand is measured in number of passengers; all other components are measured per served passenger.

We optimize ride-sharing operations by leveraging and adapting the approach from Bertsimas and Yan (2021), as described in EC.4.3. We use the same experimental setup as in Section 4. Recall that, since Manhattan represents a high-density region with good transit options, the results can be seen as conservative estimates of the impact of microtransit in other, lower-density areas with fewer transit options. Again, all our insights hold in the MiND-DAR, as shown in EC.1.3.

5.1. Value of Microtransit Flexibility

Microtransit operations. We first compare transit vs. microtransit operations, for a given set of reference lines. Figure 4 illustrates the MiND-DAR operations along two lines in Midtown Manhattan. By design, the transit system follows the reference line; in contrast, the microtransit system deviates from the reference line in all but one checkpoint pair. As a result, the microtransit system serves more passengers (24 versus 8), at the cost of a longer distance traveled (8,545 vs. 5,502 meters). Still, the higher vehicle loads leads to a smaller distance traveled per passenger (356 vs. 699 meters per passenger). These results indicate benefits of microtransit for the operator (lower costs), for passengers (lower fares) and for the environment (smaller footprint).

We now evaluate the operating performance of microtransit vs. transit in the MiND-VRP with the same set of 50 reference lines; we vary vehicle capacities (10, 15, 20 passengers), the maximum deviation (low vs. high, or $\Delta = 600$ vs. $\Delta = 1,200$ meters) and whether vehicles can skip a checkpoint ($K = 0$ vs. $K = 1$). Table 4 reports average operating performance—coverage, vehicle utilization, distance, distance per pickup—and level of service—walk, wait, detour, and delay.

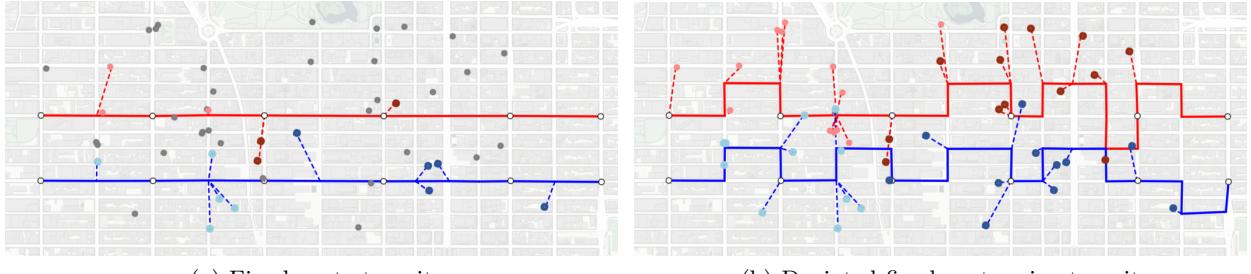


Figure 4 Illustration of transit and microtransit operations in the MiND-DAR for two reference lines [light (resp. dark) blue/red circles: origins (resp. destinations) of passengers served by the blue/red line; grey circles: origins and destinations of unserved passengers; white circles: checkpoints].

Table 4 Average operating performance and level of service for fixed-route transit and microtransit.

Cap.	Mode	Operating model		Average operating performance				Average level of service			
		Dev.	Skip?	#pass./vehicle	Util.	Dist.	Dist./pass.	Walk	Wait	Detour	Delay
10	Transit	—	—	8.05	80.50%	14.78	2.66	2.19	6.77	152.06%	-0.87
	Microtransit	Low	$K = 0$	8.61	86.12%	16.48	2.53	1.48	5.69	154.36%	-0.47
	Microtransit	High	$K = 0$	8.63	86.32%	16.57	2.51	1.44	5.58	153.62%	-0.44
	Microtransit	Low	$K = 1$	8.78	87.81%	17.08	2.54	1.17	4.51	156.50%	-0.28
	Microtransit	High	$K = 1$	8.99	89.87%	17.39	2.38	1.03	4.33	153.72%	-0.28
15	Transit	—	—	10.72	71.47%	15.06	2.40	2.27	6.88	150.73%	-1.29
	Microtransit	Low	$K = 0$	12.20	81.32%	17.17	2.17	1.57	5.90	151.82%	-0.49
	Microtransit	High	$K = 0$	12.29	81.96%	17.34	2.14	1.50	5.74	151.20%	-0.46
	Microtransit	Low	$K = 1$	12.56	83.70%	17.78	2.15	1.31	4.83	154.83%	-0.26
	Microtransit	High	$K = 1$	12.89	85.95%	18.15	1.97	1.15	4.63	151.74%	-0.31
20	Transit	—	—	12.24	61.21%	15.16	2.34	2.30	6.94	150.38%	-1.84
	Microtransit	Low	$K = 0$	15.28	76.42%	17.52	2.02	1.69	6.21	150.77%	-0.52
	Microtransit	High	$K = 0$	15.46	77.32%	17.72	1.98	1.62	6.04	150.08%	-0.50
	Microtransit	Low	$K = 1$	15.90	79.49%	18.13	1.96	1.48	5.25	153.51%	-0.28
	Microtransit	High	$K = 1$	16.16	80.78%	18.57	1.81	1.43	5.39	151.17%	-0.33

“Cap.” – Capacity; “Pass.” – Passenger; “Util.” – Utilization; “Dist.” – Distance; “Dev.” – deviation.

Units: distance, distance per passenger – kilometers; walk, wait, delay/earliness – minutes.

Parameters: two-hour horizon; 10 weekday scenarios, maximum walk: 7 minutes, maximum wait: 10 minutes.

On average, the microtransit system can add 1-4 passengers per vehicle while reducing walking times by 50% and wait times by 2 minutes. These benefits come at a cost of a small increase in detours (+2%) and an increase in distance traveled (+15-25%). Still, due to the large increase in utilization, distance per passenger is reduced by up to 500 meters, or 23%. These benefits become stronger with larger vehicles, underlining the value of flexibility in high-capacity microtransit. Interestingly, even when microtransit vehicles are constrained to stay close to the reference lines (low deviation) and to visit all checkpoints ($K = 0$), the microtransit system can significantly improve coverage (0.5 to 3 extra passengers per vehicle, on average) and level of service (reduction in walking times by 40 seconds and in waiting times by 1 minute). In other words, even limited extents of demand-responsiveness can achieve significant performance improvements through stronger demand consolidation, higher level of service, and a smaller environmental footprint per passenger.

Figure 5 plots the average vehicle load in fixed-route transit (x axis) vs. deviated fixed-route microtransit (y axis), for each reference line broken down into low, medium and high density (colored lines) and for each vehicle capacity (dot shapes). Microtransit makes use of the deviations to improve vehicle load—all observations lie above the 45-degree line—but these improvements on population density and vehicle capacity. In low-density regions, microtransit operations can serve a few extra passengers but vehicles do not operate at capacity. In high-density regions, fixed-route transit already provides high demand coverage, so the relative improvements are more limited, especially with low-capacity vehicles. In comparison, the increase in vehicle load is strongest on medium-density lines (+0-5 and +5-10 pickups with low- and high-capacity vehicles, respectively).

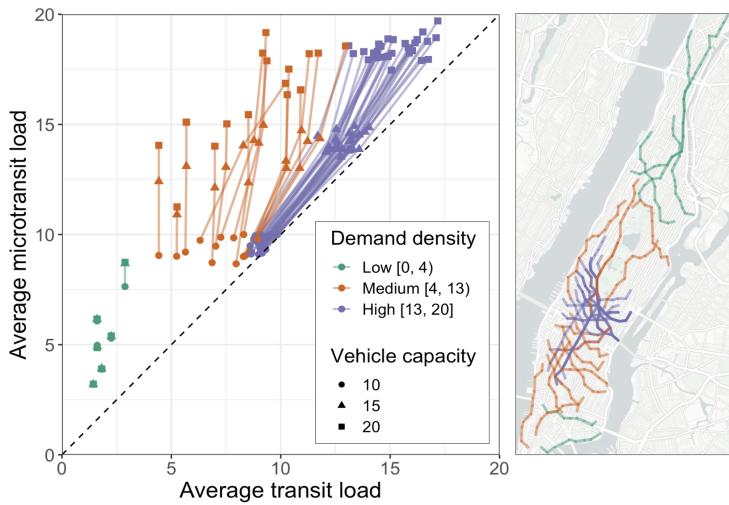


Figure 5 Value of operating flexibility ($\Delta = 1,200 \text{ m.}$, $K = 0$). Low (resp. medium, high) density: lines with maximum load less than 4 (resp. 4 to 13, more than 13) passengers on average under transit.

These results identify a medium-density regime where high-capacity microtransit can be most impactful. In low-density regions, on-demand door-to-door transportation can provide high levels of service with small-occupancy vehicles with limited detours and delays. Vice versa, high-density regions can be effectively served with fixed-route transit due to high synergies across passengers. In-between, deviated fixed-route microtransit provides the strongest benefits—increases in demand coverage and gains in passenger level of service—when population density is high enough to consolidate demand into high-occupancy vehicles but too low for fixed-route transit to be as effective.

Microtransit network design. Figure 6 depicts the optimized first-stage networks of reference lines under transit and microtransit, along with the resulting catchment areas. Reference lines are labelled as “selected” if at least one corresponding reference trip is selected over the planning horizon. The figure also depicts the number of trip options from each of Manhattan’s 21,000

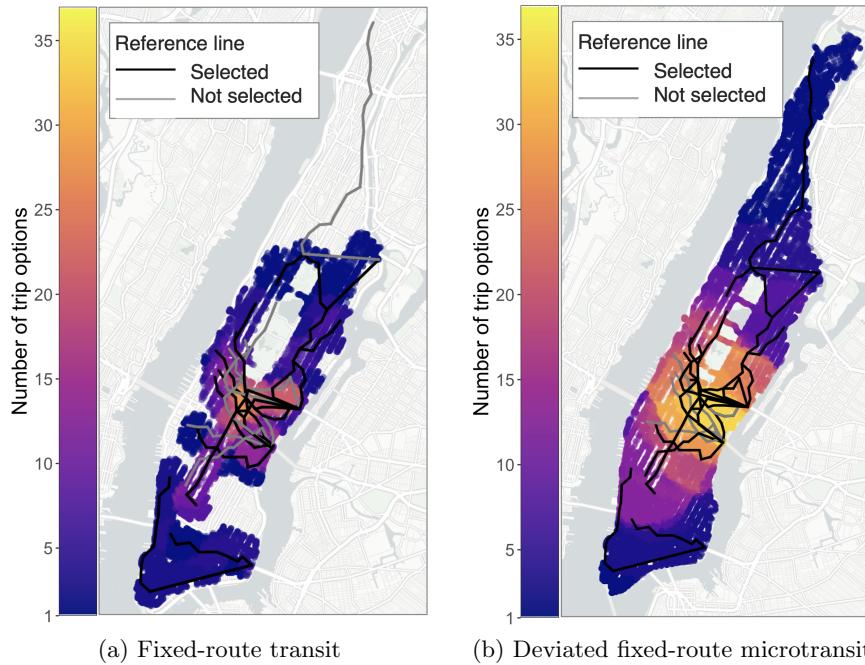


Figure 6 Reference lines and catchment areas. Parameters: 25 candidate reference lines, 2-hour horizon, 20 vehicles with a 20-passenger capacity each. Microtransit parameters: $\Delta = 1,200$ m., $K = 0$.

roadway intersections, defined as the number of reference trips throughout the planning horizon with a candidate pickup location within a 5-minute walking radius of the intersection.

Note that the microtransit network expands the catchment area from fixed-route transit. Consistently with our findings from Figure 5, the fixed-route transit system mostly selects lines in high-demand areas, where it is most effective. Due to its operating flexibility, microtransit provides more trip options in high-demand regions with fewer lines: in Midtown Manhattan for instance, the number of trip options increases from 20-25 for transit to 30-35 for microtransit; overall, microtransit increases the average number of trip options per intersection by a factor of 3 (8.31 vs. 2.61). Most importantly, the microtransit system reaches low-demand regions, such as Uptown Manhattan—microtransit covers 60% more intersections with at least one trip option (53.8% vs. 85.4%). In summary, microtransit provides more traveling options in high-density regions and expands its geographic reach to under-served regions, thus enhancing accessibility across the population.

5.2. Performance Assessment of Microtransit

We now compare the performance of microtransit against fixed-route transit and ride-sharing. To establish an apples-to-apples comparison, we fix total seating capacity across all systems (e.g., 10 transit/microtransit vehicles of capacity 10, ride-sharing with 100/50/25 vehicles of capacity of 1/2/4), and perform an out-of-sample assessment corresponding to five new weekdays. Unlike in Table 4 and Figure 5, we consider here the optimized network of reference lines in transit and microtransit. Table 5 reports average coverage, level of service, and distance traveled.

Table 5 Average level of service of fixed-route transit, microtransit ($\Delta = 1,200$ m., $K = 0$), and ride-sharing.

Mode	Design	Coverage	Walk	Wait	Detour	Delay	Distance
Transit	5 candidate lines	13.9%	2.06	7.06	158.56%	-1.17	356
	10 candidate lines	20.4%	2.21	6.91	146.22%	-0.79	384
	25 candidate lines	29.8%	2.03	6.8	136.98%	0.13	435
	50 candidate lines	33.6%	2.03	6.65	137.34%	-0.06	472
Microtransit	5 candidate lines	22.3%	1.68	6.22	159.99%	-0.01	419
	10 candidate lines	30.0%	1.68	6.22	146.11%	-0.15	462
	25 candidate lines	35.6%	1.53	5.82	138.52%	-0.16	471
	50 candidate lines	36.6%	1.36	5.55	141.00%	0.03	468
Rideshare	Cap. 4	36.3%	0	4.2	150.68%	13.4	1,883
	Cap. 2	44.7%	0	3.74	124.60%	8.17	3,359
	Cap. 1	50.5%	0	1.79	100.00%	1.79	5,671

Coverage: percentage of served requests; distance in kilometers; walk, wait, delay/earliness in minutes.

These results confirm that microtransit increases demand coverage and reduces walk and wait times as compared to fixed-route transit, at virtually no cost in terms of detours and delays. At the other extreme, single-occupancy ride-sharing achieves high coverage with no walking (by design) and short waits. However, single-occupancy ride-sharing results in much longer distances traveled; it also leads to higher passenger delays due to the on-demand dispatches. Thus, microtransit defines a middle ground between transit and ride-sharing: less walk and less wait for passengers than in transit and less delays than in ride-sharing, and intermediate ridership and operating costs.

Another interesting observation stems from the comparison of microtransit to ride-pooling. Both modes leverage on-demand operations to consolidate demand into multi-occupancy vehicles, but do so differently. By relying on on-demand dispatch and door-to-door transportation, ride-pooling results in no walk and low wait times but increases detours and delays—underscoring the impact of spatiotemporal externalities across users, even with small-occupancy vehicles. By consolidating demand into high-capacity vehicles along reference lines, deviated fixed-route microtransit reduces distance traveled by a factor of 4 but reaches similar demand coverage and achieves a comparable level of service—no delay, smaller detours, moderate walking times, and slightly longer wait times. These results identify deviated fixed-route microtransit as a promising pathway to provide efficient and convenient urban mobility options at scale with high-capacity vehicles.

Figure 7 provides a system-wide assessment of each mode by plotting total distance traveled (Figures 7a) and vehicle loads (Figures 7b), averaged out of 5 out-of-sample scenarios. Total distance traveled is used as a proxy of operating costs and environmental footprint; it includes both the “internal” distance for all served passengers plus the “external” distance from single-occupancy trips for all unserved passengers (assuming, e.g., that unserved passengers take a taxi to the airport).

These results show that microtransit reduces total distance by 10-15% versus fixed-route transit, by 20-30% versus ride-pooling, and by 50% versus single-occupancy ride-sharing. Recall that microtransit involves a much smaller internal distance but lower demand coverage than ride-sharing

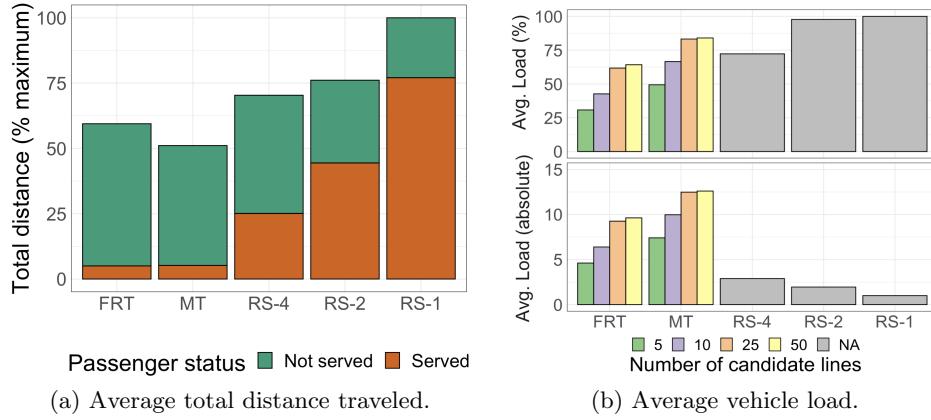


Figure 7 System-wide assessment of fixed-route transit (FRT), microtransit ($\Delta = 1,200 \text{ m.}$, $K = 0$) (MT), and ride-sharing systems with capacities 4, 2 and 1 (RS-4, RS-2 and RS-1, respectively).

(Table 5). Figure 7a shows that microtransit still results in much smaller total distance after accounting for unserved passengers. Vice versa, microtransit involves higher demand coverage but slightly longer internal distances than fixed-route transit (Table 5); Figure 7a shows that the increase in utilization outweighs the longer distances due to on-demand deviations, leading to a smaller distance per passenger and thus a smaller total distance. Altogether, these results suggest that deviated fixed-route microtransit can contribute to more affordable on-demand mobility and to mitigating the environmental footprint of the mobility sector, by achieving stronger demand consolidation than ride-sharing and ride-pooling (higher vehicle loads in Figure 7b) and higher demand coverage than fixed-route transit (smaller distance for unserved passengers in Figure 7a).

Results in EC.1.3 confirm the findings from Table 5 and Figure 7 for the MiND-DAR: deviated fixed-route microtransit increases demand coverage versus fixed-route transit, improves demand consolidation versus ride-sharing and ride-pooling, and reduces total distance versus all benchmarks (by 5-15% vs. fixed-route transit, by 40-50% vs. ride-pooling, and by over 100% vs. single-occupancy ride-sharing). Thus, our insights are robust across vehicle routing and dial-a-ride settings.

