Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000, EISSN 0000-0000

Stochastic Bilevel Optimization for the Network Design of Multimodal Transit Systems with Heterogeneous Rider Preferences under Uncertain Travel Times and Demand

Suri Liu
State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian, China, liusuri@mail.dlut.edu.cn

Yiling Zhang
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN, yiling@umn.edu

Beste Basciftci
Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, IA, beste-basciftci@uiowa.edu

Wenyuan Wang
State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian, China, wangwenyuan@dlut.edu.cn

Abstract. Designing efficient and user-friendly multimodal transit networks is critical for modern urban mobility. We study a novel stochastic multimodal transit network design problem that integrates fixed-route services with on-demand shuttles, explicitly accounting for heterogeneous rider preferences, uncertain travel times, and passenger demand. The hierarchical decision-making process is modeled using a two-stage stochastic bilevel optimization problem, where the transit agency (leader) determines the network design, and riders (followers) select their preferred routes based on realized traffic conditions. The model inherits the complexity of a nonconvex bilevel structure with stochastic programming, posing significant computational challenges. To address this, we first develop an equivalent single-level mixed integer linear programming (MILP) reformulation by introducing a response search algorithm that efficiently enumerates critical follower route choices. To further enhance scalability, we propose a decomposition method that combines a relaxed formulation with a subset of follower responses and iteratively strengthens it with valid cutting planes. Computational experiments on instances derived from a public transit network in Dalian, China, demonstrate the efficiency and effectiveness of our approaches, achieving over 10 times speedups compared to existing single-level reformulations. Additionally, a comprehensive case study on the Ann Arbor/Ypsilanti region in Michigan highlights practical benefits of the proposed framework, yielding up to 12% cost savings and up to 7% improvements in route convenience, demonstrating the value of the proposed stochastic bilevel model over deterministic or single-level counterparts.

Key words: Multimodal Transit, Network Design, Stochastic Bilevel Optimization, Two-stage Stochastic Programming, Integer Programming, Decomposition Methods

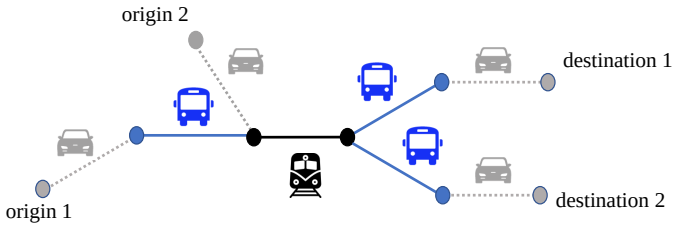
1. Introduction

Optimizing transit network designs is a fundamental challenge in transportation systems and urban planning, which aims to identify optimal designs of routes and services such as buses and rail (Borndörfer et al. 2007, Farahani et al. 2013). With the rise of the Mobility-as-a-Service (MaaS), the integration of various transport modes, such as transit systems with shared mobility services, into unified and multimodal

networks has attracted significant attention (Shaheen and Chan 2016). These integrated systems offer considerable benefits for both riders and transit agencies by providing affordable, flexible, and convenient travel options compared to traditional transit. Reflecting this growing interest, pilot programs have been launched worldwide to investigate the effectiveness of such systems. Notable examples include the Federal Transit Administration (FTA) Mobility on Demand (MOD) Sandbox Program, which comprises eleven projects across the US (Federal Transit Administration 2023), the MARTA Reach program in Atlanta, Georgia (Van Hentenryck et al. 2023), TransLink's first Shared Mobility in Canada (Abotalebi and Petrunić 2021), and the RIDE2RAIL project in the EU (Rataj et al. 2025), demonstrating the potential and impact of the multimodal transit systems through real-world programs.

A typical multimodal transit system combines fixed-route bus and/or rail services connecting transit hubs with shared mobility services, such as ridesharing, carsharing, or bike-sharing, to transport riders from their origins to destinations, potentially via multiple intermediate hub nodes. Riders book trips online (e.g., via mobile apps), receiving pickups and drop-offs at hubs or their own locations, often with one or more bus and/or rail legs in their trip itinerary. Figure 1 provides an example of a multimodal transit system, where origins and destinations of two trips are illustrated, which can be served with routes including a combination of different travel modes. In this regard, an On-Demand Multimodal Transit System (ODMTS) is a pertinent transportation system, in which a transit agency operates on-demand shuttles along with high-frequency bus and rail services between transit hubs, addressing the first- and last-mile challenges in urban mobility (Mahéo et al. 2019, Basciftci and Van Hentenryck 2023). Such systems enhance both accessibility and convenience by bridging connectivity gaps between fixed transit hubs and origins and destinations of the riders.

Figure 1 **An example multimodal transit system with potential routes of two trips from their origins to destinations. Dashed lines represent legs that can connect origin/destinations with hubs through shared mobility services. Solid lines represent legs that can connect hubs through bus and/or rail services.**



For ensuring effective operations of the transit systems, a critical problem is to design the network between the transit hubs while considering the travel demand from different origins and destinations and rider preferences. In practice, stochastic factors such as travel time variability, fluctuating demand, and diverse user preferences critically influence transit network adoption and user satisfaction. Properly accounting for these uncertainties is essential to capture traffic congestion and rider behaviors, thereby improving system utilization. By ignoring such variabilities in network design problems, deterministic models

often risk providing suboptimal network designs that perform poorly under uncertain real-world conditions, potentially limiting ridership and overall system effectiveness. These necessitate the development of optimization frameworks that can account for systemic uncertainties and hierarchical decision-making of the transit agency and riders during network design.

To address these challenges, this paper introduces a stochastic multimodal transit network design framework using a bilevel hierarchical structure. The model captures strategic interaction between the transit agency (leader), which designs the network, and heterogeneous riders (followers), who select routes under real-time travel conditions. Our main contributions are fourfold:

1. We develop a two-stage stochastic bilevel optimization approach for designing multimodal transit systems that effectively captures heterogeneous rider preferences under travel time uncertainty. The bilevel model consists of (i) a leader problem that optimizes the transit network design by incorporating rider travel preferences, system disutilities, and travel time uncertainty; and (ii) follower problems that identify riders' most cost-efficient and convenient route choices under realized travel times.
2. We propose an exact single-level reformulation by introducing a preprocessing response search algorithm that efficiently enumerates all critical follower responses with guaranteed finite termination. These responses are integrated into the leader's problem using integer programming techniques via special ordered sets of type 1 (SOS1). Importantly, the response search algorithm runs independently across uncertainty scenarios and riders, naturally enabling efficient parallel implementation that achieves up to 13 times speedup with 16 cores compared to the serial implementation. While the proposed bilevel framework adopts an optimistic setting, selecting rider responses most favorable to the network operator's objective, the proposed single-level reformulation can be readily extended to a pessimistic setting.
3. While the proposed reformulation runs at least 10 times faster than alternative single-level reformulations in the literature on moderate-sized instances, the model size grows substantially for larger networks. To further improve efficiency and scalability, we develop a novel decomposition method that leverages a relaxed formulation with a subset of rider responses. The proposed method iteratively strengthens the relaxation with valid cutting planes and further provides computational enhancements, reducing the optimality gap from above 96% to less than 1.5% across all large instances under the same time limits.
4. Furthermore, a comprehensive case study of the Ann Arbor/Ypsilanti region in Michigan highlights real-world benefits, including up to 12% total cost savings and improved route convenience up to 7%, demonstrating the value of our stochastic bilevel framework over deterministic or single-level alternatives.

Notably, while both the single-level reformulation and the algorithms are developed for the network design problems, they can be readily extended to other bilevel optimization problems involving binary leader decisions, demonstrating broad applicability beyond the multimodal transit design context.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents the problem settings and the resulting stochastic bilevel multimodal transit network design problems with heterogeneous rider preferences and uncertain travel times. Section 4 introduces the response search algorithm, along with its properties of finite termination and enumerative completeness, and develops an equivalent mixed-integer linear programming (MILP) reformulation. Section 5 presents the decomposition algorithm, derives valid cuts, and discusses practical computational enhancements. Section 6 demonstrates the performance of the proposed approaches in real-world network instances. Finally, Section 7 concludes the paper with final remarks and future research directions.

2. Motivation and Related Literature

In this section, we review two streams of relevant literature on multimodal transit systems and stochastic bilevel optimization in Sections 2.1 and 2.2, respectively, and discuss the contributions of our paper in each section to these relevant areas.

2.1. Multimodal Transit Systems

Multimodal transit systems are increasingly recognized as vital components of modern urban mobility, integrating different transportation modes. Platforms enabling MaaS exemplify this integration, offering riders more flexible, convenient, and affordable travel options compared to traditional transit networks. Designing such integrated systems poses complex challenges, as network configurations and service provisions require careful coordination to balance planning and operational costs, and rider convenience (Stiglic et al. 2018, Liu and Ouyang 2021, Najmi et al. 2023, Mahéo et al. 2019, Dalmeijer and Van Hentenryck 2020, Auad-Perez and Van Hentenryck 2022).

Given the involvement of multiple stakeholders in multimodal transit systems, bilevel optimization has emerged as a natural and expressive framework to capture the hierarchical interactions inherent in transit network design. Here, the transit agency (leader) determines network and service designs, while riders (followers) respond by selecting services and routes based on these decisions (Yu et al. 2015, Yao and Zhang 2024). This approach explicitly accounts for differing interests of the stakeholders. For instance, Basciftci and Van Hentenryck (2020, 2023) develop bilevel models to capture latent riders who decide service adoption depending on route suggestions. Specifically, the leader problem designs transit networks and services, while the follower problem identifies routes suggested to riders, who then accept or reject these routes based on choice models.

The bilevel nature of these problems poses significant computational challenges (Kleinert et al. 2020, Fischetti et al. 2017, Jeroslow 1985). To address this, Basciftci and Van Hentenryck (2023) propose a combinatorial Benders decomposition algorithm enhanced with different cut generation approaches and valid

inequalities. Despite these advances, large-scale instances remain difficult to solve optimally. To improve scalability, Guan et al. (2024) propose a path-based single-level reformulation of this bilevel problem, while Guan et al. (2026) introduce heuristic algorithms to obtain high-quality network designs with performance guarantees that align well with the transit agency objectives. Although these studies advance both modeling and computational aspects of multimodal transit system design, they assume deterministic settings, thus ignoring inherent system uncertainties.

In practice, uncertainty in travel times and ridership demand is critical in transit planning, significantly influencing system performance and user satisfaction (Chen et al. 2011). While stochastic programming methods have been explored in transit contexts, such as bus timetabling under travel time uncertainty (Wu et al. 2015) and rapid transit design under uncertain demand (An and Lo 2016), fewer studies address these uncertainties in multimodal transit system design. Uchida et al. (2015) considers a multimodal network design problem while considering uncertainties in travel demand and road capacities due to the unexpected weather conditions impacting travel times. Luo et al. (2021) studies a Mobility-on-Demand Transit (MoDT) system that considers a combination of fixed transit systems with ride-hailing services by providing a two-stage single-level stochastic program to model the transit network while maximizing the revenue under uncertain demand. Yet, integrating stochasticity with bilevel hierarchical decision making remains a challenging and underexplored area in multimodal transit system design.

In this paper, we develop a novel stochastic bilevel optimization framework for multimodal transit network design that explicitly models heterogeneous rider preferences and accounts for uncertainties in travel times and demand. While demonstrated on ODMTS instances, our framework and solution methods are broadly applicable to multimodal transit network design problems that involve hierarchical decision-making under uncertainty. The proposed bilevel framework adopts an optimistic perspective, selecting follower responses most favorable to the network operator's objective, consistent with aforementioned literature. The optimistic problem can be justified for two reasons: the transit agency (i) can directly influence rider behavior by controlling which routes are displayed and suggested to riders, and (ii) may be able to offer small side incentives to nudge rider choices. Nevertheless, the solution methods developed in this paper can be readily extended to a pessimistic setting, where the operator hedges against the least favorable rider responses, offering a more risk-averse alternative for practical deployment.

2.2. Stochastic Bilevel Optimization

Our problem setting, pertaining to transit network design, naturally fits within a bilevel optimization framework, where the leader (the transit agency) makes upper-level network design decisions that influence the followers' (riders') route choices in the lower level. These follower decisions, in turn, depend endogenously on the network design.

Classic solution approaches often rely on single-level reformulations that exploit optimality conditions of the follower's problem (Kleinert et al. 2021, Dempe 2002). However, these reformulations typically lead

to nonlinear and nonconvex formulations due to complementarity constraints and bilinear terms, requiring linearization techniques to derive MILP reformulations. In the stochastic settings, the integration with uncertainties further compounds this complexity, as these nonlinearities are introduced for each scenario, significantly increasing model size and computational burden. Thus, scalability becomes a major challenge when applying these methods to large-scale stochastic bilevel problems (Beck et al. 2023).

While our problem inherits scenario-wise block structures typical of stochastic programming (see, e.g., Birge and Louveaux 2011, Shapiro and Xu 2008), classical decomposition methods, such as Benders decomposition (Benders 1962), cannot directly be applied. This is due to the nonconvexity of the second-stage problem with respect to the leader's decision, which fails the convex second-stage problem assumption needed for typical duality-based cuts (Henkel 2014). Efficient solution approaches for such problems still remain scarce.

Existing literature explores regularization schemes primarily for continuous leader decisions (Burtscheidt et al. 2020, Shapiro and Xu 2008), and value function-based approaches for integer leader and follower decisions under uncertainty affecting only the right-hand sides (Zhang and Özaltın 2021). However, these methods do not directly extend to the transit network design setting considered in this paper, where binary leader decisions interact with objective uncertainties at both leader and follower levels, such as travel times and ridership demand, which are critical in transit network design.

We further note that while Burtscheidt et al. (2020) incorporate on the risk-averse objectives to hedge against unfavorable outcomes, the source of adversity arises from stochastic uncertainty rather than antagonistic follower behavior, and therefore does not correspond to the pessimistic bilevel setting. In the pessimistic setting, the follower selects responses least favorable to the leader among its optimal solution(s), introducing an additional layer of optimization which renders the problem intrinsically difficult to solve (Lampariello et al. 2019). While several studies (Yanıkoglu and Kuhn 2018, Goyal et al. 2023) investigate stochastic pessimistic bilevel optimization, the combination of stochastic uncertainty with pessimistic follower behavior remains relatively unexplored.

To address this emerging and practical challenge, we propose a novel single-level reformulation that leverages critical follower responses identified through a prepossessing search algorithm, leading to a more scalable MILP reformulation compared to classical approaches. Notably, as mentioned in Section 2.1, the proposed single-level reformulation approach can be extended to the pessimistic setting. Building on this, we further develop a cutting-plane-based decomposition algorithm that iteratively strengthens a relaxed version of the single-level reformulation, yielding significant computational speedups. While designed for multimodal transit network design, our methodology can be readily extended to other bilevel problems with binary leader decisions.