In conclusion, results suggest that deviated fixed-route microtransit can contribute to more efficient, equitable, and sustainable mobility. Efficiency stems from high levels of service, low operating costs and high demand coverage enabled by the combination of reference lines and on-demand operating flexibility. Equity stems from a microtransit design with broader geographic reach, hence higher accessibility, enabled by the higher demand coverage of each reference line. Sustainability stems from a smaller distance traveled per passenger, hence a lower environmental footprint, enabled by high demand coverage and demand consolidation into high-capacity vehicles.

6. Conclusion

This paper optimizes the design and operations of a deviated fixed-route microtransit system endowed with advance planning capabilities along reference lines (as in public transit) and on-demand

adjustments in response to passenger demand (as in ride-sharing). We formulated a multi-objective *Microtransit Network Design (MiND)* model via two-stage stochastic optimization to minimize operating costs, maximize demand coverage, and maximize passenger level of service—including walking times, waiting times, in-vehicle travel times, and arrival delay. The model leverages subpath-based variables between checkpoints on the reference lines to maintain a tight second-stage formulation over a load-expanded network. We developed a double-decomposition solution algorithm, leveraging Benders decomposition to decompose the problem per scenario and per reference trip, as well as subpath-based column generation to further decompose operations between checkpoints.

Using New York City data, results showed that the methodology scales to real-world and otherwise-intractable problems, with up to 100 candidate reference lines, hundreds of stations, thousands of requests, and 5-20 demand scenarios. From a practical standpoint, even limited on-demand flexibility can provide significant operating benefits by improving demand consolidation over ride-sharing—pooling passenger demand into high-capacity vehicles along reference lines—and demand coverage over fixed-route transit—leveraging on-demand deviations to enhance passenger level of service and reach more passengers. At a time where hybrid solutions are emerging to design new mobility services combining the strengths of public transit and ride-sharing, this paper suggests that deviated fixed-route microtransit can contribute to efficient mobility (high demand coverage, low operating costs, high levels of service), equitable mobility (high accessibility with broad geographic reach), and sustainable mobility (low environmental footprint). Based on these results, we have been collaborating with transit operators toward the deployment of deviated fixed-route microtransit, with a pilot implementation targeted for Summer 2024.

Acknowledgments

The authors thank Joseph Kajon for helpful assistance with data analyses and experimentation.

References

- Allen DJ (2017) *Lost in the transit desert: Race, transit access, and suburban form* (Routledge).
- Alonso-Mora J, Samaranayake S, Wallar A, Frazzoli E, Rus D (2017) On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences* 114(3):462–467.
- Banerjee S, Hssaine C, Périvier N, Samaranayake S (2021) Real-time approximate routing for smart transit systems. *arXiv preprint arXiv:2103.06212* .
- Bertsimas D, Delarue A, Martin S (2019a) Optimizing schools' start time and bus routes. *Proceedings of the National Academy of Sciences* 116(13):5943–5948.
- Bertsimas D, Jaillet P, Martin S (2019b) Online vehicle routing: The edge of optimization in large-scale applications. *Operations Research* 67(1):143–162.

- Bertsimas D, Ng YS, Yan J (2021) Data-driven transit network design at scale. *Operations Research* 69(4):1118–1133.
- Bertsimas D, Yan J (2021) The edge of optimization in large-scale vehicle routing for paratransit. *Preprint*.
- Blanchard M, Jacquillat A, Jaillet P (2023) Probabilistic bounds on the k –traveling salesman problem and the traveling repairman problem. *Mathematics of Operations Research*.
- Borndörfer R, Grötschel M, Pfetsch ME (2007) A column-generation approach to line planning in public transport. *Transportation Science* 41(1):123–132.
- Ceder A, Wilson NH (1986) Bus network design. *Transportation Research Part B: Methodological* 20(4):331–344.
- Chopra S, Martin S, Mishra PS, Smilowitz K (2023) Mobility-on-demand meets shuttles on the same mile. *Available at SSRN 4322824*.
- Cummings K, Vaze V, Ergun Ö, Barnhart C (2023) Multimodal transportation alliance design with endogenous demand: Large-scale optimization for rapid gains. *arXiv preprint arXiv:2301.03414*.
- Desaulniers G, Hickman MD (2007) Public transit. *Handbooks in operations research and management science* 14:69–127.
- Dunning I, Huchette J, Lubin M (2017) JuMP: A modeling language for mathematical optimization. *SIAM review* 59(2):295–320.
- Galarza Montenegro BD, Sörensen K, Vansteenwegen P (2022) A column generation algorithm for the demand-responsive feeder service with mandatory and optional, clustered bus-stops. *Networks* 80(3):274–296.
- Guan H, Basciftci B, Van Hentenryck P (2023) Path-based formulations for the design of on-demand multimodal transit systems with adoption awareness. *arXiv preprint arXiv:2301.07292*.
- Karsten CV, Ropke S, Pisinger D (2018) Simultaneous optimization of container ship sailing speed and container routing with transit time restrictions. *Transportation Science* 52(4):769–787.
- Liu X, Qu X, Ma X (2021) Improving flex-route transit services with modular autonomous vehicles. *Transportation Research Part E: Logistics and Transportation Review* 149:102331.
- Ma TY, Rasulkhani S, Chow JY, Klein S (2019) A dynamic ridesharing dispatch and idle vehicle repositioning strategy with integrated transit transfers. *Transportation Research Part E* 128:417–442.
- Maheo A, Kilby P, Van Hentenryck P (2019) Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science* 53(1):77–88.
- Marín ÁG, Jaramillo P (2009) Urban rapid transit network design: accelerated benders decomposition. *Annals of Operations Research* 169(1):35–53.
- McKinsey & Co (2018) Travel and logistics: data drives the race for customers. Technical report.

- McKinsey & Co (2021) Shared mobility: Where it stands, where it's headed. Technical report.
- Muter I, Birbil Şİ, Bülbül K (2013) Simultaneous column-and-row generation for large-scale linear programs with column-dependent-rows. *Mathematical Programming* 142(1-2):47–82.
- NYC Taxi & Limousine Commission (2021) TLC Trip Record Data. Available at: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.
- Ortega FA, Pozo MA, Puerto J (2018) On-line timetable rescheduling in a transit line. *Transportation Science* 52(5):1106–1121.
- Quadrifoglio L, Dessouky MM, Ordóñez F (2008) Mobility allowance shuttle transit (MAST) services: MIP formulation and strengthening with logic constraints. *EJOR* 185(2):481–494.
- Quadrifoglio L, Dessouky MM, Palmer K (2007) An insertion heuristic for scheduling mobility allowance shuttle transit (MAST) services. *Journal of Scheduling* 10(1):25–40.
- Quadrifoglio L, Hall RW, Dessouky MM (2006) Performance and design of mobility allowance shuttle transit services: bounds on the maximum longitudinal velocity. *Transportation science* 40(3):351–363.
- Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, Ratti C (2014) Quantifying the benefits of vehicle pooling with shareability networks. *PNAS* 111(37):13290–13294.
- Schaller B (2018) The new automobility: Lyft, uber and the future of american cities. Technical report, Schaller Consulting.
- Silva DF, Vinel A, Kirkici B (2022) On-demand public transit: A markovian continuous approximation model. *Transportation Science* 56(3):704–724.
- Statista (2023) Average quarterly cost per ride of ridesharing services. Technical report.
- Steiner K, Irnich S (2020) Strategic planning for integrated mobility-on-demand and urban public bus networks. *Transportation Science* 54(6):1616–1639.
- Stiglic M, Agatz N, Savelsbergh M, Gradisar M (2018) Enhancing urban mobility: Integrating ride-sharing and public transit. *Computers & Operations Research* 90:12–21.
- Sun L, Xie W, Witten T (2023) Distributionally robust fair transit resource allocation during a pandemic. *Transportation science* 57(4):954–978.
- Szufel, Przemysllaw *et al* (2023) OpenStreetMapX.jl. <https://github.com/pszufe/OpenStreetMapX.jl>.
- The Economist (2018) Public transport is in decline in many wealthy cities. www.economist.com/international/2018/06/21/public-transport-is-in-decline-in-many-wealthy-cities.
- Uber (2020) New york city: Quarterly speed statistics by hour of day (q1 2020). Acc. Nov 2022 at [https://movement.uber.com/cities/new_york/downloads/speeds?lang=en-US&tp\[y\]=2020&tp\[q\]=1](https://movement.uber.com/cities/new_york/downloads/speeds?lang=en-US&tp[y]=2020&tp[q]=1).
- US DoT (2016) Shared mobility current practices and guiding principles. Technical report.
- Vazifeh MM, Santi P, Resta G, Strogatz SH, Ratti C (2018) Addressing the minimum fleet problem in on-demand urban mobility. *Nature* 557(7706):534–538.

- Walteros JL, Medaglia AL, Riaño G (2015) Hybrid algorithm for route design on bus rapid transit systems. *Transportation Science* 49(1):66–84.
- Wan QK, Lo HK (2003) A mixed integer formulation for multiple-route transit network design. *Journal of Mathematical Modelling and Algorithms* 2(4):299–308.
- Wei K, Vaze V, Jacquillat A (2022) Transit planning optimization under ride-hailing competition and traffic congestion. *Transportation Science* 56(3):725–749.
- Westervelt M, Huang E, Schank J, Borgman N, Fuhrer T, Peppard C, Narula-Woods R (2018) Uprouted: Exploring microtransit in the united states. Technical report, Eno Center for Transportation.
- Zeighami V, Soumis F (2019) Combining benders' decomposition and column generation for integrated crew pairing and personalized crew assignment problems. *Transportation Science* 53(5):1479–1499.
- Zhang W, Jacquillat A, Wang K, Wang S (2023) Routing optimization with vehicle–customer coordination. *Management Science* .
- Zhao J, Dessouky M (2008) Service capacity design problems for mobility allowance shuttle transit systems. *Transportation Research Part B: Methodological* 42(2):135–146.

Deviated Fixed-route Microtransit: Design and Operations Electronic Companion

EC.1. Extension to the dial-a-ride setting (MiND-DAR)

EC.1.1. Modeling extension

In the dial-a-ride setting, each passenger request $p \in \mathcal{P}$ is associated with an origin $o(p)$ and a destination $d(p)$. The first-stage formulation remains unchanged, except that the set \mathcal{M}_p is re-defined as the set of reference lines that cover both the origin and the destination of request $p \in \mathcal{P}$. In the second stage, we define the sets \mathcal{P}_r^+ and \mathcal{P}_r^- ($\mathcal{P}_r = \mathcal{P}_r^+ \cup \mathcal{P}_r^-$) as the passenger requests that are picked up and dropped off, respectively, by subpath $r \in \mathcal{R}_{\ell st}$ for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$.

Level of service involves similar measures of passenger dis-utility. A subpath $r \in \mathcal{R}_{\ell st}$ is associated with a walking cost both for pickups (from the origin of passenger $p \in \mathcal{P}_r^+$ to the pickup location) and for dropoffs (from the dropoff location to the destination of passenger $p \in \mathcal{P}_r^-$); with a waiting cost for pickups; and with a delay cost for dropoffs. To capture detour costs, we denote by $T_{r(a)p}^+$ (resp. $T_{r(a)p}^-$) the pickup (resp. dropoff) time of passenger p on arc $a \in \mathcal{A}_{\ell st}$ such that $p \in \mathcal{P}_{r(a)}^+$ (resp. $p \in \mathcal{P}_{r(a)}^-$). The arc costs g_a are re-derived as follows.

$$g_a^{DAR} = \begin{cases} \sum_{p \in \mathcal{P}_{r(a)}^+} D_{ps} \left(\lambda \tau_{r(a)p}^{walk} + \mu \tau_{r(a)p}^{wait} - \sigma \frac{T_{r(a)p}^+}{\tau_p^{dir}} - M \right) + \\ \sum_{p \in \mathcal{P}_{r(a)}^-} D_{ps} \left(\lambda \tau_{r(a)p}^{walk} + \sigma \frac{T_{r(a)p}^-}{\tau_p^{dir}} + \delta \frac{\tau_{r(a)p}^{late}}{\tau_p^{dir}} + \frac{\delta}{2} \frac{\tau_{r(a)p}^{early}}{\tau_p^{dir}} \right) & \forall a \in \bigcup_{r \in \mathcal{R}_{\ell st}} \mathcal{A}_r, \\ 0 & \forall a \in \mathcal{A}_{\ell st}^v. \end{cases} \quad (\text{EC.1})$$

The MiND-DAR is then formulated as follows. The only difference with the MiND-VRP is the additional constraint ensuring that a passenger who is picked up needs to be dropped off (Equation (EC.5)). Note that the precedence constraint is captured by the set \mathcal{M}_p and therefore does not need to be enforced explicitly in the MiND-DAR formulation.

$$\min \quad \sum_{\ell \in \mathcal{L}} \sum_{t \in \mathcal{T}_\ell} h_\ell x_{\ell t} + \sum_{s \in \mathcal{S}} \pi_s \left(\sum_{\ell \in \mathcal{L}} \sum_{t \in \mathcal{T}_\ell} \sum_{a \in \mathcal{A}_{\ell st}} g_a^{DAR} y_a \right) \quad (\text{EC.2})$$

$$\text{s.t.} \quad \text{First-stage constraints: Equations (1)–(4)} \quad (\text{EC.3})$$

$$\text{Second-stage constraints: Equations (9)–(10)} \quad (\text{EC.4})$$

$$\sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}^+} y_a - \sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}^-} y_a = 0 \quad \forall s \in \mathcal{S}, p \in \mathcal{P}, (\ell, t) \in \mathcal{M}_p \quad (\text{EC.5})$$

$$\mathbf{x}, \mathbf{y}, \mathbf{z} \text{ binary} \quad (\text{EC.6})$$

EC.1.2. Algorithmic extension

Benders decomposition. For a reference trip (ℓ, t) and a scenario s , let ζ_p denote the dual variable associated to the new consistency constraint between pickup and drop-off decisions (Equation (EC.5)). The Benders dual subproblem becomes:

$$\max \quad x_{\ell t} \cdot (\varphi_{\bar{u}_{\ell st}} - \varphi_{\bar{v}_{\ell st}}) - \sum_{p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p} z_{p\ell st} \cdot \gamma_p \quad (\text{EC.7})$$

$$\text{s.t.} \quad \varphi_n - \varphi_m - \sum_{p \in \mathcal{P}_{r(a)}^+} (\gamma_p - \zeta_p) - \sum_{p \in \mathcal{P}_{r(a)}^-} \zeta_p \leq g_a^{DAR} \quad \forall a = (n, m) \in \mathcal{A}_{\ell st} \quad (\text{EC.8})$$

$$\varphi_i \in \mathbb{R} \quad \forall i \in \mathcal{V}_{\ell st} \quad (\text{EC.9})$$

$$\gamma_p \geq 0 \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (\text{EC.10})$$

$$\zeta_p \in \mathbb{R} \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (\text{EC.11})$$

Note that the new dual variables do not appear in the dual objective function, so the Benders optimality cut remains unchanged (Equation (21)).

Column generation. The restricted Benders subproblem is still obtained by restricting the decisions to a subset of arc-based variables in $\mathcal{A}'_{\ell st}$:

$$\text{RBSP}(\mathcal{A}'_{\ell st}, \mathbf{x}, \mathbf{z}) \quad \min_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{a \in \mathcal{A}'_{\ell st}} g_a^{DAR} y_a \quad (\text{EC.12})$$

$$\text{s.t.} \quad \sum_{m: (n, m) \in \mathcal{A}'_{\ell st}} y_{(n, m)} - \sum_{m: (m, n) \in \mathcal{A}'_{\ell st}} y_{(m, n)} = \begin{cases} x_{\ell t} & \text{if } n = \bar{u}_{\ell st}, \\ -x_{\ell t} & \text{if } n = \bar{v}_{\ell st}, \\ 0 & \text{otherwise,} \end{cases} \quad \forall n \in \mathcal{V}_{\ell st} \quad (\text{EC.13})$$

$$\sum_{a \in \mathcal{A}'_{\ell st} : p \in \mathcal{P}_{r(a)}^+} y_a - \sum_{a \in \mathcal{A}'_{\ell st} : p \in \mathcal{P}_{r(a)}^-} y_a = 0 \quad \forall p \in \mathcal{P}, (\ell, t) \in \mathcal{M}_p \quad (\text{EC.14})$$

$$\sum_{a \in \mathcal{A}'_{\ell st} : p \in \mathcal{P}_{r(a)}} y_a \leq z_{p\ell st} \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (\text{EC.15})$$

In the pricing problem, we split the level-of-service parameter d_{mp} into d_{mp}^+ and d_{mp}^- , corresponding to the level-of-service components associated with pickups and dropoffs, respectively. Following Section EC.1.1, we denote by \mathcal{P}_m^+ (resp. \mathcal{P}_m^-) the set of passengers that can be picked up (resp. dropped off) and by T_{mp}^+ (resp. T_{mp}^-) the pickup time (resp. dropoff time) of passenger $p \in \mathcal{P}_m^+$ (resp. $p \in \mathcal{P}_m^-$). We then define:

$$d_{mp}^+ = D_{ps} \left(\lambda \tau_{mp}^{\text{walk}} + \mu \tau_{mp}^{\text{wait}} - \sigma \frac{T_{mp}^+}{\tau_p^{\text{dir}}} - M \right) + \gamma_p - \zeta_p, \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, p \in \mathcal{P}_m^+$$

$$d_{mp}^- = D_{ps} \left(\frac{\delta \tau_{mp}^{\text{late}} + \frac{\delta}{2} \tau_{mp}^{\text{early}} + \sigma T_{mp}^-}{\tau_p^{\text{dir}}} + \lambda \tau_{mp}^{\text{walk}} \right) + \zeta_p, \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, p \in \mathcal{P}_m^-$$

Similarly, we define the following decision variables to split pickups and dropoffs:

$$f_{mq} = \begin{cases} 1 & \text{if arc } (m, q) \in \mathcal{H}_{\ell st}^{uv} \text{ is traversed in the time-expanded road segment network,} \\ 0 & \text{otherwise.} \end{cases}$$

$$w_{mp}^+ = \begin{cases} 1 & \text{if passenger } p \in \mathcal{P}_m^+ \text{ is picked up at node } m \in \mathcal{U}_{\ell st}^{uv}, \\ 0 & \text{otherwise.} \end{cases}$$

$$w_{mp}^- = \begin{cases} 1 & \text{if passenger } p \in \mathcal{P}_m^- \text{ is dropped off at node } m \in \mathcal{U}_{\ell st}^{uv}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\xi_m = \text{vehicle load in node } m \in \mathcal{U}_{\ell st}^{uv}$$

The pricing problem is them formulated as follows. Equation (EC.16) minimizes the reduced cost. Constraints (EC.17)–(EC.19) define the load at each node based on the pickups and dropoffs. Constraints (EC.20) and (EC.21) ensure that a passenger can only be picked up or dropped off in a node that is visited. Constraints (EC.22) and (EC.23) guarantee that a passenger is picked up and dropped off at most once, respectively. Constraints (32) apply flow balance in the time-expanded road segment network. The remaining constraints enforce binary requirements.

$$\min \sum_{m \in \mathcal{U}_{\ell st}^{uv}} \left(\sum_{p \in \mathcal{P}_m^+} d_{mp}^+ w_{mp}^+ + \sum_{p \in \mathcal{P}_m^-} d_{mp}^- w_{mp}^- \right) + \varphi_{end(a)} - \varphi_{start(a)} \quad (\text{EC.16})$$

$$\text{s.t. } \xi_{(u, T_{\ell t}(u))} = c_{(u, c_1)}, \quad \xi_{(v, T_{\ell t}(v))} = c_{(v, c_2)} \quad (\text{EC.17})$$

$$\xi_q - \xi_m \leq \left(\sum_{p \in \mathcal{P}_m^+} D_{ps} w_{mp}^+ - \sum_{p \in \mathcal{P}_m^-} D_{ps} w_{mp}^- \right) + C_\ell (1 - f_{mq}), \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{uv} \quad (\text{EC.18})$$

$$\xi_q - \xi_m \geq \left(\sum_{p \in \mathcal{P}_m^+} D_{ps} w_{mp}^+ - \sum_{p \in \mathcal{P}_m^-} D_{ps} w_{mp}^- \right) - C_\ell (1 - f_{mq}), \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{uv} \quad (\text{EC.19})$$

$$w_{mp}^+ \leq \sum_{q: (m, q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, \quad \forall p \in \mathcal{P}_m^+ \quad (\text{EC.20})$$

$$w_{mp}^- \leq \sum_{q: (m, q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, \quad \forall p \in \mathcal{P}_m^- \quad (\text{EC.21})$$

$$\sum_{m \in \mathcal{U}_{\ell st}^{uv}: p \in \mathcal{P}_m^+} w_{mp}^+ \leq 1 \quad \forall p \in \mathcal{P} \quad (\text{EC.22})$$

$$\sum_{m \in \mathcal{U}_{\ell st}^{uv}: p \in \mathcal{P}_m^-} w_{mp}^- \leq 1 \quad \forall p \in \mathcal{P} \quad (\text{EC.23})$$

$$\sum_{q: (m, q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} - \sum_{q: (q, m) \in \mathcal{H}_{\ell st}^{uv}} f_{qm} = \begin{cases} 1 & \text{if } m = (u, T_{\ell t}(u)), \\ -1 & \text{if } m = (v, T_{\ell t}(v)), \\ 0 & \text{otherwise.} \end{cases} \quad \forall m \in \mathcal{U}_{\ell st}^{uv} \quad (\text{EC.24})$$

$$f_{mq} \in \{0, 1\} \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{uv} \quad (\text{EC.25})$$

$$w_{mp}^+ \in \{0, 1\} \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, p \in \mathcal{P}_m^+ \quad (\text{EC.26})$$

$$w_{mp}^- \in \{0, 1\} \quad \forall m \in \mathcal{U}_{\ell st}^{uv}, p \in \mathcal{P}_m^- \quad (\text{EC.27})$$

Label setting algorithm. To distinguish pickups and dropoffs, we extend the label-setting algorithm from a two-dimensional to a three-dimensional state space. Dropoffs are treated the same way as pickups; for instance, the state transition includes checking all passenger combinations for pickups and all passenger combinations for dropoffs. This extension has two major implications that increase the computational requirements in the pricing problem. First, the dominance rule requires the dominating state to have the same set of pickups *and* the same set of dropoffs as the dominated state. Second, the set of load differential needs to be extended from $\{0, 1, \dots, C_\ell\}$ to $\{-C_\ell, \dots, -1, 0, 1, \dots, C_\ell\}$. Nonetheless, our results show that our methodology scales to meaningful practical instances of the MiND-DAR model in Manhattan, with up to 10 candidate lines, hundreds of candidate stops, thousands of passenger requests and 5 demand scenarios—resulting in over 60,000 first-stage variables and 700 second-stage problems.

EC.1.3. Experimental results

We construct a case study setting in Midtown Manhattan, with 10 candidate lines traveling West to East from the 11th to the 1st avenue along every other street between 36th and 54th. Each line contains a checkpoint at every other avenue, and each street-avenue intersection defines a station—leading to a total of 168 stations. We calibrate demand inputs by collecting all West-to-East requests in Midtown Manhattan during the morning rush from 6 to 9 am, amounting to over 3,000 passenger requests. We set up one-hour, two-hour and three-hour instances (from 6 to 7 am, 6 to 8 am, and 6 to 9 am, respectively). For each one, we run the deviated fixed-route microtransit as well as the fixed-line transit benchmark and ride-sharing benchmarks with single-occupancy, two-occupancy and four-occupancy vehicles (see EC.4.3). We consider five demand scenarios. Again, for apples-to-apples comparison, we group results by total seating capacity (e.g., 10 transit/microtransit vehicles of capacity 10, ride-sharing with 100/50/25 vehicles of capacity of 1/2/4), and perform an out-of-sample assessment corresponding to five new weekdays.

We evaluate the system-wide performance of all optimized transportation modes in Table EC.1, broken down into level of service (demand coverage and average walking time, waiting time, delay and detour), vehicle utilization (passengers served divided by vehicle capacity), and distance traveled (internal distance for served passengers plus external distance for unserved passengers). These results confirm and extend all takeaways from the MiND-VRP (Table 5 and Figure 7).

Note, first, the benefits of on-demand flexibility versus fixed-line transit: by leveraging on-demand deviations, microtransit enables significant increases in demand coverage. Specifically, microtransit serves 2 to 3 times more passengers; in the three-hour case for example, this increase translates into an improvement in vehicle utilization from 30% to 80% on average with medium system

Table EC.1 Average performance of fixed-route transit, microtransit, and ride-sharing in a dial-a-ride setting.

Setting			Average level of service				Vehicle utilization		Distance traveled (km)			
Horizon	Capacity	Mode	Coverage	Walk	Wait	Delay	Detour	Absolute	Relative	Internal	External	Total
1 hour	50	Transit	6.7%	1.68	3.57	1.45	200%	3.04	40.5%	60	881	941
		Microtransit	16.5%	1.92	3.71	1.70	144%	6.30	81.4%	94	798	893
		RS Cap. 4	23.8%	0.00	3.63	6.74	183%	3.81	95.3%	293	1,082	1,375
		RS Cap. 2	36.6%	0.00	3.56	4.30	120%	1.95	97.3%	567	942	1,509
		RS Cap. 1	53.3%	0.00	2.38	2.38	100%	1.00	100.0%	1,040	754	1,793
	100	Transit	7.6%	1.68	3.51	1.64	201%	3.18	22.7%	63	881	944
		Microtransit	22.8%	2.04	3.73	1.68	154%	8.29	57.4%	96	798	894
		RS Cap. 4	42.4%	0.00	3.48	6.92	190%	3.81	95.3%	541	833	1,374
		RS Cap. 2	63.3%	0.00	3.71	4.57	122%	1.94	97.2%	1,057	569	1,626
		RS Cap. 1	85.2%	0.00	2.47	2.47	100%	1.00	100.0%	1,876	268	2,144
2 hours	200	Transit	8.7%	1.64	3.54	2.07	199%	2.74	13.7%	84	906	990
		Microtransit	26.8%	2.05	3.80	1.65	156%	7.21	36.1%	127	755	882
		RS Cap. 4	70.7%	0.00	3.55	7.24	193%	3.86	96.6%	965	421	1,386
		RS Cap. 2	94.9%	0.00	4.01	5.05	125%	1.95	97.3%	1,708	76	1,784
		RS Cap. 1	100.0%	0.00	2.66	2.66	100%	1.00	100.0%	2,131	0	2,131
	50	Transit	6.9%	1.87	3.76	1.05	195%	3.13	52.1%	153	2,403	2,557
		Microtransit	15.2%	1.83	3.72	1.69	162%	6.62	101.5%	220	2,196	2,416
		RS Cap. 4	19.2%	0.00	3.88	6.77	182%	3.82	95.4%	592	3,039	3,631
		RS Cap. 2	29.7%	0.00	3.69	4.38	120%	1.96	98.1%	1,154	2,770	3,924
		RS Cap. 1	43.9%	0.00	2.46	2.46	100%	1.00	100.0%	2,184	2,337	4,521
	100	Transit	5.9%	0.94	3.77	1.32	196%	3.63	22.2%	120	2,428	2,548
		Microtransit	19.9%	2.12	3.95	1.65	155%	11.05	66.8%	175	2,071	2,246
		RS Cap. 4	33.9%	0.00	3.84	7.11	189%	3.84	96.0%	1,113	2,535	3,649
		RS Cap. 2	53.0%	0.00	3.78	4.61	123%	1.96	98.2%	2,252	1,948	4,200
		RS Cap. 1	74.3%	0.00	2.53	2.53	100%	1.00	100.0%	4,207	1,214	5,421
	200	Transit	8.5%	0.91	3.70	1.67	198%	3.23	16.2%	187	2,331	2,517
		Microtransit	27.3%	2.09	3.88	1.65	152%	9.42	47.1%	258	1,848	2,106
		RS Cap. 4	59.1%	0.00	3.80	7.41	194%	3.85	96.3%	2,114	1,611	3,724
		RS Cap. 2	86.1%	0.00	4.03	5.07	126%	1.96	97.9%	4,117	635	4,752
		RS Cap. 1	99.5%	0.00	2.76	2.76	100%	1.00	100.0%	6,253	30	6,283
3 hours	50	Transit	6.7%	1.88	3.77	1.13	198%	4.53	57.0%	192	3,895	4,087
		Microtransit	13.5%	1.74	3.69	1.74	154%	8.25	104.0%	268	3,600	3,868
		RS Cap. 4	16.8%	0.00	3.91	6.78	182%	3.80	95.0%	887	5,444	6,331
		RS Cap. 2	26.9%	0.00	3.80	4.47	120%	1.97	98.3%	1,750	4,996	6,746
		RS Cap. 1	39.7%	0.00	2.50	2.50	100%	1.00	100.0%	3,348	4,314	7,662
	100	Transit	7.6%	1.96	3.69	0.85	201%	5.09	30.5%	174	3,866	4,040
		Microtransit	20.7%	1.99	3.83	1.74	132%	14.01	80.0%	229	3,298	3,527
		RS Cap. 4	30.0%	0.00	3.87	7.09	189%	3.81	95.2%	1,691	4,670	6,361
		RS Cap. 2	48.2%	0.00	3.89	4.69	122%	1.96	98.0%	3,451	3,753	7,204
		RS Cap. 1	68.4%	0.00	2.56	2.56	100%	1.00	100.0%	6,558	2,520	9,077
	200	Transit	8.4%	1.85	3.83	1.93	195%	3.66	18.3%	262	3,678	3,940
		Microtransit	26.1%	2.13	3.99	1.69	127%	10.37	51.8%	365	2,991	3,356
		RS Cap. 4	53.0%	0.00	3.82	7.41	195%	3.84	95.9%	3,264	3,235	6,499
		RS Cap. 2	80.3%	0.00	4.03	5.05	126%	1.96	97.9%	6,549	1,561	8,110
		RS Cap. 1	97.8%	0.00	2.78	2.78	100%	1.00	100.0%	10,751	226	10,977

Walk, wait, delay and detour are averaged across all passengers. Walk, wait, and delay are in minutes.

capacity and from 18% to 52% with high system capacity. Unlike in the MiND-VRP, higher demand coverage comes with a slight increase in passenger walking and waiting, primarily due to an adverse selection effect—by serving passengers with pickup or drop-off locations further away from the reference lines, for example. Nonetheless, level of service remains comparable to fixed-line transit, with walking and waiting times around 2–3 minutes on average.

Next, results underscore the impact of demand consolidation: by relying on higher-capacity vehicles along reference lines, microtransit serves fewer passengers but travels much shorter dis-

tances than ride-sharing systems. As expected, ride-sharing results in higher demand coverage with no walking and short wait times. On the other hand, ride-sharing induces longer delays because of on-demand dispatches. Four-occupancy ride-pooling can also result in higher detours than microtransit, due to the negative externalities of door-to-door transportation—even with small-occupancy vehicles—and the comparative benefits of line regularization in microtransit. Moreover, the microtransit system travels much smaller (internal) distances by using higher-capacity vehicles.

At the aggregate level, microtransit induces strong system-wide improvements against all benchmarks. As compared to fixed-line transit, on-demand deviations increase distance traveled but this effect is more than compensated by the increase in demand coverage—leading to a decrease in distance per passenger by a factor of 1.4 to 2.3. As compared to ride-sharing, microtransit decreases distance traveled by a much higher factor than the corresponding loss in demand coverage, leading to a smaller distance per passenger by a factor of 4–11 (resp. 3–6) as compared to single-occupancy ride-sharing (resp. four-occupancy ride-pooling). When accounting for the “external” distance from single-occupancy trips for all unserved passengers (assuming for instance that all unserved passengers take a taxi to their destination), microtransit reduces total distance from fixed-line transit by 5%, 13% and 15% in the three-hour case with small, medium and high system capacity, respectively; it reduces total distance from four-occupancy ride-pooling by 39%, 45% and 48%; and it reduces total distance from single-occupancy ride-sharing by 98%, 157% and 227%.

These results confirm the potential of deviated fixed-route microtransit to improve demand coverage as compared to fixed-line transit—thanks to demand-responsive operations—and to improve demand consolidation as compared to ride-sharing—thanks to high-occupancy vehicles. These combined effects can induce strong reductions in distance traveled per passenger, which can ultimately contribute to creating more effective and more affordable mobility options and to mitigating the environmental footprint of urban mobility.

EC.2. Details on Model Formulations

EC.2.1. Notation Tables

Table EC.2 summarizes all notation for the MiND-VRP formulation.

EC.2.2. Segment-based Benchmark for Second-stage Problem

Throughout the section, we fix first-stage decisions \boldsymbol{x} and \boldsymbol{z} , as well as scenario $s \in \mathcal{S}$. The time horizon is discretized into $T_S + 1$ intervals in the set $\mathcal{T}^S = \{0, 1, \dots, T_S\}$, from the departure of the first trip ($t = 0$) to the arrival of the last trip ($t = T_S$).

To capture time and capacity constraints without relying on big- M constraints—therefore retaining a tight second-stage formulation—we build a time-load-expanded network $(\bar{\mathcal{V}}_{\ell st}, \bar{\mathcal{A}}_{\ell st})$. A dummy sink node $\bar{v}_{\ell st}$ represents the end of a trip. Each other node $n \in \bar{\mathcal{V}}_{\ell st}$ is associated with a

Component	Type	Description
\mathcal{N}	Set	Stations: checkpoints and pickup locations
\mathcal{E}	Set	Directed arcs in $\mathcal{N} \times \mathcal{N}$ corresponding to roadways
\mathcal{L}	Set	Candidate reference lines
\mathcal{P}	Set	Passenger types
\mathcal{S}	Set	Demand scenarios
\mathcal{C}_ℓ	Set	Vehicle loads on reference line $\ell \in \mathcal{L}$
\mathcal{I}_ℓ	Set	Checkpoints for line $\ell \in \mathcal{L}$, of cardinality I_ℓ
$\mathcal{I}_\ell^{(i)}$	Set	i^{th} stop in reference line $\ell \in \mathcal{L}$ for $i = 1, \dots, I_\ell$
Γ_ℓ	Set	Subset of checkpoint pairs in $\mathcal{I}_\ell \times \mathcal{I}_\ell$ for line $\ell \in \mathcal{L}$ that skip up to K checkpoints in between
\mathcal{N}_{uv}	Set	Subset of nodes in \mathcal{N} representing possible stations between checkpoints $u, v \in \mathcal{I}_\ell$ for each line $\ell \in \mathcal{L}$
\mathcal{T}_ℓ	Set	Allowable departure times of a vehicle from the beginning of line $\ell \in \mathcal{L}$
$\mathcal{T}_{\ell t}^{uv}$	Set	Time intervals between the scheduled times $T_{\ell t}(u)$ and $T_{\ell t}(v)$ for checkpoint pair $(u, v) \in \Gamma_\ell$
\mathcal{M}_p	Set	Compatible trips in $\mathcal{L} \times \mathcal{T}_\ell$ for passenger type $p \in \mathcal{P}$
$\mathcal{R}_{\ell st}$	Set	Subpaths corresponding to reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ in scenario $s \in \mathcal{S}$.
		Each subpath $r \in \mathcal{R}_{\ell st}$ starts in $u_r \in \mathcal{I}_\ell$ and ends in $v_r \in \mathcal{I}_\ell$.
$(\mathcal{V}_{\ell st}, \mathcal{A}_{\ell st})$	Graph	Load-expanded network of trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ in scenario $s \in \mathcal{S}$. Every trip starts at $u_{\ell st} \in \mathcal{V}_{\ell st}$ and ends at $v_{\ell st} \in \mathcal{V}_{\ell st}$
\mathcal{A}_r	Set	Arcs in $\mathcal{A}_{\ell st}$ corresponding to subpath $r \in \mathcal{R}_{\ell st}$ for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
$\mathcal{A}_{\ell st}^v$	Set	Arcs in $\mathcal{A}_{\ell st}$ connecting line destination to sink node for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
$(\mathcal{U}_{\ell st}^{uv}, \mathcal{H}_{\ell st}^{uv})$	Graph	Time-expanded network from $(u, T_{\ell t}(u))$ to $(v, T_{\ell t}(v))$. Node $m \in \mathcal{U}_{\ell st}^{uv}$ is characterized by a location-time tuple (k_m, t_m)
\mathcal{P}_m	Set	Passengers in \mathcal{P} that can be picked up in node $m \in \mathcal{U}_{\ell st}^{uv}$
\mathcal{P}_r	Set	Passenger types in \mathcal{P} picked up by subpath $r \in \mathcal{R}_{\ell st}$ for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
K	Parameter	Number of consecutive checkpoints that can be skipped (0 or 1)
C_ℓ	Parameter	Vehicle capacity on reference line $\ell \in \mathcal{L}$
F	Parameter	Fleet size
h_l	Parameter	Cost to operate one trip via line $\ell \in \mathcal{L}$
D_{ps}	Parameter	Number of passengers of type $p \in \mathcal{P}$ in scenario $s \in \mathcal{S}$
$T_{\ell t}(n)$	Parameter	Time at which trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ must visit stop $n \in \mathcal{I}_\ell$
π_s	Parameter	Probability of scenario $s \in \mathcal{S}$
g_a	Parameter	Cost of arc $a \in \mathcal{A}_{\ell st}$ for trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, scenario $s \in \mathcal{S}$ (Equation (7))
Δ	Parameter	Maximum vehicle deviation from reference line
Ω	Parameter	Maximum walking distance for passengers
Ψ	Parameter	Maximum waiting time for passengers
α	Parameter	Time window radius around passengers' requested drop-off times to build \mathcal{M}_p
$\omega_{o,d}$	Parameter	Walking distance between locations o and d
$\psi_{o,d}$	Parameter	Walking time between locations o and d
τ_{rp}^{walk}	Parameter	Walk time of passenger $p \in \mathcal{P}_r$ via subpath $r \in \mathcal{R}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
τ_{rp}^{wait}	Parameter	Wait time of passenger $p \in \mathcal{P}_r$ via subpath $r \in \mathcal{R}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
$\tau_{rp}^{\text{travel}}$	Parameter	In-vehicle time of passenger $p \in \mathcal{P}_r$ via subpath $r \in \mathcal{R}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}$
$\tau_{\ell t p}^{\text{late}}$	Parameter	Delay of passenger type $p \in \mathcal{P}$ when taking trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$
$\tau_{\ell t p}^{\text{early}}$	Parameter	Earliness of passenger type $p \in \mathcal{P}$ when taking trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$
τ_p^{dir}	Parameter	Direct travel time for passenger type $p \in \mathcal{P}$
$tt(e)$	Parameter	Travel time corresponding to road segment $e \in \mathcal{E}$
τ_{mp}^{walk}	Parameter	Walk time of passenger $p \in \mathcal{P}_m$ when picked up at node $m \in \mathcal{U}_{\ell st}^{uv}$
τ_{mp}^{wait}	Parameter	Wait time of passenger $p \in \mathcal{P}_m$ when picked up at node $m \in \mathcal{U}_{\ell st}^{uv}$
$\tau_{mp}^{\text{travel}}$	Parameter	In-vehicle travel time of passenger $p \in \mathcal{P}_m$ when picked up at node $m \in \mathcal{U}_{\ell st}^{uv}$
M	Parameter	Reward for each passenger pickup
$\lambda, \mu, \sigma, \delta$	Parameters	Penalties on passenger walk time, wait time, detour, and displacement
κ	Parameter	Target vehicle load in the first-stage network design problem