3. Problem Formulation

This section presents the stochastic bilevel optimization model for the network design problem of a multimodal transit system, which is illustrated over an ODMTS. The ultimate objective of this problem is to design a bus route network that balances the interests of the transit agency, acting as the leader, and riders, corresponding as the followers. The leader and follower problems have different objectives, reflecting their respective perspectives on route cost and convenience. The model explicitly incorporates uncertainty in both travel times and demand (i.e., the number of passengers on each trip), represented through a finite set of scenarios from historical data or generated via sample average approximation techniques (e.g., Shapiro and Xu 2008). Specifically, the leader designs the network by determining which bus legs to open for the ODMTS service, considering fixed infrastructure costs and the expected operating costs to serve all trips, where these costs include the cost of operating the transit network with buses and on-demand shuttles, respectively. Each follower, defined by a trip under a particular uncertain scenario, is then suggested a route, from its origin to destination, which can involve on-demand shuttles, buses, or a combination of both, with the goal of minimizing a weighted measure of cost and convenience.

Section 3.1 details the stochastic bilevel model with multiple followers, and Section 3.2 provides alternative single-level reformulations of the proposed model before introducing a novel single-level reformulation in Section 4. For clarity, boldface letters are used to denote matrices and vectors of variables or values.

3.1. Stochastic Bilevel Model

The public transportation network is represented by a set of nodes N, which corresponds to a set of stops where riders can start and end their trips by being picked up and dropped off by on-demand shuttles. A subset of these nodes, $H \subseteq N$, is designated as potential transit hub locations, where buses operate between open hubs that are determined by the resulting network design, and the set $\mathcal{H} = \{(h,l) : h,l \in H\}$ denotes all potential bus legs between hubs. In the remainder of this paper, we refer to them as bus legs and hub legs interchangeably. Each trip $r \in T$ is characterized by its origin o^r and destination d^r stops, and the number of riders taking that trip $p^{r,\omega}$ under scenario $\omega \in \Omega$. Riders of each trip $r \in T$ are suggested with routes which include either a direct shuttle from o^r to d^r, or utilize available bus legs as intermediate segments on the route, which can result in a multimodal trip utilizing on-demand shuttles and multiple legs between hubs. The distance between nodes $i,j \in N$ is denoted by $e_{i,j}$ and the travel time under each scenario $\omega \in \Omega$ is denoted by $t^{\omega}_{i,j}$. To account for the trade-off between affordability and convenience, we adopt a convex combination approach following Basciftci and Van Hentenryck (2023) and Guan et al. (2024). Specifically, to capture the different perspectives of the transit agency (leader) and the riders (followers) in evaluating the cost and convenience of suggested routes, we introduce distinct parameters $\theta^l, \theta^f \in [0,1]$. The trade-off is expressed as a convex combination in their respective objective functions, where θ^l, θ^f weight convenience, measured by travel time, and $1-\theta^l, 1-\theta^f$ weight affordability, measured by travel distance. Thus, larger values of θ^l, θ^f reflect a stronger preference for convenience over affordability.

The leader problem considers the operating costs of buses, where the weighted fixed cost of opening a bus leg between hubs $h, l \in H$ is calculated by $\beta_{h,l} = (1-\theta^l)nae_{h,l}$, where n denotes the number of buses operating in each open leg within the planning horizon, and a indicates the cost of using a bus per mile. The objectives of the leader and follower problems then consider the weighted cost and convenience of using bus legs and on-demand shuttle legs, which depend on each scenario realization by considering travel time uncertainty. As the fixed cost of operating bus legs is considered in the leader problem, the weighted convenience of using the bus leg between hubs h, l is calculated by $\tau^{\omega}_{h,l}(\theta) = \theta(t^{\omega}_{h,l} + W)$, where W is the average waiting time of a bus. We note that beyond waiting time, transfer-related penalties, such as the convenience of transfer paths and the availability of real-time information, can significantly influence users' willingness to accept a route (Garcia-Martinez et al. 2018, Guimarães and Oliveira-Neto 2026). Since such information is typically unavailable at the network design stage, the transfer penalty is approximated by the average waiting time. If more detailed information becomes available, the waiting time parameter W can be extended to an arc-dependent parameter W_{hl} for transfers between hub legs h and l. The weighted cost and convenience of using the on-demand shuttle leg between nodes $i, j \in N$ is denoted by $\gamma^{\omega}_{i,j}(\theta) = (1-\theta)me_{i,j} + \theta t^{\omega}_{i,j}$, where m indicates the cost of using a shuttle per mile.

To construct the optimization model, we define the binary variable $z_{h,l}$ for the leader problem to indicate whether a bus leg between hubs $h, l \in H$ is open. For each trip $r \in T$ and scenario $\omega \in \Omega$, we define binary variables $x^{r,\omega}_{h,l}$ and $y^{r,\omega}_{i,j}$ for the corresponding follower problems, which indicate whether the route suggested from o^r to d^r includes the bus leg between hubs h, l and shuttle legs between nodes i, j, respectively. The resulting optimization model can be presented as follows:

$$\min_{\boldsymbol{z}} \sum_{h,l \in H} \beta_{h,l} z_{h,l} + \sum_{\omega \in \Omega} \rho^{\omega} \sum_{r \in T} p^{r,\omega} f^{r,\omega}(\boldsymbol{z}) \tag{1a}$$

$$s.t. \quad \sum_{l \in H} z_{h,l} = \sum_{l \in H} z_{l,h}, \quad \forall h \in H, \tag{1b}$$

$$z_{h,l} \in \{0,1\}, \quad \forall h, l \in H, \tag{1c}$$

where ρ^{ω} indicates the probability of scenario $\omega \in \Omega$. Constraints (1b) ensure weak connectivity in the network by guaranteeing that the number of incoming and outgoing open bus legs is equal to each other for each hub.

The objective of the leader problem (1a) considers the weighted fixed cost of operating bus legs and the weighted cost and convenience of trip r in scenario ω as perceived by the transit agency under the network design $\boldsymbol{z}$. This is characterized through the following *disutility function* of the leader piroblem, which incorporates key quality-of-service (QoS) attributes associated with each leg, including travel time, monetary cost, and waiting time, and reflects the trade-off between cost and convenience:

$$f^{r,\omega}(\boldsymbol{z}) = \min_{\boldsymbol{x},\boldsymbol{y} \in \Xi^{r,\omega}(\boldsymbol{z})} \sum_{h,l \in H} \tau^{\omega}_{h,l}(\theta^l) x^{r,\omega}_{h,l} + \sum_{i,j \in N} \gamma^{\omega}_{i,j}(\theta^l) y^{r,\omega}_{i,j}. \tag{2}$$

Here, $\Xi^{r,\omega}(\boldsymbol{z})$ indicates the set of optimal solutions to the follower problem under the network design $\boldsymbol{z}$, where the corresponding follower problem for trip r and scenario ω is formulated as

$$g^{r,\omega}(\boldsymbol{z}) = \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{h,l\in H} \tau^{\omega}_{h,l}(\theta^f) x^{r,\omega}_{h,l} + \sum_{i,j\in N} \gamma^{\omega}_{i,j}(\theta^f) y^{r,\omega}_{i,j} \tag{3a}$$

$$\text{s.t.} \sum_{h\in H:(h,l),(l,h)\in\mathcal{H}} (x^{r,\omega}_{i,h} - x^{r,\omega}_{h,i}) + \sum_{j\in N} (y^{r,\omega}_{i,j} - y^{r,\omega}_{j,i}) = \begin{cases} 1 & \text{if } i = o^r \\ -1 & \text{if } i = d^r \\ 0 & \text{otherwise} \end{cases}, \quad \forall i \in N, \tag{3b}$$

$$x^{r,\omega}_{h,l} \le z_{h,l}, \quad \forall h,l \in H, \tag{3c}$$

$$x^{r,\omega}_{h,l} \in \{0,1\},\ \forall h,l \in H, \quad y^{r,\omega}_{i,j} \in \{0,1\},\ \forall i,j \in N. \tag{3d}$$

The objective of the follower problem (3a) considers the weighted cost and convenience of using bus and on-demand shuttle legs perceived by the riders, where $g^{r,\omega}(\boldsymbol{z})$ corresponds to the disutility function of the follower problem. Constraints (3b) ensure the flow balance for the bus and shuttle legs used in satisfying the path of trip r. Constraints (3c) guarantee that only open bus legs are used in the resulting path. We note that the constraint matrix of the follower problem (3) is totally unimodular, which indicates that the binary variables $\boldsymbol{x}, \boldsymbol{y}$ can be relaxed as continuous variables. This observation is helpful in obtaining single-level reformulations and designing solution algorithms for the resulting bilevel problem.

REMARK 1. We note that the proposed bilevel problem (1) adopts an optimistic formulation, where the followers select, among multiple optimal solutions (if any), one that is favorable to the leader's objective. In contrast, a risk-averse leader may consider a pessimistic formulation, where the objective of problem (2) is maximized instead. Although this paper primarily focuses on the optimistic formulation, the proposed response search algorithm and single-level reformulation can be readily extended to pessimistic formulations as well. The corresponding details are discussed separately with the algorithms and reformulation in Section 4.

REMARK 2. We note that for the special case $\theta^l = \theta^f$, the proposed bilevel problem reduces to a single-level problem as the disutility functions of the leader and the follower problems become equivalent. Although this case significantly reduces the computational complexity of the problem, it enforces the leader and follower problems to have the same goal, which might not be realistic for capturing the behaviors of the transit agency and riders. To this end, we propose a bilevel formulation that allows the leader and follower problems to have different perceptions of the paths suggested. Thus, the value of θ^f reflects followers' trade-off between monetary cost, related to distance, and time cost, which are essential to users' choice behavior in multimodal transportation systems (Kreutzberger 2008, Boarnet et al. 2024). We further note that the parameter θ^f can be selected differently for the follower problem of each trip r to consider various characteristics of riders associated with these trips. This is examined in the case study in Section 6.3, where the riders of the trips are classified with respect to their income classes and their corresponding preferences on cost and convenience are taken into account accordingly. In addition to this income level based parameter

selection, further characteristics of the riders can be integrated into this process in case corresponding data sets are available on user preferences, such as survey data (see, e.g., Vaidya and Kumar 2006, Zhu et al. 2025).

We note that transit network design encompasses a broad range of interrelated decisions, including line planning, frequency setting, routing, fleet sizing, and passenger assignment, which can be classified into strategic, tactical, and operational levels, and addressed sequentially (see, e.g., Durán-Micco and Vansteenwegen 2022). To this end, this paper focuses on strategic planning decisions which encapsulates a practically relevant scenario for the design of multimodal transit systems. For instance, a transit operator can seek to adopt a multimodal transit system built upon an *existing* fixed-route transit network, where routes, service frequency, and fleet sizing are already in place. In this context, the operator's primary decision is to identify which set of transit legs to incorporate into the multimodal system, which is a natural first step in a phased planning process. Subsequent decisions, such as fleet sizing and frequency adjustments, can be revisited and optimized in later planning stages once the network structure is established. We also note that as the multimodal systems become more widely adopted, a joint optimization across all decision levels may yield more efficient system designs, which is a promising direction for future research.

3.2. Single-level Reformulations

To solve the resulting bilevel problem, we first explore two widely-used and classic single-level reformulations (e.g., Zare et al. 2019, Dempe and Zemkoho 2020): (i) a Karush-Kuhn-Tucker (KKT)-based reformulation, which replaces the follower's problem with its optimality conditions and linearizes complementarity conditions using big-M constraints; and (ii) a strong-duality-based reformulation, which replaces complementarity conditions with a single strong duality constraint. Their detailed formulations are discussed in Appendix A. However, both methods face scalability challenges for even moderately sized instances, as demonstrated in Section 6.2.

4. Response Search Algorithm and Another Single-level Formulation

To address the computational challenges arising from the reformulations of the proposed bilevel stochastic problem, in this section, we propose a novel equivalent single-level reformulation that incorporates outcomes from a follower response search algorithm. This preprocessing algorithm identifies representative responses from the followers, corresponding to their routes from their origins to destinations, under various network designs suggested by the leader. Specifically, for each trip and scenario, the algorithm generates a set of restricted arcs that are not permitted in the leader's network design, and computes the corresponding rider paths and resulting disutilities of the leader and rider. Section 4.1 presents the response search algorithm and proves its key properties, including finite termination guarantee and complete enumeration of all possible rider responses. Based on the representative responses identified by the search algorithm, we model follower behavior using integer programming techniques and derive an equivalent single-level MILP reformulation in Section 4.2.

4.1. Response Search Algorithm

In this section, we introduce our response search algorithm and prove its properties, including finite termination and complete enumeration of all possible rider responses, which are critical for deriving the single-level reformulation of our problem in the next section.

To begin with, for each trip r and scenario ω, Algorithm 1 starts with no restricted arcs. That is, the network is fully accessible and $\mathcal{R}_1 = \emptyset$, where the set of restricted arcs at iteration n is denoted by $\mathcal{R}_n$. At iteration n, the follower's problem $g^{r,\omega}(\boldsymbol{z})$ is solved using $\boldsymbol{z} = \boldsymbol{z}^n$, where $z^n_{h,l} = 0$ for $(h,l) \in \mathcal{R}_n$ and $z^n_{h,l} = 1$ otherwise. Let $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ denote the follower's optimal solution using a set of bus legs $B_n \subset \mathcal{H} \backslash \mathcal{R}_n$. In the case where there are alternative optimal solutions for the follower problem, the algorithm chooses the one that achieves the best (lowest) leader's disutility. The process recursively explores further arc exclusions by adding an arc $(h,l) \in B_n$ to $\mathcal{R}_n$, creating a new restricted arc set $\mathcal{R}^{h,l}_n = \mathcal{R}_n \cup \{(h,l)\}$. The resulting disutilities of responses $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ in the leader's and follower's problems are recorded in sets $L^{r,\omega}$ and $F^{r,\omega}$. The process continues until all arc restriction subsets have been explored. The detailed algorithm is presented in Algorithm 1.

In Line 6, when the follower problem (3) admits multiple optimal solutions, a solution favorable to the leader's objective in (1) can be obtained using a weighted sum approach (Sherali and Soyster 1983). Specifically, one minimizes the follower objective plus a small weight multiplied by the leader objective over Constraints (3a)-(3d). For a pessimistic formulation of the leader problem, the follower objectives should instead be minimized minus a small weight times the leader objective.

Algorithm 1 A response search algorithm

1: **Input**: Trip r and scenario ω.
2: **Initialization**: A set of restricted arcs $\mathcal{R}_1 = \emptyset$. Set $\mathcal{Q} = \{\mathcal{R}_1\}$. A set of arc sets used in each response $\mathcal{B}^{r,\omega} = \emptyset$. A set of leader disutilities $L^{r,\omega} = \emptyset$ and a set of follower disutilities $F^{r,\omega} = \emptyset$ associated with responses. Set iteration $n = 1$.
3: **while** $\mathcal{Q}$ is not empty **do**
4: Select a set $\hat{\mathcal{R}}$ of arcs from $\mathcal{Q}$ and remove it from $\mathcal{Q}$. Let $\mathcal{R}_n = \hat{\mathcal{R}}$.
5: Let $\boldsymbol{z}^n$ be such that $z^n_{h,l} = 0$, for all $(h,l) \in \mathcal{R}_n$ and $z^n_{h,l} = 1$ otherwise.
6: Solve the follower's problem $g^{r,\omega}(\boldsymbol{z})$ in (3a)-(3d) with $\boldsymbol{z} = \boldsymbol{z}^n$. Denote an optimal solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$, the optimal follower's disutility value $F = g^{r,\omega}(\boldsymbol{z}^n)$. Calculate the corresponding optimal leader's disutility value $L = f^{r,\omega}(\boldsymbol{z}^n)$. Note that when alternative optimal solutions exist, select the one that achieves the lowest leader's disutility.
7: According to the optimal solution $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$, find out the set of bus legs used in the follower's solution as $B_n = \{(h,l) \in \mathcal{H} : \hat{x}_{h,l} = 1\}$.
8: Update $\mathcal{B}^{r,\omega} \leftarrow \mathcal{B}^{r,\omega} \cup \{B_n\}$. ▷ Record the set of arcs used
9: Update $L^{r,\omega} \leftarrow L^{r,\omega} \cup \{L\}$, $F^{r,\omega} \leftarrow F^{r,\omega} \cup \{F\}$. ▷ Record the leader and follower disutilities
10: **if** B_n is not empty **then**
11: **for** every $(h,l) \in B_n$ **do**
12: $\mathcal{R}^{h,l}_n = \mathcal{R}_n \cup \{(h,l)\}$. ▷ Create a new set of restricted arcs
13: **if** $\mathcal{R}^{h,l}_n$ is not explored before **then**
14: $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\mathcal{R}^{h,l}_n\}$.
15: $n \leftarrow n + 1$.
16: **Output**: The responses $\mathcal{B}^{r,\omega}$, the leader disutilities $L^{r,\omega}$, and the follower disutilities $F^{r,\omega}$.