Table EC.2 Notation for the MiND-VRP model and its decomposition.

tuple (k_n, c_n, t_n) , so that node n represents a vehicle's arrival to station $k_n \in \mathcal{N}$ at time $t_n \in \mathcal{T}^S$ with $c_n \in \mathcal{C}$ passengers. The source node is denoted by $\bar{u}_{\ell st} := (\mathcal{I}_\ell^{(1)}, 0, T_{\ell t}(\mathcal{I}_\ell^{(1)}))$. We decompose

Component	Type	Description
$\bar{\mathcal{E}}_{\ell st}$	Set	Load-augmented road segments e associated with $road(e) \in \mathcal{E}$
\mathcal{T}^S	Set	Set of time periods during the planning horizon
\mathcal{P}_e	Set	Passengers picked up on segment $e \in \bar{\mathcal{E}}_{\ell st}$
$(\bar{\mathcal{V}}_{\ell st}, \bar{\mathcal{A}}_{\ell st})$	Graph	Time-load-expanded road network of trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$ in scenario $s \in \mathcal{S}$
$\bar{\mathcal{A}}_e$	Set	Arcs in $\bar{\mathcal{A}}_{\ell st}$ corresponding to segment $e \in \bar{\mathcal{E}}_{\ell st}$ for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$, $s \in \mathcal{S}$
$\bar{\mathcal{A}}_{\ell st}^{idle}$	Set	Arcs in $\bar{\mathcal{A}}_{\ell st}$ representing an idling vehicle
$\bar{\mathcal{A}}_{\ell st}^v$	Set	Arcs in $\bar{\mathcal{A}}_{\ell st}$ connecting the line's destination to the dummy sink node
τ_{ep}^{walk}	Parameter	Walk time of passenger $p \in \mathcal{P}_e$ via segment $e \in \bar{\mathcal{E}}_{\ell st}$, $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$, $s \in \mathcal{S}$
τ_{ep}^{wait}	Parameter	Wait time of passenger $p \in \mathcal{P}_e$ via segment $e \in \bar{\mathcal{E}}_{\ell st}$, $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$, $s \in \mathcal{S}$
τ_{ep}^{travel}	Parameter	In-vehicle time of passenger $p \in \mathcal{P}_e$ via segment $e \in \bar{\mathcal{E}}_{\ell st}$, $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$, $s \in \mathcal{S}$
\bar{g}_a	Parameter	Cost of arc $a \in \bar{\mathcal{A}}_{\ell st}$ on trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_{\ell}$ in scenario $s \in \mathcal{S}$

Table EC.3 Additional inputs of the segment-based formulation.

the arc set $\bar{\mathcal{A}}_{\ell st} \subset \bar{\mathcal{V}}_{\ell st} \times \bar{\mathcal{V}}_{\ell st}$ into traveling arcs, idling arcs, and terminating arcs, by writing $\bar{\mathcal{A}}_{\ell st} = \bigcup_{e \in \bar{\mathcal{E}}_{\ell st}} \bar{\mathcal{A}}_e \cup \bar{\mathcal{A}}_{\ell st}^{idle} \cup \bar{\mathcal{A}}_{\ell st}^v$.

To characterize traveling arcs, we denote by $\bar{\mathcal{E}}_{\ell st}$ the set of possible roadways and passenger pickups. Specifically, each segment $e \in \bar{\mathcal{E}}_{\ell st}$ is associated with a roadway $road(e) \in \mathcal{E}$ and a set of passengers \mathcal{P}_e who are picked up. We define traveling arcs by duplicating $e \in \bar{\mathcal{E}}_{\ell st}$ for all load pairs that correspond to the passenger pickups, and all time pairs that correspond to the travel time:

$$\bar{\mathcal{A}}_e = \left\{ (n, m) \in \bar{\mathcal{V}}_{\ell st} \times \bar{\mathcal{V}}_{\ell st} : (k_n, k_m) = road(e), \right. \\ c_m - c_n = \sum_{p \in \mathcal{P}_e} D_{ps}, \\ \left. t_m - t_n = tt(road(e)) \right\} \quad \forall e \in \bar{\mathcal{E}}_{\ell st} \quad (\text{EC.28})$$

Next, each idling arc in $\bar{\mathcal{A}}_{\ell st}^{idle}$ connects nodes corresponding to two consecutive time intervals at the same physical stop:

$$\bar{\mathcal{A}}_{\ell st}^{idle} = \{(n, m) \in \bar{\mathcal{V}}_{\ell st} \times \bar{\mathcal{V}}_{\ell st} : k_n = k_m, c_n = c_m, t_m - t_n = 1\}. \quad (\text{EC.29})$$

Finally, each terminating arc in $\bar{\mathcal{A}}_{\ell st}^v$ connects the line's destination to the dummy sink node:

$$\bar{\mathcal{A}}_{\ell st}^v = \{(n, m) \in \bar{\mathcal{V}}_{\ell st} \times \bar{\mathcal{V}}_{\ell st} : k_n = \mathcal{I}_{\ell}^{(I_{\ell})}, m = \bar{v}_{\ell st}^S\}. \quad (\text{EC.30})$$

Again, we can prune the time-load-expanded network by excluding disconnected nodes and all incident arcs. We define a segment-based cost \bar{g}_a for each $a \in \bar{\mathcal{A}}_{\ell st}$ analogously to Equation (7) to capture passenger walking times, waiting times, and relative arrival delays:

$$\bar{g}_a = \begin{cases} \sum_{p \in \mathcal{P}_e} D_{ps} \left(\lambda \tau_{ep}^{walk} + \mu \tau_{ep}^{wait} + \sigma \frac{\tau_{ep}^{travel}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{etp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{etp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right) & \text{if } e \in \bar{\mathcal{E}}_{\ell st}, a \in \bar{\mathcal{A}}_e \\ 0 & \text{if } a \in \bar{\mathcal{A}}_{\ell st}^{idle} \cup \bar{\mathcal{A}}_{\ell st}^v. \end{cases} \quad (\text{EC.31})$$

We define decision variables to select arcs in the time-load-expanded segment network:

$$\xi_a = \begin{cases} 1 & \text{if arc } a \text{ is selected, for } (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}, a \in \bar{\mathcal{A}}_{\ell st}, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{EC.32})$$

Recall that $\Gamma_\ell \subset \mathcal{I}_\ell \times \mathcal{I}_\ell$ denotes the set of checkpoint pairs with up to K skipped checkpoints:

$$\Gamma_\ell = \left\{ (\mathcal{I}_\ell^{(i)}, \mathcal{I}_\ell^{(j)}) \in \mathcal{I}_\ell \times \mathcal{I}_\ell : 1 \leq i < j \leq I_\ell, j - i \leq K + 1 \right\}, \quad \forall \ell \in \mathcal{L}$$

We define additional decision variables to select the set of checkpoint pairs that are visited:

$$\beta_{uv} = \begin{cases} 1 & \text{if checkpoints } (u, v) \in \Gamma_\ell \text{ are visited in sequence, and intermediate checkpoints are not visited,} \\ 0 & \text{otherwise.} \end{cases}$$

Recall that \mathcal{N}_{uv} denotes the set of stations that can be visited between checkpoints u and v , and $\mathcal{T}_{\ell t}^{uv}$ denotes the valid arrival times. We link the β_{uv} decisions with the ξ_a decisions, so that the vehicle route abides by the deviation limits imposed by the reference schedule. Altogether, the segment-based formulation exhibits a double flow structure—flow from checkpoint to checkpoint along the reference line, and flow from station to station between checkpoints—with linking constraints to ensure the consistency of these two sets of decisions.

The second-stage segment-based formulation is given as follows for scenario $s \in \mathcal{S}$.

$$\min \sum_{(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell} \sum_{a \in \bar{\mathcal{A}}_{\ell st}} \bar{g}_a \xi_a \quad (\text{EC.33})$$

$$\text{s.t. } \sum_{j:(i,j) \in \bar{\mathcal{A}}_{\ell st}} \xi_{(i,j)} - \sum_{j:(j,i) \in \bar{\mathcal{A}}_{\ell st}} \xi_{(j,i)} = \begin{cases} x_{lt} & \text{if } i = \bar{u}_{\ell st}, \\ -x_{lt} & \text{if } i = \bar{v}_{\ell st}, \\ 0 & \text{otherwise,} \end{cases} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall i \in \bar{\mathcal{V}}_{\ell st} \quad (\text{EC.34})$$

$$\sum_{e \in \bar{\mathcal{E}}_{\ell st}} \sum_{a \in \bar{\mathcal{A}}_e : p \in \mathcal{P}_e} \xi_a \leq z_{plt} \quad \forall p \in \mathcal{P}, \forall (\ell, t) \in \mathcal{M}_p \quad (\text{EC.35})$$

$$\sum_{v:(u,v) \in \Gamma_\ell} \beta_{uv} - \sum_{v:(v,u) \in \Gamma_\ell} \beta_{vu} = \begin{cases} x_{\ell t} & \text{if } u = \mathcal{I}_\ell^{(1)} \\ -x_{\ell t} & \text{if } u = \mathcal{I}_\ell^{(I_\ell)}, \\ 0 & \text{otherwise} \end{cases}, \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall u \in \mathcal{I}_\ell \quad (\text{EC.36})$$

$$\sum_{\substack{(i,j) \in \bar{\mathcal{A}}_{\ell st}: \\ k_j=v, t_j=T_{\ell t}(v)}} \xi_{(i,j)} \geq \sum_{w \in \mathcal{I}_\ell : (w,v) \in \Gamma_\ell} \beta_{wv} \quad \forall v \in \mathcal{I}_\ell \setminus \mathcal{I}_\ell^{(1)} \quad (\text{EC.37})$$

$$\xi_{(n,m)} \leq \sum_{\substack{(u,v) \in \Gamma_\ell: \\ k_n, k_m \in \mathcal{N}_{uv}^{(1)}, \\ t_n, t_m \in \mathcal{T}_{\ell t}^{uv}}} \beta_{uv} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall (n, m) \in \bar{\mathcal{A}}_{\ell st} \quad (\text{EC.38})$$

$$\xi_a \in \{0, 1\} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, a \in \bar{\mathcal{A}}_{\ell st} \quad (\text{EC.39})$$

$$\beta_{uv} \in \{0, 1\} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall (u, v) \in \Gamma_\ell \quad (\text{EC.40})$$

Equations (EC.33)–(EC.35) are analogous to Equations (8)–(10). Constraint (EC.36) ensures that the vehicle does not skip more than K stops in a row by selecting checkpoint pairs that form a

valid path along the reference line. Equations (EC.37) and (EC.38) serve as the linking constraints, ensuring that selected checkpoints are visited at the time specified by the reference schedule, and that the vehicle visits any intermediate locations with the correct chronology. In other words, we can only select a segment if (i) its endpoints correspond to stations in \mathcal{N}_{uv} between selected checkpoints, and (ii) its visit times fall within the reference schedule window defined by $T_{\ell t}(u)$ and $T_{\ell t}(v)$. Constraints (EC.39)–(EC.40) apply the binary requirements to the decision variables.

EC.2.3. Path-based Integer Optimization Formulation for Second-Stage Deviations

Component	Type	Description
$\mathcal{Q}_{\ell st}$	Set	Valid paths for reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$
\mathcal{P}_q	Set	Passenger pickup set corresponding to each path $q \in \mathcal{Q}_{\ell st}$
τ_{qp}^{walk}	Parameter	Walk time of passenger $p \in \mathcal{P}_r$ via path $q \in \mathcal{Q}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, $s \in \mathcal{S}$
τ_{qp}^{wait}	Parameter	Wait time of passenger $p \in \mathcal{P}_r$ via path $q \in \mathcal{Q}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, $s \in \mathcal{S}$
$\tau_{qp}^{\text{travel}}$	Parameter	In-vehicle time of passenger $p \in \mathcal{P}_r$ via path $q \in \mathcal{Q}_{\ell st}$, for $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$, $s \in \mathcal{S}$
g_q^Q	Parameter	Cost of path $q \in \mathcal{Q}_{\ell st}$ on trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ in scenario $s \in \mathcal{S}$

Table EC.4 Additional inputs of the path-based formulation.

Throughout the section, we fix first-stage decisions \mathbf{x} and \mathbf{z} , as well as scenario $s \in \mathcal{S}$.

Let $\mathcal{Q}_{\ell st}$ denote the set of all valid paths to reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and scenario $s \in \mathcal{S}$. Each path $q \in \mathcal{Q}_{\ell st}$ corresponds to a sequence of road segments that starts at the beginning of the line, end at its destination, satisfies flow balance in between, skips at most K checkpoints in a row, does not pick up more than C_ℓ passengers, and satisfies the reference schedule at the checkpoints. For each $q \in \mathcal{Q}_{\ell st}$, we store the passenger pickups in $\mathcal{P}_q \subset \mathcal{P}$. By definition, $\sum_{p \in \mathcal{P}_q} D_{ps} \leq C_\ell$. The cost g_q^Q of each path is defined analogously to Equation (7) to capture passenger level of service:

$$g_q^Q = \sum_{p \in \mathcal{P}_q} D_{ps} \left(\lambda \tau_{qp}^{\text{walk}} + \mu \tau_{qp}^{\text{wait}} + \sigma \frac{\tau_{qp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{\ell tp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{\ell tp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right), \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, q \in \mathcal{Q}_{\ell st}. \quad (\text{EC.41})$$

We define the following decision variables:

$$\zeta_q = \begin{cases} 1 & \text{if path } q \text{ is selected, for } (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, s \in \mathcal{S}, q \in \mathcal{Q}_{\ell st}, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{EC.42})$$

The path-based formulation is given as follows for scenario $s \in \mathcal{S}$.

$$\min \sum_{(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell} \sum_{q \in \mathcal{Q}_{\ell st}} g_q^Q \zeta_q \quad (\text{EC.43})$$

$$\text{s.t. } \sum_{q \in \mathcal{Q}_{\ell st}} \zeta_q = x_{\ell t} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell \quad (\text{EC.44})$$

$$\sum_{q \in \mathcal{Q}_{\ell st} : p \in \mathcal{P}_q} \zeta_q \leq z_{plt} \quad \forall p \in \mathcal{P}, \forall (\ell, t) \in \mathcal{M}_p \quad (\text{EC.45})$$

$$\zeta_q \in \{0, 1\} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, q \in \mathcal{Q}_{\ell st} \quad (\text{EC.46})$$

Equations (EC.43) is analogous to Equation (8). Constraints (EC.44) ensure that exactly one path is selected for each selected reference trip. Constraints (EC.45) ensure that selected paths only serve passengers that have been assigned to that trip, analogously to Equation (10).

EC.2.4. Proof of Proposition 1

Throughout this proof, we fix the first-stage decisions \mathbf{x} , \mathbf{z} . We consider a fixed reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ as well as a fixed scenario $s \in \mathcal{S}$.

Equivalence of the path-based and subpath-based formulations.

Constructing a load-expanded subpath solution from a path solution. Let us consider a feasible solution $\hat{\zeta}$ to the path-based formulation (Equations (EC.43)–(EC.46)) and build a feasible solution to the subpath-based formulation with the same objective value.

Assume that $x_{\ell t} = 1$, and let $q \in \mathcal{Q}_{\ell st}$ be the selected path with $\hat{\zeta}_q = 1$ (which exists by Equation (EC.44)). By definition, the path corresponds to a sequence of road segments that starts at the beginning of line ℓ , ends at its destination, picks up at most C_ℓ passengers, visits checkpoints without skipping more than K in a row, and arrives at each checkpoint at the scheduled times. With a slight abuse of notation, let $\mathcal{I}_\ell^q := \{\nu_1, \dots, \nu_Q\} \subseteq \mathcal{N}$ identify the ordered set of Q checkpoints visited by path q . Similarly, we decompose path q into an ordered sequence of $Q - 1$ subpaths $\mathcal{R}_q := \{r_1, \dots, r_{Q-1}\}$. The subpaths in \mathcal{R}_q partition the served passengers \mathcal{P}_q on path q , so that $\mathcal{P}_q = \bigcup_{r \in \mathcal{R}_q} \mathcal{P}_r$. Each subpath $r_i \in \mathcal{R}_q$ induces a unique arc $a_i := (n, m) \in \mathcal{A}_{\ell st}$ in the load-expanded network, such that (i) the arc corresponds to the subpath: $r(a_i) = r_i$; (ii) the loads are consistent with pickups: $c_n = 0$ if $i = 1$, and $c_n = \sum_{j=1}^{i-1} |\mathcal{P}_{r_j}|$ otherwise, and $c_m = \sum_{j=1}^i |\mathcal{P}_{r_j}|$. Let us collect these load-expanded subpath arcs into the set $\mathcal{A}_q := \{a_1, \dots, a_{Q-1}\} \subset \mathcal{A}_{\ell st}$.