REMARK 3. Algorithm 1 does not rely on specific structural properties of the follower problem beyond the binary nature of the leader's decisions that defines the follower's feasible region. In the context of the multimodal transit network design problem, the follower solves a shortest path problem, which can be efficiently solved using numerical algorithms such as Dijkstra's algorithm. Furthermore, since the response search algorithm runs independently for every trip and scenario, it naturally supports parallel implementation. The computational advantages of this parallelization are demonstrated in Section 6.2.

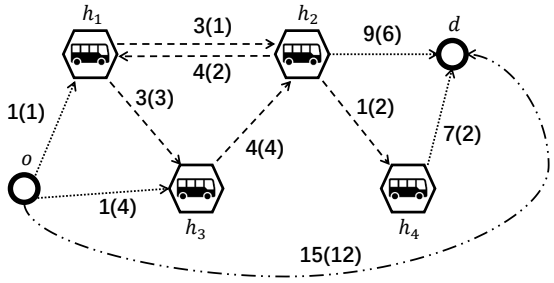
REMARK 4. The resulting responses can be further reduced by excluding restricted arc sets for which enforcing $z_{h,l} = 0$ on the arcs in the restricted arc set leads to violation of the weak connectivity Constraints (1b), resulting in no feasible network designs. To ensure the feasibility, a feasibility check procedure can be added when selecting a restricted arc set $\hat{\mathcal{R}}$ in Line 4.

To illustrate how Algorithm 1 works, we present the following example:

EXAMPLE 1 (A 4-HUB NETWORK). We demonstrate how Algorithm 1 works on a 4-hub network under a given trip with origin o and destination d and a specific scenario realization, shown in Figure 2.

Here, for illustration purposes, we only consider a subset of bus legs and on-demand shuttle legs that can be used to satisfy this trip. Specifically, we let $\mathcal{H} = \{(h_1, h_2),\ (h_2, h_1),\ (h_2, h_4),\ (h_3, h_2), (h_1, h_3)\}$ by considering these arcs as potential bus legs to be considered for the network design. Additionally, we consider the following legs as potential on-demand shuttle legs $\{(o, h_1), (o, h_3), (h_2, d), (h_4, d), (o, d)\}$.

Figure 2 **A 4-hub network with origin and destination of a trip. Each arc is labeled with two numbers: the first represents the follower's objective coefficient, while the number in parentheses represents the leader's objective coefficient.**



We provide the detailed iterations of the response search algorithm, and Table 1 summarizes the key outputs at each iteration n. Column "Restricted arcs $\mathcal{R}_n$" shows the set of arcs currently restricted by the leader at iteration n. Column "Response B_n" indicates the set of bus legs used by the follower in their optimal solution under the current restriction $\mathcal{R}_n$. Columns "Leader L" and "Follower F" present the objective values of the leader and follower, respectively, associated with the follower solution.

Iteration $n = 1$: No arcs are restricted ($\mathcal{R}_1 = \emptyset$) and the network is fully accessible. The follower selects the lowest-disutility path $o - h_1 - h_2 - h_4 - d$ with bus legs $B_1 = \{(h_1, h_2),\ (h_2, h_4)\}$, yielding follower objective $F = 12$ and leader objective $L = 6$. Two new restricted arc sets are created: $\mathcal{R}_2 = \{(h_1, h_2)\}$ and $\mathcal{R}_3 = \{(h_2, h_4)\}$.

Iteration $n = 2$: With $\mathcal{R}_2 = \{(h_1, h_2)\}$, the follower selects path $o - h_3 - h_2 - h_4 - d$ using bus legs $B_2 = \{(h_3, h_2),\ (h_2, h_4)\}$, achieving $F = 13$ and $L = 12$. Two new restricted arc sets are created: $\mathcal{R}_4 = \mathcal{R}_2 \cup \{(h_3, h_2)\} = \{(h_1, h_2),\ (h_3, h_2)\}$ and $\mathcal{R}_5 = \mathcal{R}_2 \cup \{(h_2, h_4)\} = \{(h_1, h_2),\ (h_2, h_4)\}$.

Iteration $n = 3$: With $\mathcal{R}_3 = \{(h_2, h_4)\}$, the follower selects path $o - h_1 - h_2 - d$, using bus leg $B_3 = \{(h_1, h_2)\}$, with $F = 13$ and $L = 8$. The resulting new restricted arc set coincides with $\mathcal{R}_5$.

Iteration $n = 4$: With $\mathcal{R}_4 = \{(h_1, h_2),\ (h_3, h_2)\}$, the follower's lowest-disutility path is a direct trip from the origin to the destination, resulting in a follower objective $F = 15$ and a leader objective $L = 12$. Since no bus legs are used ($B_6 = \emptyset$), and no new restricted arc sets are created.

Iteration $n = 5$: With $\mathcal{R}_5 = \{(h_1, h_2),\ (h_2, h_4)\}$, the follower travels via $o - h_3 - h_2 - d$ using $B_5 = \{(h_3, h_2)\}$, with $F = 14$ and $L = 14$. One new restricted arc set is created $\mathcal{R}_6 = \{(h_1, h_2),\ (h_2, h_4),\ (h_3, h_2)\}$.

Iteration $n = 6$: With $\mathcal{R}_6 = \{(h_1, h_2),\ (h_2, h_4),\ (h_3, h_2)\}$, as $\mathcal{R}_4 \subset \mathcal{R}_6$, $B_6 = B_4 = \emptyset$ with $F = 15$ and $L = 12$. No new restricted arc sets are created.

Table 1 Output of Algorithm 1 on the 4-hub network

n	Restricted arcs $\mathcal{R}_n$	Responses B_n	Disutility	
			Leader L	Follower F
1	$\emptyset$	$(h_1,h_2),\ (h_2,h_4)$	6	12
2	(h_1,h_2)	$(h_3,h_2),\ (h_2,h_4)$	12	13
3	(h_2,h_4)	(h_1,h_2)	8	13
4	$(h_1,h_2),\ (h_3,h_2)$	$\emptyset$	12	15
5	$(h_1,h_2),\ (h_2,h_4)$	(h_3,h_2)	14	14
6	$(h_1,h_2),\ (h_2,h_4),\ (h_3,h_2)$	$\emptyset$	12	15

Next, we establish two key properties of Algorithm 1: finite termination and complete enumeration. These properties are fundamental to the equivalent single-level reformulation developed in the following section.

THEOREM 1 **(Finite termination)**. *Algorithm 1 terminates after a finite number of iterations.*

Proof of Theorem 1: We prove finite termination by showing that the total number of restricted arc sets $\mathcal{R}_n$ generated and explored by Algorithm 1 is finite. Recall that $\mathcal{H}$ denotes the set of all hub arcs (bus legs). At each iteration, the algorithm selects a restricted arc set $\mathcal{R}_n$ from the queue $\mathcal{Q}$ and solves the follower's problem with network design $\boldsymbol{z}^n$, where arcs in $\mathcal{R}_n$ are forced to be unused, i.e., $z^n_{h,l} = 0$ for $(h,l) \in \mathcal{R}_n$. Based on the optimal follower solution, the set $B_n \subset \mathcal{H} \backslash \mathcal{R}_n$ of used hub arcs is identified in the follower problem. For each such arc $(h,l) \in B_n$, a new restricted arc set ${\mathcal{R}_n}^{h,l} = \mathcal{R}_n \cup \{(h,l)\}$ is created and added to the queue $\mathcal{Q}$.

Now, consider a restricted arc set $\hat{\mathcal{R}}$ of size $m < |\mathcal{H}|$, and let $B \subseteq \mathcal{H} \backslash \hat{\mathcal{R}}$ be the corresponding follower response. Since arcs in $\hat{\mathcal{R}}$ are restricted, the follower can only use arcs in $\mathcal{H} \backslash \hat{\mathcal{R}}$, so $|B| \leq |\mathcal{H}| - m$. From $\hat{\mathcal{R}}$, the algorithm generates at most $|B|$ restricted arc sets of size $m+1$, one for each arc in B. Let K denote the number of distinct restricted arc sets of size m; note that $K = 1$ when $m = 0$, corresponding to the initial iteration. Then the total number of restricted arc sets of size $m+1$ generated from those is at most $\sum_{k=1}^{K} |B^k| \leq K(|\mathcal{H}| - m)$, where B^k is the follower response associated with the k-th restricted arc set of size m. Since $|\mathcal{H}|$ is finite, and the algorithm proceeds by incrementally constructing restricted arc sets of size at most $|\mathcal{H}|$, the number of such sets is finite. □

REMARK 5. Note that the worst-case number of iterations can be as large as $2^{|\mathcal{H}|}$ when riders can potentially use all available hub legs. Such a worst-case scenario would require a network structure in which, under every possible network design, each rider has many alternative routes with similar objective values. In other words, removing or restricting a subset of hub legs would still leave numerous routes with nearly identical utilities. However, such situations are unlikely to arise in practice, where only a limited subset of routes typically provides competitive travel costs. Consequently, the response search algorithm usually terminates after exploring only a small fraction of the theoretical worst-case possibilities. For example, in Example 1, although the theoretical worst-case number of iterations for the studied 4-hub network is $2^{|\mathcal{H}|} = 32$, the algorithm completes after only six iterations. Similar early termination behavior is also observed in the

computational experiments discussed in Section 6.2. This is mainly due to two reasons: (i) the algorithm only explores hub legs that are relevant to the follower's routing decisions, and (ii) only a limited subset of hub legs impacts the follower's optimal response.

THEOREM 2 **(Complete enumeration)**. *Given any (feasible) network design, the set of all possible rider (i.e., follower) responses for trip r under scenario ω is contained in the response set $\mathcal{B}^{r,\omega}$ generated by Algorithm 1.*

Proof of Theorem 2: Suppose there exists a feasible network design $\boldsymbol{z}^*$ corresponding to a restricted arc set $\mathcal{R}^*$, i.e., $z^*_{h,l} = 0$ for $(h,l) \in \mathcal{R}^*$, and an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ to the follower's problem $g^{r,\omega}(\boldsymbol{z}^*)$, such that the set of hub arcs used in this solution, denoted $B^* \subseteq \mathcal{H} \backslash \mathcal{R}^*$. We show that this response set B^* is always included in $\mathcal{B}^{r,\omega}$.

We proceed by induction on the size of $m = |\mathcal{R}^*|$ of the restricted arc set.

- Base case $m = 0$: When $\mathcal{R}^* = \emptyset$, the network is fully accessible. The algorithm initializes with this network configuration and solves the follower problem, obtaining B_1, which is added to $\mathcal{B}^{r,\omega}$.
- Inductive step: Assume that for all restricted arc sets $\mathcal{R}^{m-1} \subset \mathcal{H}$ of size $m-1$, the corresponding follower bus leg usage B^{m-1} is included in $\mathcal{B}^{r,\omega}$. Now consider the restricted set $\mathcal{R}^* = \mathcal{R}^{m-1} \cup \{h_m\}$ with one restricted arc set $\mathcal{R}^{m-1}$ of size $m-1$ and let B^{m-1} denote the corresponding follower response. Three cases arise:
  1. $B^{m-1} \neq \emptyset$ and $h_m \in B^{m-1}$: Closing hub leg h_m removes an arc used in the current response B^{m-1}. This triggers the algorithm (Lines 10–14) to generate $\mathcal{R}^*$ as a new restricted arc set added to the queue $\mathcal{Q}$, and it is later explored. The resulting response set B^* is then added $\mathcal{B}^{r,\omega}$.
  2. $B^{m-1} \neq \emptyset$ and $h_m \notin B^{m-1}$: Since the additional restricted arc h_m is not used in B^{m-1}, the same follower response remains feasible and optimal. Thus, $B^* = B^{m-1} \in \mathcal{B}^{r,\omega}$ by the inductive assumption.
  3. $B^{m-1} = \emptyset$: The follower cannot travel under $\mathcal{R}^{m-1}$, and closing an additional arc h_m cannot create a new feasible follower solution. So $B^* = B^{m-1} = \emptyset \in \mathcal{B}^{r,\omega}$.

In all three cases, the response set $B^* \subset \mathcal{B}^{r,\omega}$. By induction, for any restricted arc sets $\mathcal{R}^* \subset \mathcal{H}$, the corresponding follower response B^* is included in $\mathcal{B}^{r,\omega}$. Finally, since every feasible network design $\boldsymbol{z}^*$ corresponds to some such restricted arc set $\mathcal{R}^*$, this proves that Algorithm 1 exhaustively enumerates all optimal follower responses for all network designs. □

This enumeration completeness arises from the recursive construction of the restricted arc sets. At each iteration n, the algorithm identifies a feasible path $B \subseteq \mathcal{H} \backslash \mathcal{R}_n$ used by the follower, and systematically explores all arc exclusion combinations that may alter the follower's path selection. This process ensures that for every feasible network design decision z, there exists a corresponding follower response $B \in \mathcal{B}^{r,\omega}$ captured in the output. In this way, the algorithm covers the entire decision space of the follower across all relevant leader designs, enabling an exhaustive and compact representation of the leader-follower interaction in the next section.

4.2. Response Processing and Single-level MILP Reformulation

To derive the single-level MILP reformulation, we first introduce Algorithm 2, which processes the output of the response search algorithm. Algorithm 2 sorts the responses $\mathcal{B}^{r,\omega}$ in ascending order of the follower disutility values $F^{r,\omega}$. In the case of ties, responses are further sorted by the leader disutility values $L^{r,\omega}$ in ascending order under the optimistic leader. For a pessimistic leader, the responses are instead sorted in descending order by the leader disutility values. After sorting, duplicate responses are removed to ensure a compact representation.

Algorithm 2 Response sorting

1: **Input**: Response set $\mathcal{B}^{r,\omega}$, leader disutility values $L^{r,\omega}$, and follower disutility values $F^{r,\omega}$.
2: Create a tuple set $S = \{(B_i^{r,\omega}, L_i^{r,\omega}, F_i^{r,\omega}) \mid i = 1, ..., |\mathcal{B}^{r,\omega}|\}$.
3: Sort S by ascending $F_i^{r,\omega}$ (follower disutility value). In the case of ties, sort by leader disutility $L_i^{r,\omega}$.
4: Remove duplicate tuples from S.
5: Extract the components from the sorted and deduplicated set $\hat{\mathcal{B}}^{r,\omega} = \{B_i^{r,\omega} | (B_i^{r,\omega}, L_i^{r,\omega}, F_i^{r,\omega}) \in S\}$, $\hat{L}_i^{r,\omega} = \{L_i^{r,\omega} | (B_i^{r,\omega}, L_i^{r,\omega}, F_i^{r,\omega}) \in S\}$, $\hat{F}_i^{r,\omega} = \{F_i^{r,\omega} | (B_i^{r,\omega}, L_i^{r,\omega}, F_i^{r,\omega}) \in S\}$.
6: **Output**: Sorted and deduplicated sets $\hat{\mathcal{B}}^{r,\omega}$, $\hat{L}^{r,\omega}$, and $\hat{F}^{r,\omega}$.

As an illustration, we apply Algorithm 2 to the output from Example 1.

EXAMPLE 2 (A 4-HUB NETWORK (CONTINUED)).

Step 1: Create the tuple set

$$S = \{(B_1, 6, 12),\ (B_2, 12, 13),\ (B_3, 8, 13),\ (B_4, 12, 15),\ (B_5, 14, 14),\ (B_6, 12, 15)\},$$

where each tuple contains a follower response B_i, the leader disutility L_i, and the follower disutility F_i.

Step 2: Sort the tuples in S by ascending follower disutility F_i,

$$S = \{(B_1, 6, 12),\ (B_2, 12, 13),\ (B_3, 8, 13),\ (B_5, 14, 14),\ (B_4, 12, 15),\ (B_6, 12, 15)\}.$$

For ties in follower disutility, between $(B_2, 12, 13)$ and $(B_3, 8, 13)$, and between $(B_4, 12, 15)$ and $(B_6, 12, 15)$, sort by ascending leader disutility L_i,

$$S = \{(B_1, 6, 12),\ (B_3, 8, 13),\ (B_2, 12, 13),\ (B_5, 14, 14),\ (B_4, 12, 15),\ (B_6, 12, 15)\}.$$

Step 3: Remove duplicates to obtain the final sorted and deduplicated response set shown in Table 2.

Next, we reformulate the original stochastic bilevel problem (1a)-(1c) as a single-level MILP, based on the sorted and deduplicated set of follower responses $\hat{\mathcal{B}}^{r,\omega}$ generated via Algorithm 2. For each response $B \in \hat{\mathcal{B}}^{r,\omega}$, and for every trip r and scenario ω, we introduce a binary variable $\chi_B^{r,\omega}$ to indicate whether all bus legs used in response B are available under the current network design $\boldsymbol{z}$. Specifically,

$$\chi_B^{r,\omega} \geq \sum_{(h,l)\in B} z_{h,l} - |B| + 1,\ B \in \hat{\mathcal{B}}^{r,\omega},\ r \in T,\ \omega \in \Omega \tag{4a}$$

$$\chi_B^{r,\omega} \leq z_{h,l},\ (h,l) \in B,\ B \in \hat{\mathcal{B}}^{r,\omega},\ r \in T,\ \omega \in \Omega \tag{4b}$$

$$\chi_B^{r,\omega} \in \{0,1\},\ \ B \in \hat{\mathcal{B}}^{r,\omega},\ r \in T,\ \omega \in \Omega. \tag{4c}$$

Table 2 Sorted and deduplicated sets obtained from Algorithm 2

i	Response B_i	Disutility	
		Leader L	Follower F
1	$(h_1,h_2),\ (h_2,h_4)$	6	12
2	(h_1,h_2)	8	13
3	$(h_3,h_2),\ (h_2,h_4)$	12	13
4	(h_3,h_2)	14	14
5	$\emptyset$	12	15

Constraints (4a)-(4c) ensure that $\chi_B^{r,\omega}=1$ if and only if all required bus legs in response B are open.

Given a network design, there can be multiple responses in which the required bus legs are open. Next, we define binary variables $\phi_B^{r,\omega}$ for each response $B\in\hat{\mathcal{B}}^{r,\omega}$ to indicate whether response B is selected as the optimal follower response for trip r under scenario ω:

$$\phi_B^{r,\omega}\le\chi_B^{r,\omega},\ B\in\hat{\mathcal{B}}^{r,\omega},\ r\in T,\ \omega\in\Omega \tag{5a}$$

$$\sum_{B\in\hat{\mathcal{B}}^{r,\omega}}\phi_B^{r,\omega}=1,\ r\in T,\ \omega\in\Omega \tag{5b}$$

$$\chi_{B^i}^{r,\omega}-\phi_{B^i}^{r,\omega}\le 1-\phi_{B^j}^{r,\omega},\ B^i,B^j\in\hat{\mathcal{B}}^{r,\omega},\ i<j,\ r\in T,\ \omega\in\Omega \tag{5c}$$

$$\phi_B^{r,\omega}\in\{0,1\},\ B\in\hat{\mathcal{B}}^{r,\omega},\ r\in T,\ \omega\in\Omega. \tag{5d}$$

Constraints (5a) ensure a response is only selected if all required arcs are available. Constraints (5b) enforce that exactly one response is selected per trip and scenario. Constraints (5c) together with Constraints (5b) ensure that among all feasible responses, the one with the lowest follower disutility value (according to the sorted order) is preferred. Specifically, if two responses B^i and B^j are feasible ($\chi_{B^i}^{r,\omega}=\chi_{B^j}^{r,\omega}=1$) and $i<j$, then only the one with the smaller follower disutility is preferred. Lastly, we introduce a lower bound $\delta^{r,\omega}$ on the leader's disutility value for each trip r and scenario ω. For each selected response $B\in\hat{\mathcal{B}}^{r,\omega}$, the leader disutility $\delta^{r,\omega}=L_B^{r,\omega}$ is enforced only if B is selected, i.e., $\phi_B^{r,\omega}=1$ via the following constraints

$$\delta^{r,\omega}\ge L_B^{r,\omega}-M^{r,\omega}(1-\phi_B^{r,\omega}),\ B\in\hat{B}^{r,\omega},\ r\in T,\ \omega\in\Omega, \tag{6}$$

where $M^{r,\omega}$ is a sufficiently large constant such as $M^{r,\omega}=\max\{L_B^{r,\omega}\mid B\in\hat{B}^{r,\omega}\}$. Putting all components together, we are ready to present the equivalent single-level MILP reformulation.

THEOREM 3 **(Equivalent single-level reformulation).** *The stochastic bilevel network design problem* (1a)-(1c) *admits an equivalent single-level MILP reformulation*

$$\min_{z\in Z,\delta,\chi,\phi}\left\{\sum_{h,l\in H}\beta_{h,l}z_{h,l}+\sum_{\omega\in\Omega}\rho^{\omega}\sum_{r\in T}p^{r,\omega}\delta^{r,\omega}:\ (4a)-(4c),(5a)-(5d),(6)\right\}, \tag{7}$$

where $Z=\{\boldsymbol{z}:\ (1b)-(1c)\}$ *is the leader's feasible region.*

Proof of Theorem 3: The equivalence follows from the enumeration completeness of the response search algorithm (Algorithm 1) and the ascending order of the responses imposed by Algorithm 2. □

REMARK 6. As illustrated in Example 1, the practical number of follower responses (or iterations) required is significantly fewer than the worst-case bound $2^{|\mathcal{H}|}$. The responses reported in Table 2 are further reduced through a deduplication step in Algorithm 2 that removes duplicate responses. Consequently, the number of constraints and binary variables introduced in the single-level reformulation remains compact.

REMARK 7. The proposed MILP reformulation derived from the response search algorithm does not rely total unimodularity of the follower problems, which is, however, required by both the KKT- and strong-duality-based reformulations. Instead, it relies solely on the follower problem admitting a finite number of potential optimal responses, which holds under more general conditions. This flexibility enables several practical extensions of the proposed framework. For instance, additional constraints on user route choice, such as an upper bound on the number of transfers permitted per route, implementable via a transfer-expanded graph (Dalmeijer and Van Hentenryck 2020), can be incorporated into the follower's problem. Similarly, extending the leader's problem to treat fleet sizing and service frequency as decision variables would introduce additional structure into both the leader and follower problems while preserving the finite solution set property of the follower. Decision-dependent congestion, on the other hand, introduces nonlinearity into the follower's objective; although it preserves response finiteness, the resulting problem would require fundamentally different solution techniques.

5. Decomposition Approach

The proposed MILP formulation (7) provides an effective approach for solving the multimodal network design problem. However, as the number of trips, scenarios, and follower responses increases, the model size grows substantially. This growth results in considerable preprocessing times and computational challenges, potentially making the approach intractable for large-scale instances (see Section 6.2 for detailed computational results).

To address this, rather than incorporating all follower responses for every trip and scenario, we consider a relaxed problem that includes only a subset of trips and scenarios. This reduces problem size and facilitates an iterative approach to strengthening the relaxation progressively via decomposition techniques. Specifically, in this section, we assume that a subset $S \subset T \times \Omega$ of trips and scenarios, which are fully explored, is available. Such a subset can be obtained either (i) by imposing a maximum iteration cap $n_{\max}$ to limit the number of follower responses generated or (ii) by restricting the set of trips and scenarios explored.

In Section 5.1, we present an MILP corresponding to the subset S which provides a lower bound relaxation of the original multimodal network design problem (1a)-(1c), and show how to iteratively strengthen it using cutting planes via a decomposition framework. Section 5.2 discusses practical computational enhancements on algorithm stability.

5.1. Compact Formulation and the Cutting-Plane Algorithm

The multimodal network design problem (1a)-(1c) can be expressed compactly as

$$\min_{\boldsymbol{z}\in Z} \beta^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} f^{r,\omega}(\boldsymbol{z}),$$

where the disutility function $f^{r,\omega}(\boldsymbol{z})$ of trip r and scenario ω is defined as $f^{r,\omega}(\boldsymbol{z}) = \min_{\boldsymbol{u}\in\Xi^{r,\omega}(\boldsymbol{z})} (\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}$ corresponding to (2) with the follower optimal solution set $\Xi^{r,\omega}(\boldsymbol{z}) = \arg\min_{\boldsymbol{u}} \{(\boldsymbol{c}^{r,\omega})^\top \boldsymbol{u} :\ \boldsymbol{G}\boldsymbol{u} \le \boldsymbol{b}^r(\boldsymbol{z}),\ \boldsymbol{u} \ge \boldsymbol{0}\}$ associated with the follower problem defined in (3). Here the right-hand side $\boldsymbol{b}^r(\boldsymbol{z})$ is linear in $\boldsymbol{z}$, i.e., $b^r(\boldsymbol{z}) = \boldsymbol{b}_0^r + \boldsymbol{Q}^r \boldsymbol{z}$. An equivalent formulation introduces variables $\delta^{r,\omega}$ to represent the leader's optimal value:

$$\min_{\boldsymbol{z}\in Z,\delta} \left\{ \boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} \delta^{r,\omega} :\ \delta^{r,\omega} \ge f^{r,\omega}(\boldsymbol{z}),\ r\in T,\ \omega\in\Omega \right\}.$$

Assuming the response search algorithm fully explores a subset $S \subset T\times\Omega$ of the trips and scenarios, the above formulation is equivalent to

$$\min_{\substack{\boldsymbol{z}\in Z,\boldsymbol{\delta} \\ (\boldsymbol{\chi}^{r,\omega},\boldsymbol{\phi}^{r,\omega},\forall (r,\omega)\in S)\in\Phi_S}} \left\{ \boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} \delta^{r,\omega} :\ \delta^{r,\omega} \ge f^{r,\omega}(\boldsymbol{z}),\ (r,\omega)\in S^C \right\}, \tag{8}$$

where the set Φ_S includes Constraints (4a)-(4c), (5a)-(5d), (6) associated with the subset S, and $S^C = (T\times\Omega)\backslash S$ is the complement of S corresponding to the set of trips and scenarios that are not explored by this algorithm.

To develop the cutting-plane algorithm, we first derive an equivalent dual reformulation of the disutility function.

PROPOSITION 1. *The disutility function $f^{r,\omega}(\boldsymbol{z})$ can be equivalently represented as*

$$f^{r,\omega}(\boldsymbol{z}) = \max_{(\boldsymbol{\pi},t)\in\Pi^{r,\omega}} \boldsymbol{b}^r(\boldsymbol{z})^\top \pi + t \max_{\boldsymbol{\vartheta}\in\Theta^{r,\omega}} \boldsymbol{b}^r(\boldsymbol{z})^\top \boldsymbol{\vartheta}, \tag{9}$$

where $\Pi^{r,\omega} = \{(\boldsymbol{\pi},t) :\ \boldsymbol{G}^\top\boldsymbol{\pi} + t\boldsymbol{c}^{r,\omega} \le \boldsymbol{v}^{r,\omega},\ \boldsymbol{\pi}\le\boldsymbol{0},\ t\le 0\}$ and $\Theta^{r,\omega} = \{\boldsymbol{\vartheta} :\ \boldsymbol{G}^\top\boldsymbol{\vartheta} \le \boldsymbol{c}^{r,\omega}, \boldsymbol{\vartheta}\le\boldsymbol{0}\}$.

Proof of Proposition 1. To establish the equivalence, we define the optimal value function of the follower problem as $C^{r,\omega}(\boldsymbol{z}) = \min_u \{(\boldsymbol{c}^{r,\omega})^\top \boldsymbol{u} :\ \boldsymbol{G}\boldsymbol{u}\le\boldsymbol{b}^r(\boldsymbol{z}),\ \boldsymbol{u}\ge\boldsymbol{0}\}$. The corresponding follower optimal solution set can be written as $\Xi^{r,\omega}(\boldsymbol{z}) = \{\boldsymbol{u} :\ (\boldsymbol{c}^{r,\omega})^\top\boldsymbol{u} \le C^{r,\omega}(\boldsymbol{z}),\ \boldsymbol{G}\boldsymbol{u}\le\boldsymbol{b}^r(\boldsymbol{z}),\ \boldsymbol{u}\ge\boldsymbol{0}\}$. Next, introducing dual variables $t\le 0, \boldsymbol{\pi}\le\boldsymbol{0}$ associated with the constraints in this set, the dual formulation of $\min_{\boldsymbol{u}\in\Xi^{r,\omega}(\boldsymbol{z})} (\boldsymbol{v}^{r,\omega})^\top\boldsymbol{u}$ is

$$\max_{(\boldsymbol{\pi},t)\in\Pi^{r,\omega}} \boldsymbol{b}^r(\boldsymbol{z})^\top\boldsymbol{\pi} + tC^{r,\omega}(\boldsymbol{z}), \tag{10}$$

where $\Pi^{r,\omega}$ is defined in Proposition 1. Finally, by replacing the optimal value function $C^{r,\omega}(\boldsymbol{z})$ with its dual form by introducing the dual variable $\boldsymbol{\vartheta}\le\boldsymbol{0}$ and the corresponding set $\Theta^{r,\omega}$, we obtain the reformulation in (9). □

Following Proposition 1, constraints $\delta^{r,\omega} \geq f^{r,\omega}(\boldsymbol{z}),\ \forall (r,\omega) \in S^C$, in (8) can be equivalently replaced by the semi-infinite constraints

$$\delta^{r,\omega} \geq \boldsymbol{b}^r(\boldsymbol{z})^\top \boldsymbol{\pi} + t\boldsymbol{b}^r(\boldsymbol{z})^\top \boldsymbol{\vartheta},\ \boldsymbol{\vartheta} \in \Theta^{r,\omega},\ \forall (\boldsymbol{\pi}, t) \in \Pi^{r,\omega}.$$

Since imposing all $(\boldsymbol{\pi}, t)$ and $\boldsymbol{\vartheta}$ can be generally intractable due to the semi-infinite structure, we solve the resulting problem iteratively by proposing the following approach:

(i) Start with a finite subset of dual variables $\{(\boldsymbol{\pi}_\ell^{r,\omega}, t_\ell^{r,\omega}),\ \ell \in \mathcal{L}^{r,\omega}\} \subset \Pi^{r,\omega}$ for each trip r and scenario ω, where the index set $\mathcal{L}^{r,\omega}$ can be empty initially.