We can construct a feasible solution to the load-expanded subpath formulation.

$$\hat{y}_a = \begin{cases} 1 & \text{if } a \in \bigcup_{q \in \mathcal{Q}_{\ell st}: \hat{\zeta}_q=1} \mathcal{A}_q, \\ 0 & \text{otherwise,} \end{cases} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall a \in \mathcal{A}_{\ell st}.$$

This solution satisfies the flow balance constraints in Equation (9):

- If $x_{\ell t} = 0$, no path in $\mathcal{Q}_{\ell st}$ is selected, so no arc in $\mathcal{A}_{\ell st}$ is selected either. Therefore,

$$\sum_{m:(n,m) \in \mathcal{A}_{\ell st}} \hat{y}_{(n,m)} - \sum_{m:(m,n) \in \mathcal{A}_{\ell st}} \hat{y}_{(m,n)} = 0 = \begin{cases} x_{\ell t} & \text{if } n = u_{\ell st}, \\ -x_{\ell t} & \text{if } n = v_{\ell st}, \\ 0 & \text{otherwise.} \end{cases}$$

- If $x_{\ell t} = 1$, we have, for each node $n \in \mathcal{V}_{\ell st}$:

$$\sum_{m:(n,m) \in \mathcal{A}_{\ell st}} \hat{y}_{(n,m)} - \sum_{m:(m,n) \in \mathcal{A}_{\ell st}} \hat{y}_{(m,n)} = \begin{cases} \hat{y}_{a_1} - 0 = 1 & \text{if } n = u_{\ell st}, \\ 0 - \hat{y}_{a_{Q-1}} = -1 & \text{if } n = v_{\ell st}, \\ \hat{y}_{a_i} - \hat{y}_{a_{i-1}} = 1 - 1 = 0 & \text{if } k_n = \nu_i \in \mathcal{I}_\ell^q \setminus \{\nu_1, \nu_Q\}, \\ c_n = \sum_{j=1}^{i-1} |\mathcal{P}_{r_j}|, & \\ 0 - 0 = 0 & \text{otherwise.} \end{cases}$$

The solution also satisfies the passenger linking constraints in Equations (10).

- Consider a passenger request $p \in \mathcal{P} \setminus \bigcup_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \mathcal{P}_q$. The set of pickups on the selected paths $\bigcup_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \mathcal{P}_q$ induces the set of pickups on the selected subpaths, so that $p \in \mathcal{P} \setminus \bigcup_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \bigcup_{r \in \mathcal{R}_q} \mathcal{P}_r$. Thus, $\hat{y}_a = 0$ for each arc $a \in \mathcal{A}_{\ell st}$ with $p \in \mathcal{P}_{r(a)}$, and

$$\sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}} \hat{y}_a = 0 \leq z_{\ell pst}.$$

- Consider passenger request $p \in \mathcal{P}$ served by some path $q' \in \mathcal{Q}_{\ell st}$, so that $\hat{\zeta}_{q'} = 1$ and $p \in \mathcal{P}_{q'}$. Each pickup set $\mathcal{P}_{q'}$ has been partitioned into pickup subsets at the subpath level, so there exists $r' \in \mathcal{R}_{q'}$ with $p \in \mathcal{P}_{r'}$. This subpath has been mapped to a unique arc $a_{r'} \in \mathcal{A}_{q'}$, so that

$$\sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}} \hat{y}_a = \hat{y}_{a_{r'}} = 1 = \hat{\zeta}_{q'} = \sum_{q \in \mathcal{Q}_{\ell st} : p \in \mathcal{P}_{q'}} \hat{\zeta}_q \leq z_{\ell pst}$$

where the first three equalities come from the construction of paths and subpaths, the fourth equality stems from the fact that passenger p can be picked up by one subpath, and the final one stems from Equation (EC.45).

Therefore, the solution $\hat{\mathbf{y}}$ is feasible in the subpath-based formulation. We now show that it achieves the same objective value as $\hat{\zeta}$:

$$\begin{aligned} \sum_{a \in \mathcal{A}_{\ell st}} g_a \hat{y}_a &= \sum_{a \in \mathcal{A}_{\ell st} : \hat{y}_a=1} g_a \\ &= \sum_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \sum_{a \in \mathcal{A}_q} g_a \\ &= \sum_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \sum_{r_i \in \mathcal{R}_q} g_{a_i} \\ &= \sum_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \sum_{r \in \mathcal{R}_q} \sum_{p \in \mathcal{P}_r} D_{ps} \left(\lambda \tau_{rp}^{\text{walk}} + \mu \tau_{rp}^{\text{wait}} + \sigma \frac{\tau_{rp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{ltp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{ltp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right) \\ &= \sum_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} \sum_{p \in \mathcal{P}_q} D_{ps} \left(\lambda \tau_{qp}^{\text{walk}} + \mu \tau_{qp}^{\text{wait}} + \sigma \frac{\tau_{qp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{ltp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{ltp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right) \\ &= \sum_{q \in \mathcal{Q}_{\ell st} : \hat{\zeta}_q=1} g_q^Q \\ &= \sum_{q \in \mathcal{Q}_{\ell st}} g_q^Q \hat{\zeta}_q \end{aligned} \tag{EC.47}$$

The first two equalities come from the construction of paths and subpaths; the third equality leverages the uniqueness of the load-expanded subpath arc induced by the subpath sequence; the fourth equality is due to the definition of a load-expanded subpath arc cost; the fifth is due to the partition of $\mathcal{P}_q = \bigcup_{r \in \mathcal{R}_q} \mathcal{P}_r$; and the last two equalities stem from the definition of path costs g_q^Q .

In conclusion, any path solution can be mapped into a feasible subpath solution with the same objective value. Therefore, the subpath-based formulation achieves an objective that is at most equal to an optimum of the path-based formulation.

Constructing a path solution from a load-expanded subpath solution. Let us consider a feasible solution $\hat{\mathbf{y}}$ to the subpath-based formulation (Equations (9)–(11)), and build a feasible solution $\hat{\zeta}$ to the path-based formulation (Equations (EC.43)–(EC.46)) with the same objective value.

Assume that $x_{\ell t} = 1$. We leverage Equation (9) to construct a path from $u_{\ell st}$ to $v_{\ell st}$ in the load-expanded subpath network $(\mathcal{V}_{\ell st}, \mathcal{A}_{\ell st})$. Beginning from the source, we select the unique arc $a_1 \in \mathcal{A}_{\ell st}$ incident with $u_{\ell st}$ for which $\hat{y}_{a_1} = 1$, proceeding sequentially along the directed network until $v_{\ell st}$ is reached and $Q - 1$ arcs are retrieved. A unique outgoing arc is guaranteed at every intermediate node by Equation (9).

Each arc $a_i \in \mathcal{A}_q := \{a_1, \dots, a_{Q-1}\}$ corresponds to a subpath $r_i := r(a_i) \in \mathcal{R}_{\ell st}$ and a passenger pickup set \mathcal{P}_{r_i} . The sequence of subpaths $\mathcal{R}_q := \{r_1, \dots, r_{Q-1}\}$ defines a path q from $u_{\ell st}$ to $v_{\ell st}$ (by Equation (9)), skipping at most K checkpoints in a row (by definition of the subpaths $r_i \in \mathcal{R}_{\ell st}$), cohering with the scheduled arrival times associated with trip (ℓ, t) (again by definition of r_i), obeying the vehicle's capacity (by definition of the node set in the load-expanded network $\mathcal{V}_{\ell st}$), and picking up the passengers in $\mathcal{P}_q := \bigcup_{i=1}^{Q-1} \mathcal{P}_{r_i}$ (which is unique due to Equations (10)). Thus, $\hat{\mathbf{y}}$ defines a unique and valid path in $\mathcal{Q}_{\ell st}$ for reference trip (ℓ, t) if $x_{\ell t} = 1$.

Let us collect all such paths in the set $\mathcal{Q}(\hat{\mathbf{y}})$. We construct solution $\hat{\zeta}$ from $\hat{\mathbf{y}}$:

$$\hat{\zeta}_q = \begin{cases} 1 & \text{if } q \in \mathcal{Q}(\hat{\mathbf{y}}), \\ 0 & \text{otherwise,} \end{cases} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall q \in \mathcal{Q}_{\ell st}.$$

By construction, the solution satisfies Equations (EC.44). If $x_{\ell t} = 1$, we constructed a single path based on the subpath solution. If $x_{\ell t} = 0$, there was no path to construct, as no arcs were selected from $u_{\ell st}$ to $v_{\ell st}$ by Equation (9). Therefore:

$$\sum_{q \in \mathcal{Q}_{\ell st}} \hat{\zeta}_q = \sum_{q \in \mathcal{Q}_{\ell st}} \mathbb{1}(q \in \mathcal{Q}(\hat{\mathbf{y}})) = x_{\ell t}, \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell.$$

The solution also satisfies Equation (EC.45). For each reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and some passenger $p \in \mathcal{P}$, we obtain, from the construction of the path solution and Equation (10):

$$\begin{aligned} \sum_{q \in \mathcal{Q}_{\ell st}: p \in \mathcal{P}_q} \hat{\zeta}_q &= \sum_{q \in \mathcal{Q}(\hat{\mathbf{y}})} \mathbb{1}(p \in \mathcal{P}_q) \\ &= \sum_{q \in \mathcal{Q}(\hat{\mathbf{y}})} \sum_{r \in \mathcal{R}_q} \mathbb{1}(p \in \mathcal{P}_r) \\ &= \sum_{q \in \mathcal{Q}(\hat{\mathbf{y}})} \sum_{a \in \mathcal{A}_q} \mathbb{1}(p \in \mathcal{P}_{r(a)}) \end{aligned}$$

$$\begin{aligned}
&= \sum_{a \in \mathcal{A}_{\ell st} : \hat{y}_a = 1} \mathbb{1}(p \in \mathcal{P}_{r(a)}) \\
&= \sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}} \hat{y}_a \\
&\leq z_{\ell pst}
\end{aligned}$$

Finally, the solutions $\hat{\zeta}$ and $\hat{\mathbf{y}}$ achieve the same objective values, which can be shown similarly to Equations (EC.47). Therefore, any subpath solution can be mapped into a feasible path solution with the same objective value, and the path-based formulation achieves an objective that is at most equal to the optimum of the subpath-based formulation. This concludes the proof of equivalence of the path-based and subpath-based formulations.

Equivalence of path-based and subpath-based relaxations. The arguments employed in this proof do not require the integrality of the path solution $\hat{\zeta}$ and of the subpath solution $\hat{\mathbf{y}}$. By following the same steps as above, we can map any non-integral path solution $\hat{\zeta}$ into a feasible subpath solution with the same objective value, as follows:

$$\hat{y}_a = \sum_{q \in \mathcal{Q}_{\ell st} : a \in \mathcal{A}_q} \hat{\zeta}_q \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall a \in \mathcal{A}_{\ell st}.$$

Similarly, we can map any non-integral subpath solution $\hat{\mathbf{y}}$ into a feasible path solution with the same objective value. Alternatively, we can observe that the path-based formulation is a Dantzig-Wolfe reformulation of the subpath-based formulation where Equations (9) are convexified into Equations (EC.44). Since Equations (9) already form an integral polyhedron, both formulations contain the same convex hull. This proves that the path-based and subpath-based formulations define the same linear relaxations.

Equivalence of the segment-based and subpath-based formulations.

Constructing a time-load-expanded segment solution from a load-expanded subpath solution. Let us consider a feasible solution $\hat{\mathbf{y}}$ to the subpath-based formulation (Equations (8)–(11)) and build a feasible solution to the segment-based formulation with the same objective value.

Assume that $x_{\ell t} = 1$, and let $a \in \mathcal{A}_{\ell st}$ be a selected subpath-based arc with $\hat{y}_a = 1$ (which exists by Equation (9)). By definition, the subpath-based arc corresponds to the load expansion of a subpath that traverses a sequence of road segments starting at checkpoint $u \in \mathcal{I}_\ell$ at time $T_{\ell t}(u)$, ending at checkpoint $v \in \mathcal{I}_\ell$ at time $T_{\ell t}(v)$, skipping up to K checkpoints in-between (i.e., $(u, v) \in \Gamma_\ell$), carrying $c_{start(a)}$ passengers in u , and carrying $c_{end(a)}$ passengers in v . Let us store the stations visited by subpath r in an ordered set $\mathcal{N}_r := \{\nu_1, \dots, \nu_N\} \subseteq \mathcal{N}$, where $\nu_1 = u$, $\nu_N = v$, and ν_2, \dots, ν_{N-1} denote intermediate stations. Similarly, we decompose subpath r into a sequence of $N - 1$ segments $\mathcal{E}_r := \{e_1, \dots, e_{N-1}\}$, where segment e_i connects stations ν_i and ν_{i+1} with travel time tt_{e_i} (potentially with idling time). The segments in \mathcal{E}_r partition the passengers in \mathcal{P}_r : $\mathcal{P}_r = \bigcup_{e \in \mathcal{E}_r} \mathcal{P}_e$.

To obtain the corresponding segment solution, we need to specify an appropriate time discretization. Due to the adherence to the reference schedule, the discretization in the segment-based formulation does not introduce errors as long as all viable subpaths are feasible in that formulation. We show that there exists a discrete time unit for which this is the case, in the following lemma.

LEMMA EC.1. *Assume that the elapsed time between the scheduled arrival times at the checkpoints along the reference line are strictly larger than the travel times of the corresponding subpaths. Then, there exists a discrete time unit such that, in the corresponding time-expanded network, all feasible subpaths have an estimated travel time that is less than the elapsed time between the corresponding checkpoints' scheduled arrival times.*

Let $\Delta_{uv} := T_{\ell t}(v) - T_{\ell t}(u)$ denote the travel time between checkpoints u and v , determined by the scheduled arrival times at both checkpoints, and let us denote the travel time of subpath r by $\Delta_r := \sum_{e \in \mathcal{E}_r} tt_e$. Due to the maximum deviation from the reference line, the number of passenger pickups, and the upper bound on passengers' walking distance, the set of potential subpaths $\mathcal{R}_{\ell st}^{uv}$ between checkpoints u and v is finite. For convenience, let us denote this subset by

$$\mathcal{R}_{\ell st}^{uv} := \{r \in \mathcal{R}_{\ell st} : u_r = u, v_r = v\}.$$

By assumption, all subpaths $r \in \mathcal{R}_{\ell st}$ satisfy $\Delta_r \leq T_{\ell t}(v_r) - T_{\ell t}(u_r)$, so that $\Delta_r < \Delta_{uv}$ for each $r \in \mathcal{R}_{\ell st}^{uv}$. We define the discrete time unit between checkpoints u and v as:

$$\rho_{uv} = \min_{r \in \mathcal{R}_{\ell st}^{uv}} \frac{\Delta_{uv} - \Delta_r}{|\mathcal{E}_r|} > 0. \quad (\text{EC.48})$$

Without loss of generality, we assume that ρ_{uv} is rational; otherwise, we can define it as the largest rational number bounded from above by the minimum given in Equation (EC.48). We define the universal discrete time unit as

$$\rho = \text{GCD}(\{\rho_{uv} : (u, v) \in \Gamma_\ell\}), \quad (\text{EC.49})$$

where GCD denotes the greatest common divisor. By construction, for each $(u, v) \in \Gamma_\ell$, there exists $R_{uv} \in \mathbb{Z}_+$ such that $\rho_{uv} = R_{uv}\rho \geq \rho$.

In the segment-based formulation, travel times are rounded up to the nearest discrete time step on each segment. The estimated travel time on each segment $e \in \mathcal{E}_r$, denoted by $\bar{\Delta}_e$, is therefore

$$\bar{\Delta}_e = \left\lceil \frac{tt_e}{\rho} \right\rceil \cdot \rho.$$

The travel time estimate of subpath $r \in \mathcal{R}_{\ell st}^{uv}$ in the segment-based formulation, denoted by $\bar{\Delta}_r$, is then given by:

$$\bar{\Delta}_r = \sum_{e \in \mathcal{E}_r} \bar{\Delta}_e = \sum_{e \in \mathcal{E}_r} \left\lceil \frac{tt_e}{\rho} \right\rceil \rho.$$

We make use of the following property:

$$\left\lceil \frac{tt_e}{\rho} \right\rceil \cdot \rho \leq \left\lceil \frac{tt_e}{\rho_{uv}} \right\rceil \cdot \rho_{uv} \leq tt_e + \rho_{uv}.$$

The first inequality stems from the fact that $\lceil a/R_{uv} \rceil \leq \lceil a \rceil / R_{uv}$ for any $a > 0$. The second inequality follows from the definition of the ceiling function. Thus, we obtain:

$$\bar{\Delta}_r \leq \sum_{e \in \mathcal{E}_r} (tt_e + \rho_{uv}) = \Delta_r + \sum_{e \in \mathcal{E}_r} \rho_{uv} \leq \Delta_r + \sum_{e \in \mathcal{E}_r} \left(\frac{\Delta_{uv} - \Delta_r}{|\mathcal{E}_r|} \right) = \Delta_{uv}, \quad \forall r \in \mathcal{R}_{\ell st}^{uv}. \quad (\text{EC.50})$$

This completes the proof of the lemma. \square

Lemma EC.1 shows that there exists a discrete time unit for which all feasible subpaths in the subpath-based formulation are also feasible in the segment-based formulation. With this discretization, each segment $e_i \in \mathcal{E}_r$ induces a unique arc $\bar{a}_i := (n, m) \in \bar{\mathcal{A}}_{\ell st}$ in the time-load-expanded network, such that: (i) the arc corresponds to the segment: $e(\bar{a}_i) = e_i$; (ii) the capacities are consistent with pickups: $c_n = c_{start(a)}$ if $i = 1$, and $c_n = c_{start(a)} + \sum_{j=1}^{i-1} |\mathcal{P}_{e_j}|$ otherwise, and $c_m = c_{start(a)} + \sum_{j=1}^i |\mathcal{P}_{e_j}|$; and (iii) the time is consistent with travel times: $t_n = T_{\ell t}(u)$ if $i = 1$, and $t_n = T_{\ell t}(u) + \rho_{uv} \cdot \sum_{j=1}^{i-1} \lceil \frac{tt_{e_j}}{\rho_{uv}} \rceil$ otherwise, and $t_m = T_{\ell t}(u) + \rho_{uv} \cdot \sum_{j=1}^i \lceil \frac{tt_{e_j}}{\rho_{uv}} \rceil$. Let us collect these time-load-expanded segment arcs into the set $\bar{\mathcal{A}}_a := \{\bar{a}_1, \dots, \bar{a}_{N-1}\} \subset \bar{\mathcal{A}}_{\ell st}$.