(ii) Solve the master problem based on the subsets and the full follower-response constraints Φ_S on the subset S of trips and scenarios, where the master problem can be represented as follows

$$\min_{\substack{\boldsymbol{z}\in Z, \boldsymbol{\delta}, \boldsymbol{\vartheta}, \\ (\boldsymbol{x}^{r,\omega}, \boldsymbol{\phi}^{r,\omega}, \forall (r,\omega)\in S) \in \Phi_S}} \Big\{ \boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} \delta^{r,\omega} : \delta^{r,\omega} \geq \boldsymbol{b}^r(\boldsymbol{z})^\top \boldsymbol{\pi}^{\hat{r},\omega_\ell} + \hat{t}_\ell^{r,\omega} \boldsymbol{b}^r(\boldsymbol{z})^\top \boldsymbol{\vartheta}^{r,\omega}, \\ \ell \in \mathcal{L}^{r,\omega},\ \boldsymbol{\vartheta}^{r,\omega} \in \Theta^{r,\omega},\ (r,\omega) \in S^C \Big\}, \tag{11}$$

Note that, for each r and ω, we do not require $|\mathcal{L}^{r,\omega}|$ copies of $\boldsymbol{\vartheta}$-variables as the optimal solution of the inner maximization problem in (9) over $\boldsymbol{\vartheta}$ is independent of the solution of the dual $\boldsymbol{\pi}$- and t-variable. The bilinear terms involving products $\boldsymbol{z}\boldsymbol{\vartheta}^{r,\omega}$ can be exactly linearized using McCormick inequalities (McCormick 1976) due to the binary nature of $\boldsymbol{z}$.

(iii) For the remaining trips r and scenarios ω in the complement S^C, given a network design $\hat{\boldsymbol{z}}$ obtained by solving the master problem, the follower problem is solved as a subproblem to generate violated cuts and new $(\boldsymbol{\pi}, t)$ pairs. Specifically, the subproblem (9) can be solved by sequentially solving two linear problems. Given $\hat{\boldsymbol{z}}$, evaluate the optimal value function $C^{r,\omega}(\boldsymbol{z})$ for every trip r and scenario ω by solving the follower's shortest path problem. Then solve the equivalent dual problem in the form of (10), which is also a linear problem, to obtain the dual solutions.

The iterative cutting-plane method is described in Algorithm 3.

PROPOSITION 2. *Algorithm 3 terminates in a finite number of iterations and converges to an optimal solution when the tolerance* $\epsilon = 0$.

Proof of Proposition 2. Finite termination follows from the fact that set $\Pi^{r,\omega}$ of the dual variables has a finite number of extreme points. Only finitely many cuts can be generated in the subproblems. Consequently, the cutting-plane algorithm converges to an optimal solution as $\epsilon = 0$. □

Algorithm 3 The cutting-plane algorithm

1: **Input:** Threshold ϵ, subset of trips and scenarios S, initial index set $\mathcal{L}^{r,\omega}$ for all $(r,\omega) \in S$.
2: **Initialization**: Set LB $= 0$, UB $= +\infty$.
3: **while** $\frac{\text{UB}-\text{LB}}{\text{UB}} > \epsilon$ **do**
4: Solve the master problem (11) with index set $\mathcal{L}^{r,\omega}$ for all $(r,\omega) \in S$.
5: Obtain an optimal solution $\hat{\boldsymbol{z}}$, $\hat{\delta}^{r,\omega},\ \forall r \in T, \omega \in \Omega$, and the corresponding objective value obj.
6: Update LB $= obj$.
7: **for** $(r,\omega) \in S^C$ **do**
8: Solve the subproblem $f^{r,\omega}(\hat{\boldsymbol{z}})$ using formulation (9).
9: Obtain an optimal solution $\hat{\boldsymbol{\pi}}$, $\hat{t}$ and the optimal value $\hat{f}^{r,\omega}$.
10: **if** $\hat{f}^{r,\omega} > \hat{\delta}^{r,\omega}$ **then**
11: Let $\ell = |\mathcal{L}^{r,\omega}| + 1$ and update $\mathcal{L}^{r,\omega} \leftarrow \mathcal{L}^{r,\omega} \cup \{\ell\}$. Set $\hat{\boldsymbol{\pi}}_\ell^{r,\omega} = \hat{\boldsymbol{\pi}}$ and $\hat{t}_\ell^{r,\omega} = \hat{t}$.
12: Update UB $= \min\left\{\text{UB},\ \boldsymbol{\beta}^\top \hat{\boldsymbol{z}} + \sum_{(r,\omega)\in S} \rho^\omega p^{r,\omega} \hat{\delta}^{r,\omega} + \sum_{(r,\omega)\in S^C} \rho^\omega p^{r,\omega} \hat{f}^{r,\omega}\right\}$.
13: **Output**: The optimal solution $\hat{\boldsymbol{z}}$, and the optimal objective value UB.

REMARK 8. Classical decomposition techniques such as Benders decomposition (Benders 1962) are not directly applicable to the stochastic network design problem (1) due to the bilevel structure and the nonconvexity introduced by the second-stage follower problems (Henkel 2014). Unlike standard stochastic problems with linear second-stage problems, the corresponding optimization problem in this context is a bilevel problem, which complicates traditional dualization-based approaches. Our proposed cutting-plane method overcomes this by exploiting the binary nature of the leader's decision $\boldsymbol{z}$, enabling the construction of a valid linear lower-bound approximation of the follower's disutility function $f^{r,\omega}(\boldsymbol{z})$. This method generalizes the spirit of Benders decomposition to accommodate bilevel structures within stochastic bilevel network design problems.

5.2. Stability Enhancement

The master problem (11) involves products between binary variable $\boldsymbol{z}$ and constants such as $\hat{\boldsymbol{\pi}}_\ell^{r,\omega}$ and $\hat{t}_\ell^{r,\omega}$. Using overly large constants can cause numerical instabilities (Cococcioni and Fiaschi 2021). For instance, consider the constraint $\delta \geq \hat{\pi} z$ where z is binary and $\hat{\pi}$ takes a large value (e.g., 10^8). When $z = 0$, we expect $\delta \geq 0$. However, in practice, solver tolerances may treat z as a small positive number (e.g., 10^{-5}) rather than an exact zero. This, combined with a large $\hat{\pi}$, can force δ to take a large positive value, which is not intended, therefore causing instability and inaccurate bounds.

Observed from our numerical experiments, such numerical issues arise especially from the presence of multiple optimal solutions to the subproblem (9) due to degeneracy in linear programs (e.g., Wolfe 1963, Magnanti and Wong 1981, Sherali and Lunday 2013). To mitigate this, after solving the subproblem at Line 9 of Algorithm 3, we solve the following secondary subproblem to select the "smallest" dual solution:

$$\min_{\zeta,(t,\boldsymbol{\pi})\in\Pi^{r,\omega}} \left\{\zeta:\ \zeta \geq \|\boldsymbol{\pi}\|_\infty,\ \zeta \geq -t,\ \boldsymbol{b}^r(\hat{\boldsymbol{z}})^\top \boldsymbol{\pi} + tC^{r,\omega}(\hat{\boldsymbol{z}}) \geq \hat{f}^{r,\omega}\right\}.$$

The objective minimizes ζ, which upper bounds both $\|\boldsymbol{\pi}\|_\infty$ and the absolute value of t to seek the smallest possible magnitudes of $\boldsymbol{\pi}$ and t among all optimal dual solutions.

6. Computational Results

This section provides a comprehensive computational study for analyzing the proposed stochastic bilevel optimization model for the multimodal transit network design problem, examining both computational performance and practical applications. Specifically, Section 6.2 demonstrates the computational efficiency of the proposed single-level reformulation and decomposition approaches on a diverse set of instances derived from a public transportation network of Dalian, China, benchmarking them against existing methods. Additionally, in Section 6.3, we present a case study over a real dataset from the Ann Arbor/Ypsilanti region in Michigan by highlighting the resulting network designs and the routes suggested to the riders, and the value of the bilevel optimization and integration of uncertainty with practical insights. For both the computational experiments and the case study, uncertain parameters are generated as detailed in Section 6.1. The experiments are conducted on a Linux server running Ubuntu 20.04 LTS, where the number of CPU cores is limited to 16. The models and solution algorithms are developed in Python 3.9 using Gurobi v12.0.0.

6.1. Scenario Generation

We consider two primary sources of uncertainty: travel times and passenger demand on each trip, of which scenarios are generated as follows. Firstly, to account for travel time uncertainty, vehicle speeds on each arc are generated following a truncated normal distribution with a mean of 24.92 miles per hour (mph) and a standard deviation of 5 mph, aligning with the findings of Du et al. (2017), which reports real-world bus cruising speeds. To capture variations in traffic, similar to the modeling approach in Lu et al. (2024), we further incorporate time-of-day effects: shuttle speeds are assumed to be 1.2 to 1.5 times that of buses during off-peak hours, and 0.6 to 0.85 times during peak hours. Secondly, passenger demand on each trip $r \in T$ is generated following a truncated normal distribution with mean μ^r and standard deviation σ^r, over the interval of $[0, \mu^r + 3\sigma^r]$. In the Dalian, China dataset, μ^r is assigned to a random integer between 2 and 10, based on existing surveys (Yao et al. 2016, Tang et al. 2017), whereas μ^r in the Ann Arbor/Ypsilanti, Michigan dataset, is obtained from the actual ridership data reported by Guan et al. (2024). Furthermore, we assume a constant coefficient of variation $\sigma^r/\mu^r = 1$ across all trips, implying that variability in passenger counts scales proportionally with the expected demand.

6.2. Computational Performances of Reformulations and Solution Approaches

To evaluate the computational performance of the proposed methods, we test on a public transportation network of Dalian, China, covering approximately 18×15 square miles with 81 nodes. Three sets of instances are considered: small- and medium-sized instances with 4 hubs, and large-sized instances with 6 hubs. Tables 3–5 present the computational comparisons for the small-, medium-, and large-sized instances,

respectively. We first compare the MILP reformulation (7) based on the response search algorithm (RS) with two classical reformulations: KKT-based methods (KKT) and strong-duality-based methods (SD). For both methods, we report results with and without the valid inequality (VI) proposed by Kleinert and Schmidt (2023) to enhance computational performance. Detailed formulations of the KKT and SD methods, as well as the valid inequality, are provided in Appendix A. For the response search algorithm, we experiment with both serial and parallel implementations for all trips and scenarios. Specifically, parallelization is implemented via the `multiprocessing` module in Python 3.9, using 2, 4, 8, and 16 CPU cores.

Table 3 reports the preprocessing time of the response search algorithm under parallel implementations with different numbers of cores, along with the MILP solution times and the number of nodes explored under the studied single-level reformulations. We observe that the parallel implementation significantly reduces the runtime of the response search algorithm, achieving up to 13 times speedup with 16 cores compared to the serial implementation. The total time reported for the response-search-based MILP formulation includes both the preprocessing time (using 16 cores) and the MILP solution time. For small instances, the two KKT-based methods (KKT and KKT+VI) take substantially longer computational times compared with the other approaches, and for some instances, fail to find feasible solutions within the 600-second time limit, even resulting in no optimality gap being reported. Incorporating VI does not improve the computational performance of the KKT formulation, even leading to longer solution times. In contrast, the SD approaches with and without VI perform better than the KKT-based methods, while the RS reformulation consistently takes the least computational time across all instances, even with the addition of the preprocessing time.

Table 3 Computational comparison over small instances (Time limit 600 seconds)

Trips	Scenarios	Method	Time (seconds)							Nodes
			Preprocessing					Optimization	Total	
			Serial	2 Cores	4 Cores	8 Cores	16 Cores			
200	10	KKT						53.20	53.20	1
		KKT+VI						68.92	68.92	1
		SD						11.78	11.78	1
		SD+VI						10.13	10.13	1
		RS	17.20	9.64	3.01	2.10	2.31	2.33	4.65	1
200	30	KKT						Limit(\)	Limit	1
		KKT+VI						Limit(\)	Limit	1
		SD						170.01	170.01	1
		SD+VI						131.86	131.86	1
		RS	29.19	25.25	9.39	4.57	4.41	8.02	12.43	1
300	10	KKT						Limit(\)	Limit	431
		KKT+VI						Limit(0.53%)	Limit	5518
		SD						24.03	24.03	27
		SD+VI						15.22	15.22	1
		RS	11.86	7.49	3.58	2.07	1.46	4.09	5.55	1
300	30	KKT						Limit(\)	Limit	1
		KKT+VI						Limit(\)	Limit	1
		SD						133.86	133.86	41
		SD+VI						92.70	92.70	1
		RS	39.45	35.48	15.28	10.70	6.33	15.89	22.22	1

For medium instances, we focus on comparing the RS and SD methods (with and without VI) in Table 4. The results show that the RS method is significantly faster than the SD methods across all tested instances. Table 4 also reports the number of iterations needed by the response search algorithm (both the maximum and average iterations over all follower-scenario pairs) under the Column "RS Iters". These numbers are significantly smaller than the worst-case bound of $2^{4\times3}$. More details of the iteration distributions are presented in Appendix B over the set of studied instances with various sizes. In particular, the majority of follower-scenario pairs complete the search in fewer than five iterations, thanks to the fact that riders typically use only a limited subset of hub legs for their trips, rather than considering all possible combinations. We also observe that incorporating the VI does not improve the performance of the SD formulation for these instances and even results in longer solution time or larger optimality gaps. Finally, for all instances in Tables 3 and 4, the RS method solves them at the root node, indicating a strong linear relaxation formulation of the proposed MILP model.

Table 4 Computational comparison over medium instances (Time limit = 3600 seconds)

Trips	Scenarios	Method	RS Iters (max/avg.)	Time (seconds)							Nodes
				Preprocessing					Optimization	Total	
				Serial	2 Core	4 Core	8 Core	16 Core			
600	30	SD							346.68	346.68	25
		SD+VI							564.11	564.11	24
		RS	33/2.38	71.26	35.99	18.34	11.90	7.69	5.95	13.64	1
600	50	SD							1111.44	1111.44	18
		SD+VI							1546.52	1546.52	21
		RS	30/2.19	118.69	61.82	31.37	19.08	12.40	8.69	21.08	1
1500	30	SD							1000.45	1000.45	20
		SD+VI							1515.80	1515.80	34
		RS	40/2.38	128.67	69.82	37.15	20.39	13.45	17.32	30.78	1
1500	50	SD							Limit (0.70%)	Limit	10
		SD+VI							Limit (1.43%)	Limit	1
		RS	40/2.86	217.63	156.77	70.83	39.54	24.67	72.88	97.55	1
2000	30	SD							726.03	726.03	12
		SD+VI							873.75	873.75	13
		RS	33/2.86	214.93	85.28	51.70	27.00	18.29	50.46	68.74	1
2000	50	SD							Limit (3.95%)	Limit	1
		SD+VI							Limit(\)	Limit	1
		RS	37/3.03	479.64	224.93	139.78	79.38	36.99	183.34	220.33	1

For large instances, we increase the number of hubs to six. In our implementation, we observe that the RS method itself fails to find a reasonable solution within the time limit of one hour due to: (1) the increased complexity of enumerating all responses for larger instances and (2) the intractability of solving the resulting large MILP. To this end, we utilize the decomposition algorithm proposed in Section 5 for the solution of these instances and compare it against the RS reformulation in Table 5. To manage computational complexity for RS reformulation, we impose a maximum iteration cap $n_{\max}$ in the response search algorithm to limit the number of follower responses explored and integrated into the MILP. We set $n_{\max} = 10000$,

which results in fully exploring over 99% follower problems, calculated by $|S|/(|S|+|S^C|)$, in every case. As a benchmark, we also consider a special case of $n_{\max}=0$, in which Algorithm 3 is initialized with no pre-obtained responses, i.e., $S=\emptyset$. This case is denoted as Cuts in Table 5.

We then construct a relaxed MILP by including only the subset S of responses explored and assigning a lower bound for the unexplored responses. Specifically, for each unexplored response $(r,\omega)\notin S$ we set $\delta^{r,\omega}$ to the optimal leader disutility $f^{r,\omega}(\hat{\boldsymbol{z}})$ assuming that all bus legs are available, i.e., $\hat{z}_{h,l}=1$ for all $(h,l)\in\mathcal{H}$. Solving this relaxed MILP yields a valid lower bound (LB) to the original full problem. Let $\underline{\boldsymbol{z}}$ denote the optimal solution of the corresponding MILP.

To compute an upper bound (UB), we evaluate the actual disutilities for the unexplored responses by solving their follower problems using $\underline{\boldsymbol{z}}$ and calculating their corresponding leader's disutilities. The relative optimality gap is then calculated as (UB-LB)/UB and reported in Column "Gap" to evaluate the quality of the solutions obtained, along with the lower and upper bounds obtained in Columns "LB" and "UB", respectively. The results show that the RS method with partial responses struggles with solution quality on large instances in Table 5, exhibiting large optimality gaps exceeding 96% within the 3600-second time limit for most instances. In comparison, Cuts, which implements Algorithm 3 without any pre-obtained responses, achieves better upper bounds than RS. This can be because the master problem of Cuts contains fewer constraints, enabling a more effective exploration of feasible solutions across the branch-and-bound tree. Nevertheless, RS has more response-derived constraints, yielding tighter lower bounds, which is crucial to the computational performance of RS+Cuts.

In RS+Cuts, we implement a relaxed formulation of RS in which set S is constructed using Algorithms 1 and 2 with $n_{\max}=100$ follower responses, and the relaxation is further strengthened with the cutting planes as in Algorithm 3. Additionally, we also implement RS+Cuts with the practical stability enhancements (see Section 5.2), denoted as RS+Cuts+Stab. As shown in Table 5, the RS+Cuts method reduces these gaps to below 1.5% in all cases. With the stability enhancements (RS+Cuts+Stab), these gaps are further reduced, with even one instance (of 1000 trips and 30 scenarios) being solved optimally within the time limit, which demonstrates the effectiveness of the cutting-plane approach and stability enhancements in tightening the relaxation and significantly improving solution quality.

REMARK 9. Algorithm 3 can alternatively be implemented via a branch-and-cut framework using the callback function in Gurobi. Appendix E presents a computational comparison of the callback implementation, without and with RS constraints in the master problem. While the callback implementation without RS constraints performs comparably to the iterative Cuts implementation in Table 5, adding RS constraints to the master program can drastically deteriorate computational efficiency, yielding substantially worse optimality gaps. Due to the interplay between the RS constraints and cuts added, the solution processes remain at the root node of the Branch-and-Bound trees.

Table 5 Computational comparison over large instances (Time limit = 3600 seconds)

Trips	Scenarios	Method	$n_{\max}$	% explored	Avg. RS Iters	Preprocessing (16 Cores)	LB	UB	Gap
1000	30	Cuts	0	0.00%			659.43	712.17	7.41%
		RS	10000	99.70%	319.02	107.17	685.27	837.85	18.21%
		RS+Cuts	100	93.10%	11.23	21.53	683.12	685.55	0.35%
		RS+Cuts+Stab	100	93.10%	11.23	21.53	685.55	685.55	0.00% (2186 seconds)
1000	50	Cuts	0	0.00%			1384.24	1802.77	23.22%
		RS	10000	99.43%	443.45	221.49	1557.75	134235.29	98.84%
		RS+Cuts	100	79.79%	17.83	40.96	1547.59	1566.70	1.22%
		RS+Cuts+Stab	100	79.79%	17.83	40.96	1549.59	1566.70	1.09%
1200	30	Cuts	0	0.00%			1712.16	2079.99	17.68%
		RS	10000	99.51%	607.53	330.67	1888.20	203388.24	99.07%
		RS+Cuts	100	74.99%	21.78	42.42	1883.34	1904.05	1.09%
		RS+Cuts+Stab	100	74.99%	21.78	42.42	1883.51	1904.05	1.08%
1200	50	Cuts	0	0.00%			776.33	979.44	20.74%
		RS	10000	99.93%	210.21	114.58	896.01	64584.34	98.61%
		RS+Cuts	100	93.12%	11.71	40.86	898.06	904.19	0.68%
		RS+Cuts+Stab	100	93.12%	11.71	40.86	898.42	904.19	0.64%
1500	30	Cuts	0	0.00%			1250.19	1524.75	18.01%
		RS	10000	99.69%	437.05	218.73	1400.13	41832.73	96.65%
		RS+Cuts	100	87.71%	15.99	37.01	1397.49	1415.18	1.25%
		RS+Cuts+Stab	100	87.71%	15.99	37.01	1398.15	1415.18	1.20%
1500	50	Cuts	0	0.00%			962.51	1088.41	11.57%
		RS	10000	99.92%	248.00	218.09	1021.36	68639.08	98.51%
		RS+Cuts	100	93.34%	11.82	47.05	1017.92	1024.01	0.59%
		RS+Cuts+Stab	100	93.34%	11.82	47.05	1021.10	1024.01	0.28%

6.3. Case Study

In this section, we present results of our proposed framework over a real dataset based on the transit system AAATA, the transit agency serving the Ann Arbor/Ypsilanti region in Michigan, USA. The transit system considers 1267 bus stops, where 10 stops are designated as hubs by following the setting proposed in Basciftci and Van Hentenryck (2023), Guan et al. (2024). 1503 trips are considered with a total expected number of passengers equal to 2897, corresponding to the ridership from 6 pm to 10 pm on a specific day, where each trip can be associated with multiple riders taking that trip.

To reduce the size of the resulting problem, arc elimination strategy (Basciftci and Van Hentenryck 2023) is adopted, where the network structure of each follower problem associated with trip $r \in T$ is simplified by considering only the following arcs for the on-demand shuttles: (1) from origin o^r to destination d^r; (2) from origin o^r to all hubs $h \in H$; and (3) from any hub $h \in H$ to the destination d^r . Additionally, similar trips are aggregated to reduce the model size and alleviate computational burden by presenting a trip clustering procedure as described in Appendix C, reducing the number of trips considered to 659 while capturing 2586 riders.

To examine different behaviors of riders against the transit system, trips are classified by riders' income levels. To accomplish this, destination stops of the trips are linked to the residential addresses of the riders, as most riders are heading home between 6 pm and 10 pm. Then, we match residential addresses to regional income data based on spatial income distribution (City-Data Forum 2024, Pew Research Center 2024). As a

result, 21, 283, and 355 trips are associated with high-income, middle-income, and low-income populations with 129, 1104, and 1353 riders, respectively. For capturing these different ridership preferences in follower problems, θ^f values for low-income, middle-income and high-income trips are set to 0.005, 0.01, and 0.03, respectively. Since higher values indicate more sensitivity to the duration of the trip in comparison to its cost, higher (lower) income riders are associated with higher (lower) θ^f values. We set θ^l to 0.0001 in the leader problem in the baseline setting, representing the preferences of the transit agency. The remainder of the problem parameters are set to the following value: $a = \$7.24$ per mile, $m = \$2.86$ per mile, $n = 16$, $W = 450$ seconds.

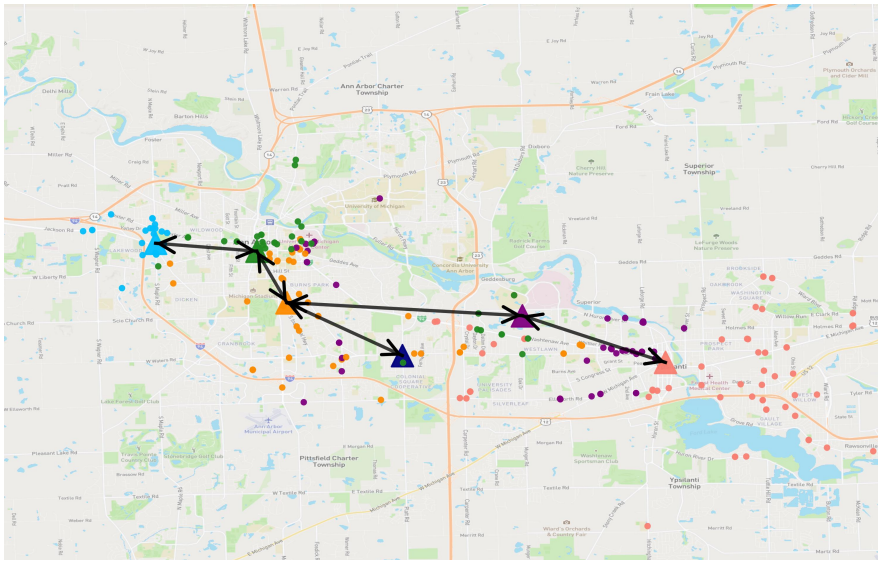
6.3.1. Network Designs We present the transit network design from the stochastic bilevel model and compare it with the network designs obtained from two benchmark approaches: i) a deterministic bilevel model using expected travel time and ridership values, and ii) a stochastic single-level model that incorporates uncertainty but removes the bilevel structure by moving the follower constraints to the leader problem and omitting follower objectives.

Figure 3 presents the resulting network designs, where each open hub is represented with a unique color, while the origins of the trips using the corresponding hub as part of their initial transfer stop are marked with the same color. The three approaches yield notably different network designs. More specifically, the stochastic bilevel model utilizes six hubs, whereas the deterministic bilevel model uses three, and the stochastic single-level model uses all of the potential 10 hubs in its designs. Consequently, the cost of operating these systems and the resulting routes suggested to the riders differ under each setting, which are examined in Sections 6.3.2 and 6.3.3, respectively.

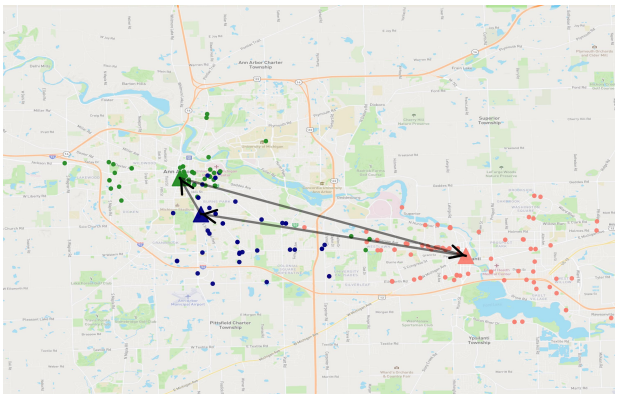
6.3.2. Value of Bilevel Optimization and Integration of Uncertainty In this section, we analyze the value of the proposed approach by evaluating the three network designs obtained in Section 6.3.1 under the stochastic bilevel model. Table 6 provides the resulting comparisons where the investment cost indicates the total weighted cost of operating bus legs, whereas the travel disutilities correspond to the disutilities realized by the transit agency and the riders as obtained from their corresponding leader and follower problems. The total objective column indicates the objective of the leader problem, which is realized by the transit agency.

In line with the network designs illustrated in Figure 3, the stochastic single-level approach requires the highest investment cost for operating the transit network, whereas the deterministic bilevel approach opens less hub legs and relies more on on-demand shuttles, which leads to highest disutility value observed by the transit agency. We observe that the stochastic bilevel approach provides a balance between these two settings with significant savings, resulting in 12.29% and 12.83% reductions in the total objective values, when compared with the network designs obtained by the deterministic bilevel and stochastic single-level approaches, respectively. Comparing the disutilities realized by the riders, the stochastic and deterministic

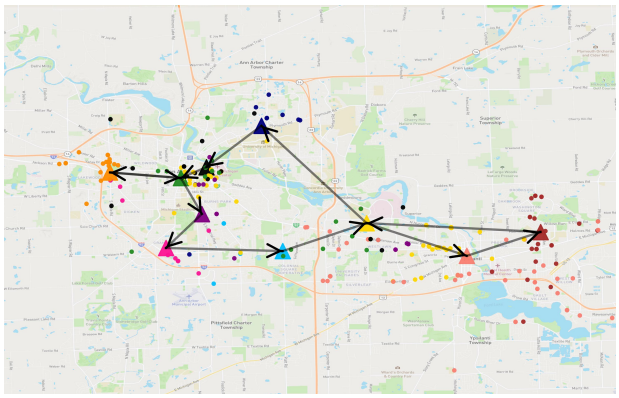
Figure 3 **Network designs under different modeling approaches.**



(a) Stochastic bilevel model



(b) Deterministic bilevel model



(c) Stochastic single-level model

bilevel approaches yield similar values, whereas the single-level counterpart results in much higher rider disutility. In particular the stochastic bilevel approach reduces rider disutility by up to 7% compared with the stochastic single-level method. These results highlight the importance of considering a bilevel model from the perspectives of both the transit agency and riders while capturing the underlying uncertainties.

Table 6 **Comparison of investment amounts and travel disutilities under different modeling approaches**

Model	Investment Cost	Travel Disutilities		Total Objective
		Transit Agency	Riders	
Stochastic bilevel	2687.09	19301.89	33046.87	21988.97
Deterministic bilevel	1720.67	23348.49	32136.59	25069.17
Stochastic single-level	3018.45	22500.45	35538.88	25518.90

6.3.3. Analyses of Suggested Paths to Riders In this section, we examine the durations of the suggested paths to the riders by the transit agency under the stochastic bilevel model. Table 7 provides these results for riders with different income levels. Since travel duration is stochastic, we report the expected travel time over all scenarios and the reported paths are based on each follower's objective, which balances cost and convenience (measured by distance and time). "Mode" column represents whether the trip is a direct shuttle ride or multimodal by involving shuttle and bus, and "Bus Transits" column indicates the average number of transfers when buses are utilized. "Distance" and "Duration" columns compare suggested paths within each income level and mode. Higher-income riders, prioritizing convenience, have the highest direct on-demand shuttle trips, whereas low-income level riders' trips utilize the transit network further. The results indicate that the transit system provides efficient paths to riders from different income levels, where the distances traveled are very similar when the travel times under the transit system and direct trips are compared, whereas travel duration for multimodal trips can be longer considering transfers and slower speeds of buses. To analyze the trade-off between cost and convenience of the paths suggested to the riders, the leader can adjust its corresponding parameter θ^l in its objective function, which is examined in Section 6.3.4, that can lead to different network designs and paths for riders.