With these arcs, we construct a feasible solution to the time-load-expanded segment formulation.

$$\begin{aligned} \hat{\xi}_{\bar{a}} &= \begin{cases} 1 & \text{if } \bar{a} \in \bigcup_{a \in \mathcal{A}_{\ell st}: \hat{y}_a=1} \bar{\mathcal{A}}_a, \\ 0 & \text{otherwise,} \end{cases} & \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall \bar{a} \in \bar{\mathcal{A}}_{\ell st} \\ \hat{\beta}_{uv} &= \begin{cases} 1 & \text{if there exists } a \in \mathcal{A}_{\ell st} : \hat{y}_a = 1, u_{r(a)} = u, v_{r(a)} = v, \\ 0 & \text{otherwise,} \end{cases} & \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall (u, v) \in \Gamma_\ell. \end{aligned}$$

This solution satisfies the flow balance constraints in Equation (EC.34).

– If $x_{\ell t} = 0$, then no subpath arc in $\mathcal{A}_{\ell st}$ is selected, so no arc in $\bar{\mathcal{A}}_{\ell st}$ is selected either. Therefore

$$\sum_{j:(i,j) \in \bar{\mathcal{A}}_{\ell st}} \hat{\xi}_{(i,j)} - \sum_{j:(j,i) \in \bar{\mathcal{A}}_{\ell st}} \hat{\xi}_{(j,i)} = 0 = \begin{cases} x_{\ell t} & \text{if } i = \bar{u}_{\ell st}, \\ -x_{\ell t} & \text{if } i = \bar{v}_{\ell st}, \\ 0 & \text{otherwise.} \end{cases}$$

– Suppose that $x_{\ell t} = 1$. Recall the sequence of subpath arcs $a \in \mathcal{A}_{\ell st}$ such that $\hat{y}_a = 1$, which exist by Equation (9). We collect the corresponding segment arcs from $\bar{\mathcal{A}}_a$ in order into set $\bar{\mathcal{A}}_{\text{all}} = \bigcup_{a \in \mathcal{A}_{\ell st}: \hat{y}_a=1} \bar{\mathcal{A}}_a := \{\bar{a}_1, \dots, \bar{a}_{M-1}\}$ and corresponding time-load-expanded nodes $\bar{\mathcal{V}}^{\text{all}} = \{n_1, \dots, n_M\}$, where $n_1 = \bar{u}_{\ell st}$ and $n_M = \bar{v}_{\ell st}$ and n_2, \dots, n_{M-1} refer to intermediate nodes. We have, for each node $n \in \bar{\mathcal{V}}_{\ell st}$:

$$\sum_{m:(n,m) \in \bar{\mathcal{A}}_{\ell st}} \hat{\xi}_{(n,m)} - \sum_{m:(m,n) \in \bar{\mathcal{A}}_{\ell st}} \hat{\xi}_{(m,n)} = \begin{cases} \hat{\xi}_{\bar{a}_1} - 0 = 1 = x_{\ell t} & \text{if } n = \bar{u}_{\ell st}, \\ 0 - \hat{\xi}_{\bar{a}_M} = -1 = -x_{\ell t} & \text{if } n = \bar{v}_{\ell st}, \\ \hat{\xi}_{\bar{a}_i} - \hat{\xi}_{\bar{a}_{i-1}} = 0 & \text{if } n = n_i \in \bar{\mathcal{V}}^{\text{all}} \setminus \{n_1, n_M\}, \\ 0 - 0 = 0 & \text{otherwise.} \end{cases}$$

The solution also satisfies the passenger linking constraints in Equations (EC.35).

- Consider a passenger request $p \in \mathcal{P} \setminus \bigcup_{a \in \mathcal{A}_{\ell st} : \hat{y}_a=1} \mathcal{P}_{r(a)}$. The set of pickups on the selected subpaths $\bigcup_{a \in \mathcal{A}_{\ell st} : \hat{y}_a=1} \mathcal{P}_a$ induces the set of pickups on the selected segments, so that $p \in \mathcal{P} \setminus \bigcup_{a \in \mathcal{A}_{\ell st} : \hat{y}_a=1} \bigcup_{e \in \mathcal{E}_{r(a)}} \mathcal{P}_e$. Thus, $\hat{\xi}_{\bar{a}} = 0$ for each arc $\bar{a} \in \bar{\mathcal{A}}_{\ell st}$ with $p \in \mathcal{P}_{e(\bar{a})}$, and

$$\sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st} : p \in \mathcal{P}_{e(\bar{a})}} \hat{\xi}_{\bar{a}} = 0 \leq z_{\ell pst}.$$

- Consider passenger request $p \in \mathcal{P}$ served by some subpath arc $a' \in \mathcal{A}_{\ell st}$, so that $\hat{y}_{a'} = 1$ and $p \in \mathcal{P}_{r(a')}$. Each pickup set $\mathcal{P}_{r(a')}$ has been partitioned into pickup subsets at the segment level, so that there exists some $e \in \mathcal{E}_{r(a')}$ with $p \in \mathcal{P}_e$. This segment has been mapped to a unique arc $\bar{a}' \in \mathcal{A}_{a'}$. Using Equation (10), we obtain:

$$\sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st} : p \in \mathcal{P}_{e(\bar{a})}} \hat{\xi}_{\bar{a}} = \hat{\xi}_{\bar{a}'} = 1 = \hat{y}_{a'} = \sum_{a \in \mathcal{A}_{\ell st} : p \in \mathcal{P}_{r(a)}} \hat{y}_a \leq z_{\ell pst}.$$

Next, the solution satisfies the flow balance between checkpoints in Equations (EC.36).

- If $x_{\ell t} = 0$, then no subpath arcs in $\mathcal{A}_{\ell st}$ are selected, so $\hat{\beta}_{uv} = 0$ for all $(u, v) \in \Gamma_\ell$. Therefore, for each checkpoint $u \in \mathcal{I}_\ell$, we have:

$$\sum_{v:(u,v) \in \Gamma_\ell} \hat{\beta}_{uv} - \sum_{v:(v,u) \in \Gamma_\ell} \hat{\beta}_{vu} = 0 = \begin{cases} x_{\ell t} & \text{if } u = \mathcal{I}_\ell^{(1)} \\ -x_{\ell t} & \text{if } u = \mathcal{I}_\ell^{(I_\ell)} \\ 0 & \text{otherwise} \end{cases}$$

- If $x_{\ell t} = 1$, then we identify the sequence of subpath arcs $a \in \mathcal{A}_{\ell st}$ such that $\hat{y}_a = 1$, which exists and defines the unique sequence of checkpoints per Equation (9). With a slight abuse of notation, this sequence is denoted by $\mathcal{I}_{\ell t} := \{\omega_1 := \mathcal{I}_\ell^{(1)}, \dots, \omega_O := \mathcal{I}_\ell^{(I_\ell)}\}$. We obtain the flow balance constraints for each checkpoint $u \in \mathcal{I}_\ell$:

$$\sum_{v:(u,v) \in \Gamma_\ell} \hat{\beta}_{uv} - \sum_{v:(v,u) \in \Gamma_\ell} \hat{\beta}_{vu} = \begin{cases} \hat{\beta}_{\omega_1, \omega_2} - 0 = 1 = x_{\ell t} & \text{if } u = \omega_1 \\ 0 - \hat{\beta}_{\omega_{O-1}, \omega_O} = -1 = -x_{\ell t} & \text{if } u = \omega_O \\ \hat{\beta}_{\omega_i, \omega_{i+1}} - \hat{\beta}_{\omega_{i-1}, \omega_i} = 1 - 1 = 0 & \text{if } u \in \mathcal{I}_{\ell t} \setminus \{\omega_1, \omega_O\} \\ 0 & \text{otherwise} \end{cases}$$

Next, the solution satisfies the checkpoint visit constraints given in Equation (EC.37):

- If $x_{\ell t} = 0$, then $\hat{\beta}_{uv} = 0$ for all $(u, v) \in \Gamma_\ell$, so the equation is trivially satisfied.
- If $x_{\ell t} = 1$, then we enumerate the set of visited checkpoints by the subpaths in $\mathcal{I}_{\ell t}$ using Equations (9). We consider a checkpoint $v \in \mathcal{I}_\ell \setminus \mathcal{I}_\ell^{(1)}$. If $\sum_{w \in \mathcal{I}_\ell \setminus \{(v, w) \in \Gamma_\ell\}} \beta_{wv} = 1$, then there exists a subpath-based arc $a \in \mathcal{A}_{\ell st}$ such that $\hat{y}_a = 1$, $u_{r(a)} = w$, and $v_{r(a)} = v$, which terminates in v . Per construction of the segment-based arcs, there exist segments $\bar{a}_1, \dots, \bar{a}_{N-1} \in \bar{\mathcal{A}}_a$ such that $\hat{\xi}_{\bar{a}_1} = \dots = \hat{\xi}_{\bar{a}_{N-1}} = 1$, corresponding to segments e_1, \dots, e_{N-1} . Then,

$$\sum_{\substack{(i,j) \in \bar{\mathcal{A}}_{\ell st} : \\ k_j=v, t_j=T_{\ell t}(v)}} \hat{\xi}_{(i,j)} = 1,$$

and the constraint is satisfied. The constraint is trivially satisfied if $\sum_{w \in \mathcal{I}_\ell : (w,v) \in \Gamma_\ell} \beta_{wv} = 0$.

Finally, the solution satisfies the checkpoint sequencing constraints given in Equation (EC.38), by construction of the $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\beta}}$ variables. Indeed, $\widehat{\beta}_{uv} = 1$ whenever there exists an arc $\bar{a} \in \bigcup_{a \in \mathcal{A}_{\ell st} : \widehat{y}_a=1} \bar{\mathcal{A}}_a$ between checkpoints u and v and between times $T_{\ell t}(u)$ and $T_{\ell t}(v)$ such that $\widehat{\xi}_{\bar{a}} = 1$.

Next, the solution $\widehat{\boldsymbol{\xi}}$ achieves the same objective value as $\widehat{\mathbf{y}}$:

$$\begin{aligned}
\sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st}} \bar{g}_{\bar{a}} \widehat{\xi}_{\bar{a}} &= \sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st} : \widehat{\xi}_{\bar{a}}=1} \bar{g}_{\bar{a}} \\
&= \sum_{a \in \mathcal{A}_{\ell st} : \widehat{y}_a=1} \sum_{\bar{a} \in \mathcal{A}_{\bar{a}}} \bar{g}_{\bar{a}} \\
&= \sum_{a \in \mathcal{A}_{\ell st} : \widehat{y}_a=1} \sum_{e \in \mathcal{E}_r} \sum_{p \in \mathcal{P}_e} D_{ps} \left(\lambda \tau_{ep}^{\text{walk}} + \mu \tau_{ep}^{\text{wait}} + \sigma \frac{\tau_{ep}^{\text{travel}}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{\ell tp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{\ell tp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right) \\
&= \sum_{a \in \mathcal{A}_{\ell st} : \widehat{y}_a=1} \sum_{p \in \mathcal{P}_r} D_{ps} \left(\lambda \tau_{rp}^{\text{walk}} + \mu \tau_{rp}^{\text{wait}} + \sigma \frac{\tau_{rp}^{\text{travel}}}{\tau_p^{\text{dir}}} + \delta \frac{\tau_{\ell tp}^{\text{late}}}{\tau_p^{\text{dir}}} + \frac{\delta}{2} \frac{\tau_{\ell tp}^{\text{early}}}{\tau_p^{\text{dir}}} - M \right) \\
&= \sum_{a \in \mathcal{A}_{\ell st} : \widehat{y}_a=1} g_a \\
&= \sum_{a \in \mathcal{A}_{\ell st}} g_a \widehat{y}_a
\end{aligned} \tag{EC.51}$$

In conclusion, any subpath solution can be mapped into a feasible segment solution with the same objective value. Therefore, the segment-based formulation achieves an objective that is at most equal to the optimum of the subpath-based formulation.

Constructing a subpath solution from a time-load-expanded segment solution.

Suppose that $\widehat{\boldsymbol{\xi}}$ is a feasible solution to the segment-based formulation (Equations (EC.33)–(EC.40)). Assume that $x_{\ell t} = 1$. We leverage Equations (EC.34) to construct a subpath between checkpoints u and v and between times $T_{\ell t}(u), T_{\ell t}(v)$. Starting from the source checkpoint u , we select the arc $\bar{a}_1 \in \bar{\mathcal{A}}_{\ell st}$ incident with $u_{\ell st}$ for which $\widehat{\xi}_a = 1$, proceeding sequentially along the directed network until reaching the node corresponding to checkpoint v at time $T_{\ell t}(v)$. An outgoing arc is guaranteed at every intermediate node by Equation (EC.34), and boundary conditions at the checkpoints are guaranteed by Equation (EC.37).

Each arc $\bar{a}_i \in \bar{\mathcal{A}}_a := \{\bar{a}_1, \dots, \bar{a}_{N-1}\}$ corresponds to a segment $e_i := e(\bar{a}_i) \in \mathcal{E}_{\ell st}$ and a passenger pickup set \mathcal{P}_{e_i} . The sequence of segments $\mathcal{E}_r := \{e_1, \dots, e_{N-1}\}$ defines a subpath r from u_r to v_r , skipping at most K checkpoints in a row (by Equations (EC.36) and the definition of checkpoint pairs Γ_ℓ), adhering to the scheduled arrival times at the checkpoints (defined by Equations (EC.37)), obeying the vehicle's capacity (by definition of the node set $\bar{\mathcal{V}}_{\ell st}$ in the time-load-expanded network), and picking up the passengers in $\mathcal{P}_r := \bigcup_{i=1}^{N-1} \mathcal{P}_{e_i}$ (who are unique due to Equations (EC.35)). Thus, we obtain a unique and valid subpath-based arc in $\mathcal{A}_{\ell st}$, induced by $\widehat{\boldsymbol{\xi}}$.

Let us collect all such subpath arcs in the set $\mathcal{A}(\widehat{\boldsymbol{\xi}})$. We use $\mathcal{A}(\widehat{\boldsymbol{\xi}})$ to construct solution $\widehat{\mathbf{y}}$ from $\widehat{\boldsymbol{\xi}}$:

$$\widehat{y}_a = \begin{cases} 1 & \text{if } a \in \mathcal{A}(\widehat{\boldsymbol{\xi}}), \\ 0 & \text{otherwise,} \end{cases} \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall a \in \mathcal{A}_{\ell st}.$$

By construction, the solution satisfies Equations (9). If $x_{\ell t} = 1$, we constructed a unique subpath-based solution for each pair $(u, v) \in \Gamma_\ell$. If $x_{\ell t} = 0$, there was no subpath to construct. Therefore:

$$\sum_{a \in \mathcal{A}_{\ell st}} \widehat{y}_a = \sum_{a \in \mathcal{A}_{\ell st}} \mathbb{1}(a \in \mathcal{A}(\widehat{\boldsymbol{\xi}})) = x_{\ell t}, \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell.$$

The solution also satisfies Equation (10). For passenger $p \in \mathcal{P}$, we obtain, from the construction of the subpath solution and Equation (EC.34):

$$\begin{aligned} \sum_{a \in \mathcal{A}_{\ell st}: p \in \mathcal{P}_{r(a)}} \widehat{y}_a &= \sum_{a \in \mathcal{A}(\widehat{\boldsymbol{\xi}})} \mathbb{1}(p \in \mathcal{P}_{r(a)}) \\ &= \sum_{a \in \mathcal{A}(\widehat{\boldsymbol{\xi}})} \sum_{e \in \mathcal{E}_{r(a)}} \mathbb{1}(p \in \mathcal{P}_e) \\ &= \sum_{a \in \mathcal{A}(\widehat{\boldsymbol{\xi}})} \sum_{\bar{a} \in \bar{\mathcal{A}}_a} \mathbb{1}(p \in \mathcal{P}_{e(\bar{a})}) \\ &= \sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st}: \widehat{\xi}_{\bar{a}} = 1} \mathbb{1}(p \in \mathcal{P}_{e(\bar{a})}) \\ &= \sum_{\bar{a} \in \bar{\mathcal{A}}_{\ell st}: p \in \mathcal{P}_{e(\bar{a})}} \widehat{\xi}_{\bar{a}} \\ &\leq z_{\ell pst} \end{aligned}$$

Finally, the solutions $\widehat{\boldsymbol{\xi}}$ and $\widehat{\mathbf{y}}$ achieve the same objective values, which can be shown similarly to Equation (EC.51). Therefore, any segment solution can be mapped into a feasible subpath solution with the same objective value, and the subpath-based formulation achieves an objective that is at most equal to the optimum of the segment-based formulation. This concludes the proof of equivalence of the subpath-based and segment-based formulations.

Proof that the subpath-based relaxation is at least as strong as the segment-based relaxation.

The subpath-based formulation is a Dantzig-Wolfe reformulation of the segment-based formulation. Alternatively, we can observe that the arguments to map a segment-based solution into a subpath-based solution do not require the integrality of the subpath solution $\widehat{\mathbf{y}}$. By following the same steps as above, we can map any non-integral subpath solution $\widehat{\mathbf{y}}$ into a feasible segment solution with the same objective value, as follows:

$$\begin{aligned} \widehat{\xi}_{\bar{a}} &= \sum_{a \in \mathcal{A}_{\ell st}: \bar{a} \in \bar{\mathcal{A}}_a} \widehat{y}_a \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall \bar{a} \in \bar{\mathcal{A}}_{\ell st} \\ \widehat{\beta}_{uv} &= \sum_{a \in \mathcal{A}_{\ell st}: u_{r(a)} = u, v_{r(a)} = u} \widehat{y}_a \quad \forall (\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell, \forall u, v \in \Gamma_\ell. \end{aligned}$$

However, a non-integral segment solution cannot be mapped directly to a subpath solution. We demonstrate this claim with an example with three checkpoints (A, B and C). Figure EC.1 shows a non-integral segment-based solution—all shown segments have a flow of 0.5. The solution satisfies the flow balance constraints from station to station given in Equation (EC.34), the flow balance constraints from checkpoint to checkpoint given in Equation (EC.36), as well as the consistency constraints between checkpoint-checkpoint flows and station-station flows given in Equations (EC.37)–(EC.38). In this solution $\hat{\beta}$ values are $\hat{\beta}_{(A,B)} = \hat{\beta}_{(B,C)} = \hat{\beta}_{(A,C)} = 0.5$, so that the flows are split between a subpath from Checkpoint A to Checkpoint B, a subpath from Checkpoint B to Checkpoint C, and a subpath from Checkpoint A to Checkpoint C, each with a flow of 0.5. The critical observation is that the solution leverages the segments (shown in solid lines) that fall outside the spatial scope of the deviations between Checkpoints A and B as part of the subpath connecting Checkpoints A and B. Specifically, there exists a segment $(n, m) \in \mathcal{A}_{\text{lst}}$ with $k_a \in \mathcal{N}_{13} \setminus \mathcal{N}_{12}$ or $k_b \in \mathcal{N}_{13} \setminus \mathcal{N}_{12}$. This solution belongs to the polyhedron defined by the segment-based formulation, because $\xi_{(n,m)} = 0.5 \leq \beta_{(A,C)} = 0.5$. However, the resulting subpath is infeasible because it connects Checkpoints A and B without adhering to the maximum deviation Δ . This proves that the subpath-based relaxation is at least as strong as the segment-based relaxation.

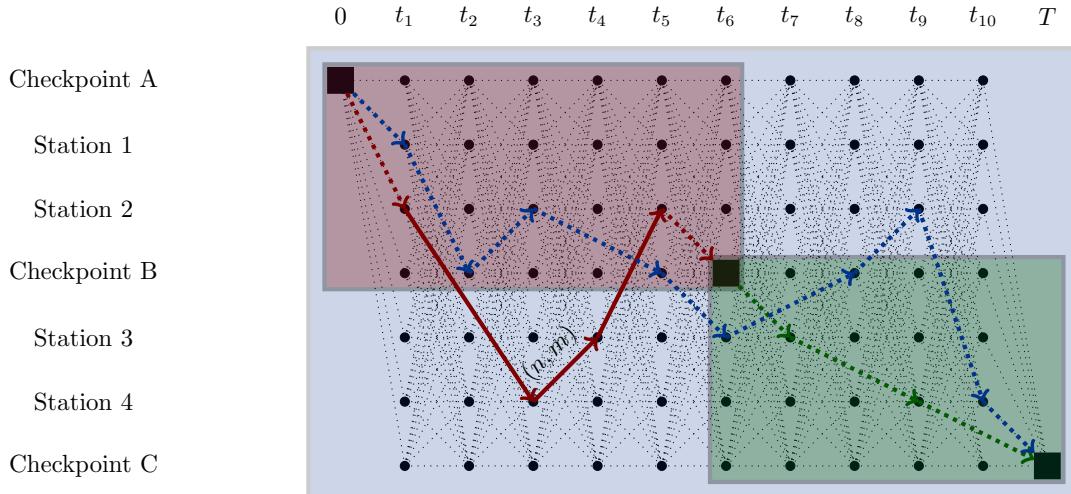


Figure EC.1 Example of a non-integral segment solution that cannot be mapped to a subpath solution. For simplicity, the load dimension of the time-load-expanded network is omitted. The black squares encode the reference schedule at the checkpoints. The red (resp. green, blue) area represents the stations that can be reached between Checkpoints A and B (resp. between Checkpoints B and C, between Checkpoints A and C). All thick segments are associated with a flow of 0.5. The solid segments are outside the allowable region in the subpath-based formulation.

EC.2.5. Proof of Proposition 2

Let D_{\min} be the minimum distance between pickup locations (a constant dictated by the station set \mathcal{N}) and let Π denote the maximum distance between any checkpoint pair (a constant dictated by the candidate reference lines and the value of $K = 0$ vs. $K = 1$). Recall that Δ denotes the maximum deviation from the reference line for microtransit vehicles, and Ω denotes the maximum walking distance. Therefore, the rectangular service area associated with a checkpoint pair has side lengths $\Pi + 2(\Delta + \Omega)$ and $2(\Delta + \Omega)$. The maximum number of stations between checkpoints is $\Xi = \lfloor \frac{(\Pi+2(\Delta+\Omega))}{D_{\min}} \rfloor \cdot \lfloor \frac{2(\Delta+\Omega)}{D_{\min}} \rfloor$. In our case study, Ξ is significantly less than $|\mathcal{N}| = 640$.

Segment-based model.

The number of variables scales in $\mathcal{O}(T_S \cdot C_\ell^2 \cdot I_\ell \cdot \Xi^2)$.

- The β variables are indexed over the set of valid directed checkpoint pairs Γ_ℓ . Since K is a small constant, the number of valid checkpoint pairs scales linearly with the number of checkpoints, so that $\mathcal{O}(|\Gamma_\ell|) = \mathcal{O}\left(\sum_{i=1}^{I_\ell-(K+1)} (K+1)\right) = \mathcal{O}(I_\ell)$.
- The ξ variables are indexed over the set $\bar{\mathcal{A}}_{\ell st}$, i.e., the arcs in the time-load-expanded network. Arcs define connections between consecutive stations between a checkpoint pair (which scale in $\mathcal{O}(I_\ell \cdot \Xi^2)$) for each time period, and they can correspond to any vehicle load pair, so the number of ξ variables scales in $\mathcal{O}(\bar{\mathcal{A}}_{\ell st}) = \mathcal{O}(T_S \cdot C_\ell^2 \cdot I_\ell \cdot \Xi^2)$.

The number of constraints scales in $\mathcal{O}(|\mathcal{P}| + T_S \cdot C_\ell \cdot |\mathcal{N}| + T_S \cdot C_\ell^2 \cdot I_\ell \cdot \Xi^2)$.

- Equations (EC.34): There is one constraint per node in the time-load-expanded network. There is one node per combination of time periods in \mathcal{T}^S , vehicle loads in \mathcal{C}_ℓ , and stations in \mathcal{N} , so that there are $\mathcal{O}(T_S \cdot C_\ell \cdot |\mathcal{N}|)$ flow balance constraints.
- Equations (EC.35): The passenger linking constraints scale with $\mathcal{O}(\sum_{p \in \mathcal{P}} |\mathcal{M}_p|)$. The cardinality of each set \mathcal{M}_p is bounded by a small constant, so there are $\mathcal{O}(|\mathcal{P}|)$ linking constraints.
- Equations (EC.36)–(EC.37): There are $\mathcal{O}(I_\ell)$ flow balance constraints for checkpoint-to-checkpoint flows and $\mathcal{O}(I_\ell)$ schedule adherence constraints.
- Equations (EC.38): There is one constraint per arc in the time-load-expanded network to ensure consistency between the station-to-station and checkpoint-to-checkpoint flows, which grows in $\mathcal{O}(\bar{\mathcal{A}}_{\ell st}) = \mathcal{O}(C_\ell^2 \cdot T_S \cdot I_\ell \cdot \Xi^2)$, as previously established.

The complexity of Equations (EC.36)–(EC.37) is dominated by that of Equations (EC.34) and (EC.38). The result follows.

Subpath-based model. The number of variables scales in $\mathcal{O}(I_\ell \cdot C_\ell \cdot 2^\Xi)$. In particular, y scales with $\mathcal{O}(|\mathcal{A}_{\ell st}|)$, the number of arcs in the load-expanded subpath network. By definition:

$$|\mathcal{A}_{\ell st}| = |\mathcal{A}_{\ell st}^v| + \sum_{r \in \mathcal{R}_{\ell st}} |\mathcal{A}_r|.$$

As for $|\mathcal{A}_{\ell st}^v|$, there are $C_\ell + 1$ arcs connecting the last stop to the sink node (one per vehicle load). Turning to \mathcal{A}_r , a subpath $r \in \mathcal{R}_{\ell st}$ is the shortest path to serve the corresponding passenger set \mathcal{P}_r . Thus, the number of subpath variables is proportional to the number of possible sets \mathcal{P}_r . The number of different passengers that can be picked up at each station is bounded by a small constant, so we use the number of stations as a proportional proxy for the number of passengers that can be picked up. There are up to $\binom{\Xi}{c}$ station combinations that pick up c passengers, each of which can be replicated $C_\ell - c + 1$ times in the arc set (corresponding to initial loads $0, 1, \dots, C_\ell - c$). Therefore, the number of subpaths is

$$\sum_{c=0}^{C_\ell} \binom{\Xi}{c} \cdot (C_\ell - c + 1) \leq (C_\ell + 1) \cdot \sum_{c=0}^{C_\ell} \binom{\Xi}{c}.$$

When $\Xi \leq C_\ell$, the binomial sum above is equal to 2^Ξ . When $\Xi > C_\ell$, it is equal to

$$\sum_{c=0}^{C_\ell} \binom{\Xi}{c} = 2^\Xi - \sum_{c=C_\ell+1}^{\Xi} \binom{\Xi}{c} \leq 2^\Xi.$$

Therefore, $\mathcal{O}(|\mathcal{A}_{\ell st}|) = \mathcal{O}(2^\Xi \cdot C_\ell \cdot I_\ell)$.

The number of constraints scales in $\mathcal{O}(|\mathcal{P}| + C_\ell \cdot I_\ell)$.

- Equations (9): There are $\mathcal{O}(\mathcal{V}_{\ell st}) = \mathcal{O}(C_\ell \cdot I_\ell)$ flow balance constraints, one per node in the load-expanded subpath network.
- Equations (10): There are $\mathcal{O}(|\mathcal{P}|)$ linking constraints, one per passenger that can be picked up by reference trip (ℓ, t) .

Path-based model. The number of variables scales in $\mathcal{O}(2^{\Xi \cdot I_\ell})$. The ζ variables are indexed over the path set $\mathcal{Q}_{\ell st}$. Each path $q \in \mathcal{Q}_{\ell st}$ can be decomposed into a sequence of subpaths in $\mathcal{R}_{\ell st}$ by partitioning the path-based passenger set \mathcal{P}_q into subpath-based passenger sets \mathcal{P}_r , that is $\mathcal{P}_q = \bigcup_{r \in \mathcal{R}_q} \mathcal{P}_r$. Recall that there are $O(2^\Xi)$ possible subpaths between each checkpoint pair, and there are $\mathcal{O}(I_\ell)$ checkpoint pairs, so we obtain $\mathcal{O}(2^{\Xi \cdot I_\ell})$ overall paths.

The number of constraints scales in $\mathcal{O}(|\mathcal{P}|)$, i.e., with the number of linking constraints (Equations (EC.45)). The model also comprises a single partitioning constraint (Equation (EC.44)), which does not affect the constraint complexity.

EC.3. Details on Solution Algorithm

EC.3.1. Proof of Proposition 3

Fix a reference trip $(\ell, t) \in \mathcal{L} \times \mathcal{T}_\ell$ and a scenario $s \in \mathcal{S}$.

Let us consider a pair of checkpoints $(u, v) \in \Gamma_\ell$ and two load values $c_1 \leq c_2 \in \mathcal{C}_\ell$. Let us define the load differential as $\varepsilon = c_2 - c_1$. By construction, the load component of the reduced cost satisfies:

$$\Delta \varphi_{\ell st}^{u, v, \varepsilon} \geq \varphi_{(u, c_1)} - \varphi_{(v, c_2)} \tag{EC.52}$$

Consider a solution \mathbf{f}^* , \mathbf{w}^* , $\boldsymbol{\xi}^*$ of the pricing problem $\text{PP}_{\ell st}^{u,v,c_1,c_2}$. With a slight abuse of notation, we also refer to its optimal value as $\text{PP}_{\ell st}^{u,v,c_1,c_2}$. By construction, the solution \mathbf{f}^* , \mathbf{w}^* defines a feasible solution to the problem defining $Z_{\ell st}^{u,v,\varepsilon}$. Indeed, the load differential satisfies

$$\sum_{m \in \mathcal{U}_{\ell st}^{u,v}} \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^* = \sum_{(m,q) \in \mathcal{H}_{\ell st}^{u,v} : f_{mq}=1} (\xi_q^* - \xi_m^*) = \xi_{(v,T_{\ell t}(v))}^* - \xi_{(u,T_{\ell t}(u))}^* = c_{(v,c_2)} - c_{(u,c_1)} = \varepsilon,$$

where the first equality is induced by Equations (28)–(29), the second equality is induced by telescoping the sum from Equation (32), the third equality is induced by Equation (27), and the last inequality is by assumption.

Therefore, the routing component of the reduced cost expression satisfies:

$$Z_{\ell st}^{u,v,\varepsilon} \leq \sum_{m \in \mathcal{U}_{\ell st}^{u,v}} \sum_{p \in \mathcal{P}_m} d_{mp} w_{mp}^* \quad (\text{EC.53})$$

From Equations (EC.52) and (EC.53), we obtain:

$$Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon} \leq \sum_{m \in \mathcal{U}_{\ell st}^{u,v}} \sum_{p \in \mathcal{P}_m} d_{mp} w_{mp}^* + \varphi_{(v,c_2)} - \varphi_{(u,c_1)} = \text{PP}_{\ell st}^{u,v,c_1,c_2}$$

By taking the minimum over all arcs with a load differential ε , we obtain:

$$Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon} \leq \min_{c_1, c_2 \in \mathcal{C}_\ell : c_2 - c_1 = \varepsilon} \text{PP}_{\ell st}^{u,v,c_1,c_2}, \quad \forall (u, v) \in \Gamma_\ell \quad (\text{EC.54})$$

Vice versa, let us consider two checkpoints $(u, v) \in \Gamma_\ell$ and a load differential $\varepsilon \in \mathcal{C}_\ell$. Consider an arc $a^* \in \mathcal{A}_{\ell st}$ that maximizes the load component of the reduced cost and a solution \mathbf{f}^* , \mathbf{w}^* that minimizes the routing component for that load differential. Specifically, the arc $a^* \in \mathcal{A}_{\ell st}$ defines a subpath that starts in checkpoint $u = k_{start(a^*)} \in \mathcal{I}_\ell$ at time $T_{\ell t}(u)$ with vehicle load $c_{start(a^*)}$, that ends in checkpoint $v = k_{end(a^*)} \in \mathcal{I}_\ell$ at time $T_{\ell t}(v)$ with load $c_{end(a^*)} = c_{start(a^*)} + \varepsilon$, and that satisfies

$$\varphi_{start(a^*)} - \varphi_{end(a^*)} = \Delta \varphi_{\ell st}^{u,v,\varepsilon}$$

The solution \mathbf{f}^* , \mathbf{w}^* satisfies Equations (30)–(33) by construction. We then define a load variable ξ_m , keeping track of the load at node $m \in \mathcal{U}_{\ell st}^{u,v}$. We initialize it with:

$$\xi_{(u,T_{\ell t}(u))} = c_{start(a^*)}.$$

Following solution \mathbf{f}^* , \mathbf{w}^* , we increase ξ_m by $\sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^*$ if we traverse $(m, q) \in \mathcal{H}_{\ell st}^{u,v}$:

$$\xi_q - \xi_m = \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^*, \quad \forall (m, q) \in \mathcal{H}_{\ell st}^{u,v} : f_{mq}^* = 1.$$

The variables ξ_m satisfy Equations (28)–(29) and (33) by construction. By combining it with Equations (32), and telescoping the sum, we obtain:

$$\begin{aligned}\xi_{(v, T_{\ell t}(v))} &= \sum_{m \in \mathcal{U}_{\ell st}^{uv}: f_{m,(v, T_{\ell t}(v))}^* = 1} \left(\xi_m + \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^* \right) \\ &= \xi_{(u, T_{\ell t}(u))} + \sum_{(m,q) \in \mathcal{H}_{\ell st}^{uv}: f_{mq}^* = 1} \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^* \\ &= c_{start(a^*)} + \varepsilon \\ &= c_{end(a^*)},\end{aligned}$$

where the third equality comes from the initialization $\xi_{(u, T_{\ell t}(u))} = c_{start(a^*)}$ and the constraint $\sum_{m \in \mathcal{U}_{\ell st}^{uv}} \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp} = \varepsilon$, and the last equality follows from the construction of $a^* \in \mathcal{A}_{\ell st}$. Therefore, the variables ξ_m also satisfy Equations (27).

Therefore, solution \mathbf{f}^* , \mathbf{w}^* , ξ defines a feasible solution for the pricing problem $\text{PP}_{\ell st}^{u,v,c_{start(a^*)},c_{end(a^*)}}$, and we have:

$$Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon} = \sum_{p \in \mathcal{P}_m} D_{ps} w_{mp}^* + \varphi_{end(a^*)} - \varphi_{start(a^*)}$$

Since, by construction, $c_{end(a^*)} - c_{start(a^*)} = \varepsilon$, we obtain:

$$Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon} \geq \min_{c_1, c_2 \in \mathcal{C}_\ell: c_2 - c_1 = \varepsilon} \text{PP}_{\ell st}^{u,v,c_1,c_2}.$$

This completes the proof that $Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon}$ is equal to the minimum reduced cost across all variables with load differential ε :

$$Z_{\ell st}^{u,v,\varepsilon} - \Delta \varphi_{\ell st}^{u,v,\varepsilon} = \min_{c_1, c_2 \in \mathcal{C}_\ell: c_2 - c_1 = \varepsilon} \text{PP}_{\ell st}^{u,v,c_1,c_2}$$

EC.3.2. Proof of Remark 1

Suppose that Equation (32) is replaced with the following constraints in the PP.

$$\sum_{q: (m,q) \in \mathcal{H}_{\ell st}^{uv}} f_{mq} - \sum_{q: (q,m) \in \mathcal{H}_{\ell st}^{uv}} f_{qm} = \begin{cases} x_{\ell t} & \text{if } m = (u, T_{\ell t}(u)), \\ x_{\ell t} & \text{if } m = (v, T_{\ell t}(v)), \\ 0 & \text{otherwise.} \end{cases} \quad \forall m \in \mathcal{U}_{\ell st}^{uv} \quad (\text{EC.55})$$

With Equation (EC.55), the resulting optimal solution to the PP could not be used to construct a subpath, which by definition is a sequence of arcs connecting checkpoints u and v .

Suppose toward a contradiction that Equation (31) is replaced with the following constraints in the PP.

$$\sum_{m \in \mathcal{U}_{\ell st}^{uv}: p \in \mathcal{P}_m} w_{mp} \leq z_{\ell pst} \quad \forall p \in \mathcal{P} : (\ell, t) \in \mathcal{M}_p \quad (\text{EC.56})$$

Consider the subset of passengers $p \in \mathcal{P}_{\ell st}$ for which $z_{\ell p st} = 0$. Then the pricing problem only constructs arcs over the following set:

$$\mathcal{A}_{\ell st}(\mathbf{z}) := \{a \in \mathcal{A}_{\ell st} : z_{p \ell st} = 1, \forall p \in \mathcal{P}_{r(a)}\}.$$

As a result, the optimal dual solution $(\boldsymbol{\varphi}, \boldsymbol{\gamma})$ to the corresponding RMP would have unknown feasibility to the following constraints:

$$\varphi_n - \varphi_m - \sum_{p \in \mathcal{P}_a} \gamma_p \leq g_{(n,m)} \quad \forall (n, m) \in \mathcal{A}_{\ell st} \setminus \mathcal{A}_{\ell st}(\mathbf{z}). \quad (\text{EC.57})$$

Thus, the solution $(\boldsymbol{\varphi}, \boldsymbol{\gamma})$ is not necessarily in $\Lambda_{\ell st}$, and the corresponding constraint (Equation (21)) would not be valid in the Benders decomposition algorithm.