Table 7 Trip duration analysis

Income	Mode	Percentage	Distance (Miles)		Duration (Minutes)		Bus Transits
			Transit System	Direct	Transit System	Direct	
low	shuttle	68.45%	1.80	1.80	3.31	3.31	\
	shuttle+bus	31.55%	6.60	5.58	43.84	10.64	1.62
	any	100.00%	3.31	2.98	16.14	5.60	0.51
middle	shuttle	79.19%	2.03	2.03	3.76	3.76	\
	shuttle+bus	20.81%	4.97	4.15	30.94	7.79	1.33
	any	100.00%	2.51	2.37	8.30	4.42	0.22
high	shuttle	84.29%	1.67	1.67	3.15	3.15	\
	shuttle+bus	15.71%	6.74	5.61	40.67	10.45	1.46
	any	100.00%	1.96	1.90	5.23	3.56	0.08

6.3.4. Sensitivity analysis on the parameter θ^l Table 8 compares the impact of θ^l value on the decisions of the leader and follower problems, where $\theta^l = 0.0001$ is the baseline setting studied in Sections 6.3.1-6.3.3, and $\theta^l \in \{0.002, 0.005\}$ correspond to the cases when the leader prioritizes travel duration further in the expense of higher operating costs of the transit network and on-demand shuttles. Table 8 provides the distances and durations of the paths suggested to riders from different income levels, followed by objective values considering to investment and travel disutilities. Additionally, network designs under $\theta^l \in \{0.002, 0.005\}$ are presented in Appendix D. The results indicate that the travel time of the paths suggested to the riders decreases as θ^l increases, which can also be seen as a reduction in travel disutilities realized by the riders. On the other hand, the total objective and the transit agency's disutility increase when θ^l increases, capturing the interplay between the cost and convenience of the trips suggested. Although the

transit network designs change as θ^l increases from 0.002 to 0.005, travel durations of the paths suggested to the riders and the realized disutilities of the riders remain similar between the two settings.

Table 8 Comparison of investment amounts, travel disutilities, and trip durations under different θ^l values

θ^l	Income	Distance (Miles)		Duration (Minutes)		Investment Cost	Travel Disutilities		Total Objective
		Transit System	Direct	Transit System	Direct		Transit Agency	Riders	
0.0001	low	3.31	2.98	16.14	5.60	2687.09	19301.89	33046.87	21988.97
	middle	2.51	2.37	8.30	4.42				
	high	1.96	1.90	5.23	3.56				
	any	2.91	2.68	12.33	5.01				
0.002	low	3.34	2.98	10.44	5.60	2677.70	23354.69	29919.75	26032.39
	middle	2.52	2.37	6.75	4.42				
	high	2.00	1.90	4.75	3.56				
	any	2.94	2.68	8.62	5.01				
0.005	low	3.34	2.98	10.18	5.60	2615.96	28731.61	29889.24	31347.57
	middle	2.53	2.37	6.51	4.42				
	high	2.01	1.90	4.58	3.56				
	any	2.94	2.68	8.38	5.01				

7. Conclusion

This paper presents a novel two-stage stochastic bilevel optimization model for multimodal transit network design that explicitly incorporates heterogeneous rider preferences, uncertain travel times and trip demand. By modeling the hierarchical decision-making process between transit agencies and riders, our approach captures realistic operational and behavioral dynamics for effective network planning.

We developed an equivalent single-level MILP reformulation by proposing the response search algorithm that efficiently enumerates critical follower route selections and then leveraging the outputs of this algorithm to reformulate the stochastic bilevel problem. To improve scalability in large-scale networks, we proposed a decomposition method that iteratively strengthens a relaxed formulation from a subset of follower responses. Notably, the proposed single-level reformulation and the response search algorithm can be extended to other stochastic bilevel optimization problems involving binary leader decisions beyond network design problems. Computational experiments on real-world datasets demonstrate significant computational efficiency and practical improvements of our methods. Additionally, our case study illustrates the practical benefits of our approaches, including cost savings and improved rider convenience.

Future research directions include extending the framework to incorporate additional sources of multimodal uncertainty, such as demand variability under different traffic scenarios, incorporation of the decision-dependent congestion effect induced by the resulting network design, and exploring robust network design strategies that hedge against undesired traffic conditions.

Acknowledgment

This work is supported by the National Science Foundation grants 2434301 and 2434302.

References

Abotalebi, E., Petrunić, J., 2021. New mobility and autonomous vehicles. URL: https://cutric-crituc.org/wp-content/uploads/2022/03/New-Mobility-and-Autonomous-Vehicles-Impacts-on-Greenhouse-Gas-Emissions-in-Metro-Vancouver.pdf. accessed: 2025-07-30.

An, K., Lo, H.K., 2016. Two-phase stochastic program for transit network design under demand uncertainty. *Transportation Research Part B: Methodological* 84, 157–181.

Auad-Perez, R., Van Hentenryck, P., 2022. Ridesharing and fleet sizing for on-demand multimodal transit systems. *Transportation Research Part C: Emerging Technologies* 138, 103594.

Basciftci, B., Van Hentenryck, P., 2020. Bilevel optimization for on-demand multimodal transit systems, in: *Integration of Constraint Programming, Artificial Intelligence, and Operations Research: 17th International Conference, CPAIOR 2020, Vienna, Austria, September 21–24, 2020, Proceedings 17*, Springer. pp. 52–68.

Basciftci, B., Van Hentenryck, P., 2023. Capturing travel mode adoption in designing on-demand multimodal transit systems. *Transportation Science* 57, 351–375.

Beck, Y., Ljubić, I., Schmidt, M., 2023. A survey on bilevel optimization under uncertainty. *European Journal of Operational Research* 311, 401–426.

Benders, J.F., 1962. Partitioning procedures for solving mixed-variables programming problems. *Numer. Math* 4, 238–252.

Birge, J.R., Louveaux, F., 2011. *Introduction to stochastic programming*. Springer Science & Business Media.

Boarnet, M.G., Shao, Q., Pilgram, C.A., 2024. Monetary cost, time cost, and mode choice: Transit and ridehailing in California. *Transportation Research Part D: Transport and Environment* 130, 104149.

Borndörfer, R., Grötschel, M., Pfetsch, M.E., 2007. A column-generation approach to line planning in public transport. *Transportation Science* 41, 123–132.

Burtscheidt, J., Claus, M., Dempe, S., 2020. Risk-averse models in bilevel stochastic linear programming. *SIAM Journal on Optimization* 30, 377–406.

Chen, A., Zhou, Z., Chootinan, P., Ryu, S., Yang, C., and, S.C.W., 2011. Transport network design problem under uncertainty: A review and new developments. *Transport Reviews* 31, 743–768.

City-Data Forum, 2024. Ann Arbor, Michigan (MI) income map, earnings map, and wages data. URL: https://www.city-data.com/income/income-Ann-Arbor-Michigan.html. accessed: 2024-06-20.

Cococcioni, M., Fiaschi, L., 2021. The Big-M method with the numerical infinite M. *Optimization Letters* 15, 2455–2468.

Dalmeijer, K., Van Hentenryck, P., 2020. Transfer-expanded graphs for on-demand multimodal transit systems, in: *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, Springer. pp. 167–175.

Dempe, S., 2002. *Foundations of bilevel programming*. Springer Science & Business Media.

Dempe, S., Zemkoho, A., 2020. Bilevel optimization, in: *Springer optimization and its applications*. Springer. volume 161.

Du, Y., Deng, F., Liao, F., Ji, Y., 2017. Understanding the distribution characteristics of bus speed based on geocoded data. *Transportation Research Part C: Emerging Technologies* 82, 337–357.

Durán-Micco, J., Vansteenwegen, P., 2022. A survey on the transit network design and frequency setting problem. *Public Transport* 14, 155–190.

Farahani, R.Z., Miandoabchi, E., Szeto, W., Rashidi, H., 2013. A review of urban transportation network design problems. *European Journal of Operational Research* 229, 281–302.

Federal Transit Administration, 2023. Mobility on demand sandbox program. URL: https://www.transit.dot.gov/research-innovation/mobility-demand-mod-sandbox-program. accessed: 2024-05-29.

Fischetti, M., Ljubić, I., Monaci, M., Sinnl, M., 2017. A new general-purpose algorithm for mixed-integer bilevel linear programs. *Operations Research* 65, 1615–1637.

Garcia-Martinez, A., Cascajo, R., Jara-Diaz, S.R., Chowdhury, S., Monzon, A., 2018. Transfer penalties in multimodal public transport networks. *Transportation Research Part A: Policy and Practice* 114, 52–66.

Goyal, A., Zhang, Y., He, C., 2023. Decision rule approaches for pessimistic bilevel linear programs under moment ambiguity with facility location applications. *INFORMS Journal on Computing* 35, 1342–1360.

Guan, H., Basciftci, B., Van Hentenryck, P., 2024. Path-based formulations for the design of on-demand multimodal transit systems with adoption awareness. *INFORMS Journal on Computing* 36, 1359–1756.

Guan, H., Basciftci, B., Van Hentenryck, P., 2026. Bilevel optimization and heuristic algorithms for integrating latent demand into the design of large-scale transit systems. *Transportation Science* .

Guimarães, C.G.C., Oliveira-Neto, F.M., 2026. Transfer perception in multimodal public transport networks. *Transportation Research Part A: Policy and Practice* 204, 104841.

Henkel, C., 2014. *An algorithm for the global resolution of linear stochastic bilevel programs*. Ph.D. thesis. Duisburg, Essen, Universität Duisburg-Essen, Diss., 2014.

Jeroslow, R.G., 1985. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming* 32, 146–164.

Kleinert, T., Labbé, M., Ljubić, I., Schmidt, M., 2021. A survey on mixed-integer programming techniques in bilevel optimization. *EURO Journal on Computational Optimization* 9, 100007.

Kleinert, T., Labbé, M., Plein, F.a., Schmidt, M., 2020. There's no free lunch: on the hardness of choosing a correct big-M in bilevel optimization. *Operations Research* 68, 1716–1721.

Kleinert, T., Schmidt, M., 2023. Why there is no need to use a big-M in linear bilevel optimization: A computational study of two ready-to-use approaches. *Computational Management Science* 20, 3.

Kreutzberger, E.D., 2008. Distance and time in intermodal goods transport networks in Europe: A generic approach. *Transportation Research Part A: Policy and Practice* 42, 973–993.

Lampariello, L., Sagratella, S., Stein, O., 2019. The standard pessimistic bilevel problem. *SIAM Journal on Optimization* 29, 1634–1656.

Liu, Y., Ouyang, Y., 2021. Mobility service design via joint optimization of transit networks and demand-responsive services. *Transportation Research Part B: Methodological* 151, 22–41.

Lu, J., Trasatti, A., Guan, H., Dalmeijer, K., Van Hentenryck, P., 2024. The impact of congestion and dedicated lanes on on-demand multimodal transit systems. *Travel Behaviour and Society* 36, 100772.

Luo, Q., Li, S., Hampshire, R.C., 2021. Optimal design of intermodal mobility networks under uncertainty: Connecting micromobility with mobility-on-demand transit. *EURO Journal on Transportation and Logistics* 10, 100045.

Magnanti, T.L., Wong, R.T., 1981. Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations Research* 29, 464–484.

Mahéo, A., Kilby, P., Van Hentenryck, P., 2019. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science* 53, 77–88.

McCormick, G.P., 1976. Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. *Mathematical Programming* 10, 147–175.

Najmi, A., Rashidi, T.H., Waller, T., 2023. A multimodal multi-provider market equilibrium model: A game-theoretic approach. *Transportation Research Part C: Emerging Technologies* 146, 103959.

Pew Research Center, 2024. The state of the American middle class. URL: https://www.pewresearch.org/race-and-ethnicity/2024/05/31/the-state-of-the-american-middle-class/. accessed: 2024-10-23.

Rataj, M., Lodi, C., Zawieska, J., Stepniak, M., Cheimariotis, I., Grosso, M., Piazza, F., Marotta, A., 2025. *New and Emerging Transport Technologies and Trends in European Research and Innovation Projects 2024*. Technical Report JRC140839. Publications Office of the European Union. Luxembourg. URL: https://data.europa.eu/doi/10.2760/1362356, doi:10.2760/1362356.

Shaheen, S., Chan, N., 2016. Mobility and the sharing economy: Potential to facilitate the first- and last-mile public transit connections. *Built Environment* 42, 573–588.

Shapiro, A., Xu, H., 2008. Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization* 57, 395–418.

Sherali, H.D., Lunday, B.J., 2013. On generating maximal nondominated Benders cuts. *Annals of Operations Research* 210, 57–72.

Sherali, H.D., Soyster, A.L., 1983. Preemptive and nonpreemptive multi-objective programming: Relationship and counterexamples. *Journal of Optimization Theory and Applications* 39, 173–186.

Stiglic, M., Agatz, N., Savelsbergh, M., Gradisar, M., 2018. Enhancing urban mobility: Integrating ride-sharing and public transit. *Computers & Operations Research* 90, 12–21.

Tang, C., Ceder, A., Ge, Y.E., 2017. Integrated optimization of bus line fare and operational strategies using elastic demand. *Journal of Advanced Transportation* 2017, 7058789.

Uchida, K., Sumalee, A., Ho, H.W., 2015. A stochastic multimodal reliable network design problem under adverse weather conditions. *Journal of Advanced Transportation* 49, 73–95.

Vaidya, O.S., Kumar, S., 2006. Analytic hierarchy process: An overview of applications. *European Journal of operational research* 169, 1–29.

Van Hentenryck, P., Riley, C., Trasatti, A., Guan, H., Santanam, T., Huertas, J.A., Dalmeijer, K., Watkins, K., Drake, J., Baskin, S., 2023. MARTA Reach: Piloting an On-Demand Multimodal Transit System in Atlanta. *arXiv preprint arXiv:2308.02681* .

Wolfe, P., 1963. A technique for resolving degeneracy in linear programming. *Journal of the Society for Industrial and Applied Mathematics* 11, 205–211.

Wu, Y., Tang, J., Yu, Y., Pan, Z., 2015. A stochastic optimization model for transit network timetable design to mitigate the randomness of traveling time by adding slack time. *Transportation Research Part C: Emerging Technologies* 52, 15–31.

Yanıkoglu, I., Kuhn, D., 2018. Decision rule bounds for two-stage stochastic bilevel programs. *SIAM Journal on Optimization* 28, 198–222.

Yao, B., Wang, Z., Cao, Q., Jin, L., Zhang, M., 2016. Express bus fare optimization based on passenger choice behavior. *Simulation* 92, 617–625.

Yao, R., Zhang, K., 2024. Design an intermediary mobility-as-a-service (MaaS) platform using many-to-many stable matching framework. *Transportation Research Part B: Methodological* 189, 102991.

Yu, B., Kong, L., Sun, Y., Yao, B., Gao, Z., 2015. A bi-level programming for bus lane network design. *Transportation Research Part C: Emerging Technologies* 55, 310–327.