EC.3.3. Proof of Proposition 5

Benders decomposition returns an optimal solution to MiND-VRP' in a finite number of iterations as long as the algorithm for solving the Benders subproblem supplies an optimal dual solution at each iteration. The column generation algorithm yields an optimal dual solution to each Benders subproblem as long as the pricing problem is correct. The pricing problem is correct as long as it enables convergence of the column generation algorithm to a valid Benders dual subproblem solution. Per Lemma EC.1, we can procure a discrete time interval that maintains the feasibility of all subpaths in the time-expanded network. In Proposition 1, the time discretization ρ ensures that each sequence of segments constructs a valid subpath, and that it is possible to recover every subpath from a sequence of time-expanded roadway segments. Here, we leverage an analogous argument to conclude that any solution to the pricing problem (a sequence of segments in the time-expanded network) yields a valid subpath, and furthermore, that it searches over the full set of valid subpaths. By Remark 1, the dual solution to the Benders subproblem is optimal upon termination of the column generation algorithm. The result follows.

EC.4. Experimental Setup

In this appendix, we provide details on the generation of the model inputs (EC.4.1); in particular, we present a breadth-first search tree approach to define candidate reference lines (EC.4.2). Figure EC.2 illustrates these inputs. We also detail our ride-sharing benchmarks (EC.4.3).

EC.4.1. Model Inputs

We developed a real-world experimental setup in Manhattan, using data from the NYC Taxi & Limousine Commission (2021). We filtered trips to the airports during the morning rush (6–9 am), leading to up to 1,900 passenger request per instance (shown in Figure EC.2a of the paper). We defined a road network and travel times using data from Google Maps, OpenStreetMap, and Uber

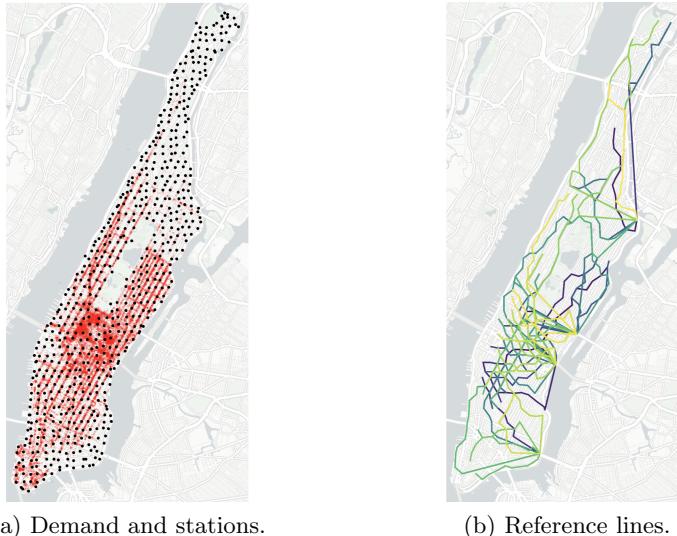


Figure EC.2 **Visualization of MiND-VRP inputs in Manhattan.**

(2020). We considered pickup stations 300 meters apart, leading to 640 stations (also shown in Figure EC.2a of the paper). We assumed passengers could originate from any of the approximately 20,000 roadway intersections in Manhattan, and that they would walk from their origin to the closest station. We obtained the mapping and routing inputs from the `fastest_route` functionality in the OpenStreetMapX package in Julia (Szufel, Przemysław *et al.* 2023). We calibrated travel time estimates to heavy Manhattan traffic using speed data from Uber (2020). We computed average speeds during the morning rush for each roadway type present in our Manhattan map (primary, secondary, tertiary, unclassified) and used these average speeds as input to the travel time estimation function, overriding default speeds provided by OpenStreetMapX.

Recall that our MiND-VRP experiments model a shuttle service from Manhattan to LaGuardia Airport with vehicles of capacity 10 to 20 passengers. Every trip leaves Manhattan and heads directly toward LaGuardia Airport via four possible exits: the Queensboro Bridge, the Williamsburg Bridge, the Kennedy Bridge, and the Midtown Tunnel. Travel times from each exit to LaGuardia were obtained via Google Maps estimates during the morning rush.

Table EC.5 reports the parameter values used in our computational experiments (Section 4), and practical experiments for the MiND-VRP (Section 5) and MiND-DAR (Appendix EC.1.3).

EC.4.2. Reference Line Generation

We describe the process of generating the set \mathcal{L} of candidate reference lines (shown in Figure EC.2b of the paper). The procedure proceeds in three steps: (i) generating a comprehensive routing graph over Manhattan; (ii) using breadth-first search (BFS) trees to generate a very large set of candidate reference lines; and (iii) clustering and filtering to obtain a small but representative final set of candidate reference lines. We describe each step in detail below.

Table EC.5 Details on input calibration for computational and practical analyses.

Model component	Section 4 value(s)	Section 5 value(s)	EC.1.3 value(s)
Ω	210 meters	420 meters	250 meters
Ψ	10 minutes	10 minutes	10 minutes
Δ	600 meters	600 or 1,200 meters	300 meters
α	5 minutes	10 minutes	10 minutes
C_ℓ	10 people	10, 15, or 20 people	5, 10 or 20 people
T_S	30 seconds	30 seconds	30 seconds
μ	1	1	1
λ	1	1	1
σ	1	1	1
δ	1	1	1
κ	1	1	1
M	10,000	10,000	10,000
h_ℓ	$T_{et}(\mathcal{I}_\ell^{(I_\ell)}) - T_{et}(\mathcal{I}_\ell^{(1)})$	$T_{et}(\mathcal{I}_\ell^{(I_\ell)}) - T_{et}(\mathcal{I}_\ell^{(1)})$	$T_{et}(\mathcal{I}_\ell^{(I_\ell)}) - T_{et}(\mathcal{I}_\ell^{(1)})$
F	$ \mathcal{L} $ vehicles	10 or 20 vehicles	5 or 10 vehicles
\mathcal{T}_ℓ	15 minute intervals	15 minute intervals	15 minute intervals
$T_{et}(\mathcal{I}_\ell^{(i+1)}) - T_{et}(\mathcal{I}_\ell^{(i)})$	120% of direct	120% of direct	110% of direct

$T_{et}(\mathcal{I}_\ell^{(i+1)}) - T_{et}(\mathcal{I}_\ell^{(i)})$: buffer time between arrival times at consecutive checkpoints $\mathcal{I}_\ell^{(i)}$ and $\mathcal{I}_\ell^{(i+1)}$.

\mathcal{T}_ℓ : the frequency set is populated with departure times at evenly spaced intervals across the demand horizon.

T_S : Time elapsed between consecutive discrete time units (between t and $t + 1$) in the discretized set \mathcal{T}^S .

Note that our procedure to construct and optimize reference lines relies on a training set of demand data. This process avoids any bias moving from design to evaluation.

Manhattan routing graph. We build a node set using discrete locations in Manhattan by generating a grid of GPS coordinates spanning Manhattan that were each 300 meters apart, and snapping each node to the closest road intersection. The outcome of this process is a list of candidate checkpoints \mathcal{N} , shown in Figure EC.3a. We then build an edge set over this routing graph by connecting each node to its six closest neighbors according to their Euclidean distance. We used OpenStreetMapX to remove any edges that were impossible for a vehicle to traverse.

BFS trees. To generate a large set of reference line candidates, we build BFS trees over the routing graph. Specifically, we let each node be the root of a BFS tree over the routing network (see Figure EC.3b). We then build reference line candidates over each BFS tree, by constructing node sequences from the root node to each leaf. Ultimately, we obtain tens of thousands of distinct candidate reference lines, across all BFS trees.

Clustering and filtering. We first filter out many candidate lines that are illogical (e.g., indirect lines, very short or very long lines). We developed several metrics of line quality to systematically filter out low-quality options:

- *Minimum number of checkpoints.* Each line must visit a minimum of 10 stations.
- *Low average and maximum detour.* For each checkpoint, we compute the relative detour as the ratio of the travel time from the checkpoint to the destination (LaGuardia) with the reference

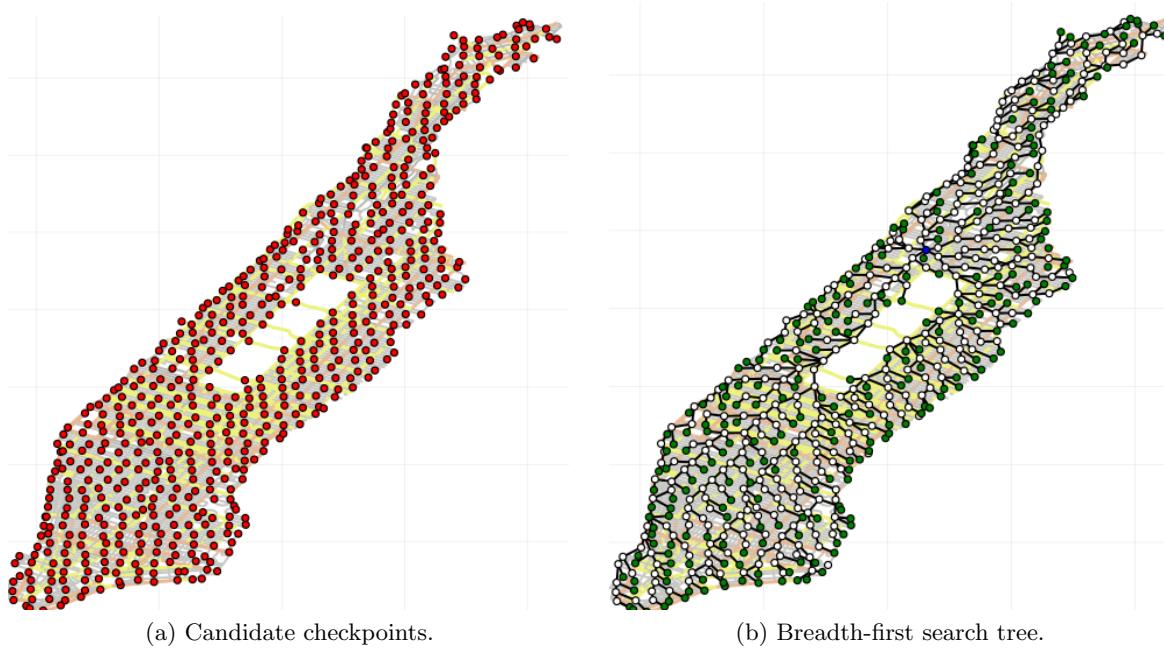


Figure EC.3 Candidate checkpoints and BFS tree (blue: root node; green: leaves; white: intermediate nodes)

line and the corresponding direct travel time. The average detour across all checkpoints should not exceed 200%, and the maximum detour should not exceed 250%.

- *Limited wrong-way travel.* To measure travel in the “wrong direction,” we measure the percentage of a reference line’s checkpoints that are farther away from LaGuardia than their immediate predecessors.
- *Demand coverage.* We assigned a popularity score to each checkpoint based on the frequency of trip requests with pickup locations close to that stop—in a training dataset. We filter out lines with a low average popularity score across its checkpoints.

Then, we remove redundancy over overlapping candidate lines, which is especially present among lines constructed from the same BFS tree. We measure the dissimilarity of two candidate lines as:

$$\text{dissim}_{k\ell} = 1 - \frac{|\mathcal{I}_k \cap \mathcal{I}_\ell|}{\min\{\mathcal{I}_k, \mathcal{I}_\ell\}}.$$

When $\text{dissim}_{k\ell} = 0$, lines k and ℓ share as many stops as possible and are therefore substitutable. We collect these substitutable pairs into an undirected graph, and define an updated set of candidate lines \mathcal{L}' by computing a minimum vertex cover over that graph.

At this point, we are left with approximately 3,000 candidate lines in \mathcal{L}' . In order to retain a tractable set of candidate lines in the optimization model, we cluster them into 100 representative and high-quality options. Specifically, we formulate a bi-objective clustering model to maximize medoid quality and diversity. Let $y_\ell \in \{0, 1\}$ indicate whether line $\ell \in \mathcal{L}'$ is selected in the final set

\mathcal{L} , and $x_{k\ell} \in \{0, 1\}$ indicate whether line $k \in \mathcal{L}'$ is assigned to the cluster with medoid line $\ell \in \mathcal{L}'$. We define a line-dependent parameter q_ℓ penalizing undesirable line characteristics based on the aforementioned metrics.

The clustering model maximizes line quality and minimizes the total dissimilarity among the line mapping (Equation (EC.58)), subject to partitioning constraints (Equation (EC.59)), consistency constraints (Equation (EC.60)) and budget constraints (Equation (EC.61)). We define the final reference line set as $\mathcal{L} := \{\ell \in \mathcal{L}' : y_\ell = 1\}$.

$$\min \quad \sum_{\ell \in \mathcal{L}'} q_\ell y_\ell + \lambda \sum_{k \in \mathcal{L}'} \sum_{\ell \in \mathcal{L}'} \text{dissim}_{k\ell} x_{k\ell} \quad (\text{EC.58})$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{L}'} x_{k\ell} = 1 \quad \forall \ell \in \mathcal{L}' \quad (\text{EC.59})$$

$$x_{k\ell} \leq |\mathcal{L}'| y_\ell \quad \forall k, \ell \in \mathcal{L}' \quad (\text{EC.60})$$

$$\sum_{\ell \in \mathcal{L}'} y_\ell = 100 \quad (\text{EC.61})$$

$$\mathbf{x} \in \{0, 1\}^{\mathcal{L}' \times \mathcal{L}'} \quad (\text{EC.62})$$

$$\mathbf{y} \in \{0, 1\}^{\mathcal{L}'} \quad (\text{EC.63})$$

We constructed three candidate line sets with 100 lines each by scaling the aforementioned quality measures with the following parameter settings:

$$q_\ell^{\text{cluster}} = 0 \quad \forall \ell \in \mathcal{L}'$$

$$q_\ell^{\text{direct}} = \frac{1}{3} \cdot (\text{maxDetour}_\ell + \text{meanDetour}_\ell + \text{wrongWay}_\ell) \quad \forall \ell \in \mathcal{L};$$

$$q_\ell^{\text{popular}} = \text{popularity}_\ell \quad \forall \ell \in \mathcal{L}'$$

Throughout the manuscript, we use q^{popular} as the default to focus on the demand coverage objective, except for Section 5.1 on microtransit network design, in which we consider the line sets corresponding to all three quality measures.

EC.4.3. Ride-sharing Benchmark

We build our ride-sharing benchmark using the cluster-then-route heuristic from Bertsimas and Yan (2021), originally built to generate paratransit itineraries with up to 4 passengers per vehicle. Their approach was itself based on the maximum weighted matching over a shareability network from Santi et al. (2014). To extend the approach from two- to four-passenger trips, Bertsimas and Yan (2021) first created a set of passenger pairs and then approximated the shareability network over two-passenger trips. We adopt a similar approach except that, instead of requiring all requests to be served, we maximize the number of served requests and then minimize travel times.

Single-occupancy ride-sharing. With single-occupancy vehicles, the clustering step is unnecessary. We simply apply the routing step from Bertsimas and Yan (2021) over the request set.

Two-occupancy ride-sharing. We build a pair-wise shareability network that encodes the pairs of trips that can share a vehicle. Let t_i denote the requested pickup time of request i , T_i the direct travel time of request i , and $tt(x, y)$ the travel time from location x to location y .

- If $t_j \leq t_i + T_i + \Psi$, then trip j can be picked up before trip i is dropped off;
- if $t_i \leq t_j + T_j + \Psi$, then trip i can be picked up before trip j is dropped off; and
- otherwise, trips i and j cannot be shared.

Then we determine whether there exists pickup times for trips i and j (in that order) such that no request is picked up early and each pickup is within Ψ of their requested times. The following conditions must hold, where x denotes the pickup time of trip i :

$$\begin{array}{ll} t_i \leq x \leq t_i + \Psi & \text{Request } i \text{ has tolerable wait time} \\ t_j \leq x + tt(o_i, o_j) \leq t_j + \Psi & \text{Request } j \text{ has tolerable wait time} \end{array}$$

which reduces to finding some x such that:

$$x \in [\max\{t_i, t_j - tt(o_i, o_j)\}, \min\{t_i + \Psi, t_j + \Psi - tt(o_i, o_j)\}].$$

The two requests can also share a vehicle if the symmetric problem holds, corresponding to the instance where trip j is picked up first:

$$x \in [\max\{t_j, t_i - tt(o_j, o_i)\}, \min\{t_j + \Psi, t_i + \Psi - tt(o_j, o_i)\}]$$

Bertsimas and Yan (2021) impose a maximum delay limit, but we remove this restriction to enable more ride-pooling. Finally, we determine the travel time associated with each version of the trip.

$$\begin{aligned} c_{i \rightarrow j} &= tt(o_i, o_j) + T_j \\ c_{j \rightarrow i} &= tt(o_j, o_i) + T_i \end{aligned}$$

If $c_{i \rightarrow j} \leq T_i + T_j$ or $c_{j \rightarrow i} \leq T_j + T_i$, then the shared trip is more efficient than serving the two requests separately. If both are efficient, then we select the best option.

The shared trips satisfying the above conditions are added to the VSN with cost $T_i + T_j - \min\{c_{i \rightarrow j}, c_{j \rightarrow i}\}$ to reflect the cost savings of pooling the requests. We solve a maximum weighted matching problem to pair requests into capacity-2 trips, with some requests potentially still served in isolation if they are not matched to any other request. We first maximize the number of served requests, and then we minimize the total travel time, subject to the fleet size limit.

Four-occupancy ride-sharing. We build a new shareability network that combines trip pairs from the pair-wise shareability network. For the MiND-VRP, we solve a simple vehicle routing problem for each candidate set of four trips to find the best sequence of stops within that set, while ensuring that no one is picked up earlier than their requested times and that none of passengers' wait times exceeds limit Ψ . For the MiND-DAR, we solve a simple dial-a-ride problem for each candidate set of four trips, which also includes precedence constraints so that each pickup occurs before the corresponding dropoff. We note that the optimal pooling configuration of two request pairs could potentially be to serve all four requests together, or to pool only a subset of these requests and serve the remaining requests separately. We proceed as in the two-occupancy case, solving a maximum weighted matching problem over the VSN to determine final trips, and then performing an identical itinerary generation procedure to the one described previously.