Zare, M.H., Borrero, J.S., Zeng, B., Prokopyev, O.A., 2019. A note on linearized reformulations for a class of bilevel linear integer problems. *Annals of Operations Research* 272, 99–117.

Zhang, J., Özaltın, O.Y., 2021. Bilevel integer programs with stochastic right-hand sides. *INFORMS Journal on Computing* 33, 1644–1660.

Zhu, G., Ye, M., Yu, X., Liu, J., Wang, M., Luo, Z., Liang, H., Zhong, Y., 2025. Optimizing route planning via the weighted sum method and multi-criteria decision-making. *Mathematics* 13, 1704.

Appendix A: Two Exact Single-Level Reformulations: KKT-based and Strong-duality-Based

In this section, we present the two widely-used single-level reformulations (Zare et al. 2019, Dempe and Zemkoho 2020): a Karush-Kuhn-Tucker-based (KKT-based) one and a strong-duality-based one. We further present a valid inequality that can be used to strengthen these formulations. To simplify the notations, we work on the equivalent compact form described in Section 5.1. Specifically, the multimodal network design problem can be expressed compactly as

$$\min_{\boldsymbol{z}\in Z} \boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} f^{r,\omega}(\boldsymbol{z}),$$

where the follower's disutility function for trip r and scenario ω is defined as

$$f^{r,\omega}(\boldsymbol{z}) = \min_{\boldsymbol{u}} \left\{ (\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u} :\ \boldsymbol{u} \in \arg\min_{\boldsymbol{u}\geq 0} \left\{ (\boldsymbol{c}^{r,\omega})^\top \boldsymbol{u} : \boldsymbol{G}\boldsymbol{u} \leq \boldsymbol{b}^r(\boldsymbol{z}) \right\} \right\},$$

with $b^r(\boldsymbol{z})$ depending affinely on $\boldsymbol{z}$.

A.1. KKT-Based Single-Level Reformulation with Big-M

The follower problem's KKT conditions include primal feasibility, dual feasibility, stationarity, and complementarity:

$$\boldsymbol{G}\boldsymbol{u}^{r,\omega} \leq \boldsymbol{b}^r(\boldsymbol{z}), \quad \boldsymbol{u}^{r,\omega} \geq \boldsymbol{0},\ \forall r\in T,\ \forall \omega\in\Omega \tag{12}$$

$$\boldsymbol{\pi}^{r,\omega} \geq \boldsymbol{0}, \quad \boldsymbol{\mu}^{r,\omega} \geq \boldsymbol{0},\ \forall r\in T,\ \forall \omega\in\Omega \tag{13}$$

$$\boldsymbol{c}^{r,\omega} + \boldsymbol{G}^\top \boldsymbol{\pi}^{r,\omega} - \boldsymbol{\mu}^{r,\omega} = \boldsymbol{0},\ \forall r\in T,\ \forall \omega\in\Omega \tag{14}$$

$$\pi_i^{r,\omega} \cdot (\boldsymbol{G}\boldsymbol{u}^{r,\omega} - \boldsymbol{b}^r(\boldsymbol{z}))_i = 0, \quad \forall i,\ \forall r\in T,\ \forall \omega\in\Omega \tag{15}$$

$$\mu_j^{r,\omega} \cdot u_j^{r,\omega} = 0, \quad \forall j,\ \forall r\in T,\ \forall \omega\in\Omega. \tag{16}$$

To linearize the complementarity conditions (15)-(16), introduce binary variables $w_i^{r,\omega}, s_j^{r,\omega} \in \{0,1\}$ and a big constant M such that

$$(\boldsymbol{G}\boldsymbol{u}^{r,\omega} - b^r(\boldsymbol{z}))_i \geq M(w_i^{r,\omega} - 1), \quad \pi_i^{r,\omega} \leq M w_i^{r,\omega}, \quad \forall i,\ \forall r\in T,\ \forall \omega\in\Omega \tag{17a}$$

$$u_j^{r,\omega} \leq M(1 - s_j^{r,\omega}), \quad \mu_j^{r,\omega} \leq M s_j^{r,\omega}, \quad \forall j,\ \forall r\in T,\ \forall \omega\in\Omega \tag{17b}$$

The single-level KKT reformulation is:

$$\min_{\boldsymbol{z}\in Z, \boldsymbol{u},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{w},\boldsymbol{s}} \left\{ \boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega} \rho^\omega \sum_{r\in T} p^{r,\omega} (\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}^{r,\omega} : (12) - (14),\ (17a) - (17b) \right\}.$$

A.2. Strong-Duality-Based Single-Level Reformulation

The dual of the follower problem is

$$\max_{\boldsymbol{\pi}^{r,\omega}\leq 0} \left\{ (\boldsymbol{b}^r(\boldsymbol{z}))^\top \boldsymbol{\pi}^{r,\omega} : \boldsymbol{G}^\top \boldsymbol{\pi}^{r,\omega} \leq \boldsymbol{c}^{r,\omega} \right\},\ \forall r\in T,\ \forall \omega\in\Omega.$$

By strong duality, at optimality

$$(\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}^{r,\omega} \leq (\boldsymbol{b}^r(\boldsymbol{z}))^\top \boldsymbol{\pi}^{r,\omega},\ \forall r\in T,\ \forall \omega\in\Omega. \tag{18}$$

Recall that $\boldsymbol{b}^r(\boldsymbol{z}) = \boldsymbol{b}_0^r + \boldsymbol{Q}^r \boldsymbol{z}$. The term $(\boldsymbol{b}^r(\boldsymbol{z}))^\top \boldsymbol{\pi}^{r,\omega}$ contains bilinear terms $z_k \pi_m^{r,\omega}$, where z_k are binary. For each such product, define auxiliary variable $w_{k,m}^{r,\omega} = z_k \pi_m^{r,\omega}$ and add McCormick inequalities:

$$\begin{cases} w_{k,m}^{r,\omega} \geq \underline{\pi}_m^{r,\omega} z_k,\ \forall m,k,\ \forall r\in T,\ \forall \omega\in\Omega, \\ w_{k,m}^{r,\omega} \leq \overline{\pi}_m^{r,\omega} z_k,\ \forall m,k,\ \forall r\in T,\ \forall \omega\in\Omega, \\ w_{k,m}^{r,\omega} \geq \pi_m^{r,\omega} - \overline{\pi}_m^{r,\omega}(1 - z_k),\ \forall m,k,\ \forall r\in T,\ \forall \omega\in\Omega, \\ w_{k,m}^{r,\omega} \leq \pi_m^{r,\omega} - \underline{\pi}_m^{r,\omega}(1 - z_k),\ \forall m,k,\ \forall r\in T,\ \forall \omega\in\Omega, \end{cases} \tag{19}$$

where $\underline{\pi}_m^{r,\omega}$ and $\overline{\pi}_m^{r,\omega}$ are lower and upper bounds on $\pi_m^{r,\omega}$. Replace all $z_k\pi_m^{r,\omega}$ with $w_{k,m}^{r,\omega}$ in the model and obtain an equivalent MILP reformulation.

$$\min_{\boldsymbol{z}\in Z,\boldsymbol{u},\boldsymbol{\pi},\boldsymbol{w}} \Big\{\boldsymbol{\beta}^\top \boldsymbol{z} + \sum_{\omega\in\Omega}\rho^\omega \sum_{r\in T} p^{r,\omega}(\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}^{r,\omega} : (19),\ \boldsymbol{G}\boldsymbol{u}^{r,\omega} \le \boldsymbol{b}^r(\boldsymbol{z}), \boldsymbol{u}^{r,\omega}\ge \boldsymbol{0}, \boldsymbol{\pi}^{r,\omega}\le \boldsymbol{0},$$
$$\boldsymbol{G}^\top \boldsymbol{\pi}^{r,\omega} \le \boldsymbol{c}^{r,\omega}, \forall r\in T, \forall \omega\in\Omega,$$
$$(\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}^{r,\omega} \le (\boldsymbol{b}_0^r)^\top \boldsymbol{\pi}^{r,\omega} + \sum_k\sum_m Q_{m,k}^r w_{k,m}^{r,\omega}\Big\}.$$

A.3. A Valid Inequality for the Single-Level Reformulation

Building on the strong duality condition (18), Kleinert and Schmidt (2023) propose to replace the nonlinear term involving the product of $\boldsymbol{z}$ and $\boldsymbol{\pi}$ on the right-hand side with its upper bound. Since $\boldsymbol{\pi}^{r,\omega}\le \boldsymbol{0}$ and variables in $\boldsymbol{z}$ are binary, setting $\boldsymbol{z}=\boldsymbol{0}$ yields the following valid inequalities, which can be added to the KKT-based or strong-duality-based single-reformulation:

$$(\boldsymbol{v}^{r,\omega})^\top \boldsymbol{u}^{r,\omega} \le \boldsymbol{b}_0^r,\ \forall r\in T,\ \forall \omega\in\Omega. \tag{20}$$

Appendix B: Distributions of Response Search Iterations

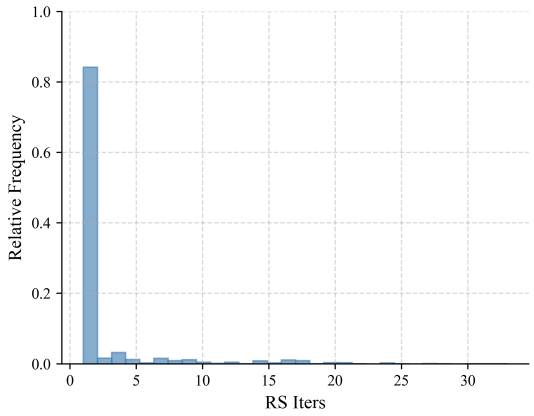
Figure 4 shows the distributions of the iterations explored in the response search algorithm for the instances reported in Table 4. Here, $|T|$ and $|\Omega|$ are the numbers of trips and scenarios, respectively. Consistent with the observations in Example 1, the number of iterations required by the algorithm is significantly smaller than the theoretical worst-case exponential enumeration, i.e., $2^{|\mathcal{H}|}=2^{12}$, of all possible network designs. Notably, the majority of follower-scenario pairs complete the search in fewer than five iterations. This practical efficiency stems from the fact that riders tend to use only a limited subset of hub legs for their trips rather than considering all possible combinations.

Appendix C: Trip Clustering

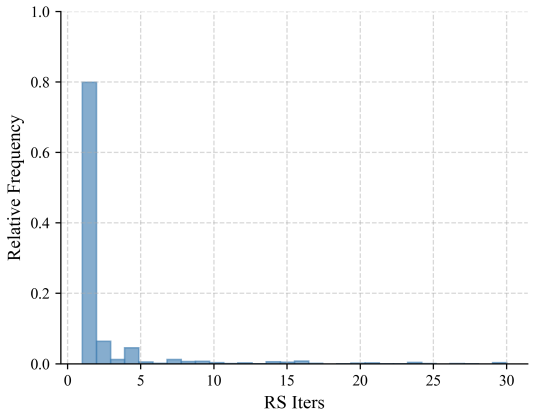
Numerical results suggest that the large number of scenarios and trips results in an exponential increment in the computational burden, making the resulting problem computationally challenging to solve. To this end, we propose a spatial-based clustering approach that reduces the total number of trips while ensuring that sufficient riders are included in the model. Firstly, the studied area is divided into several subareas according to the household income heat map (City-Data Forum 2024). Subsequently, all origins and destinations are allocated to one subarea by their geographical coordinates. Thirdly, trips originating from the subarea $\mathfrak{A}_1$ and terminating in the subarea $\mathfrak{A}_2$, where $\mathfrak{A}_1\neq\mathfrak{A}_2$ are clustered and represented by $R^{\mathfrak{A}}$, indicating a cluster of trips. To represent each cluster, a new node is introduced for identifying the origin of these trips, with its coordination $(lat^{new}, long^{new})$ in terms of its latitude and longitude calculated by its weighted average based on the ridership amount associated with these trips as follows

$$lat^{new} = \sum_{r\in R^{\mathfrak{A}}} \frac{p^r}{\sum_{r\in R^{\mathfrak{A}}} p_r} lat^r,$$

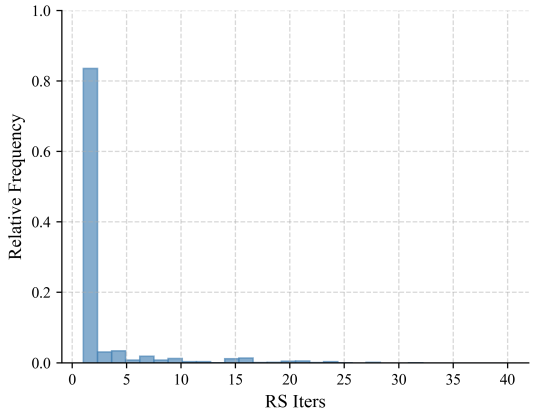
Figure 4 Distribution of the number of response search iterations of the instances in Table 4.



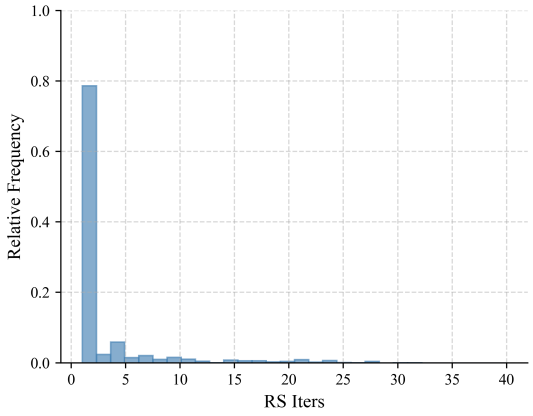
(a) $|T|$=600, $|\Omega|$=30



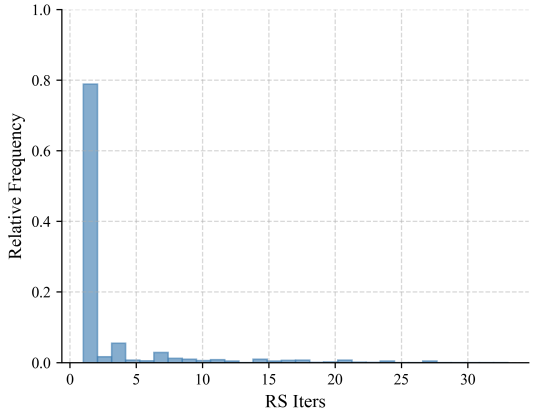
(b) $|T|$=600, $|\Omega|$=50



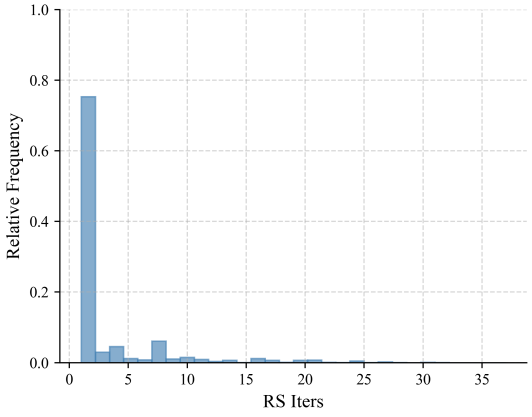
(c) $|T|$=1500, $|\Omega|$=30



(d) $|T|$=1500, $|\Omega|$=50



(e) $|T|$=2000, $|\Omega|$=30



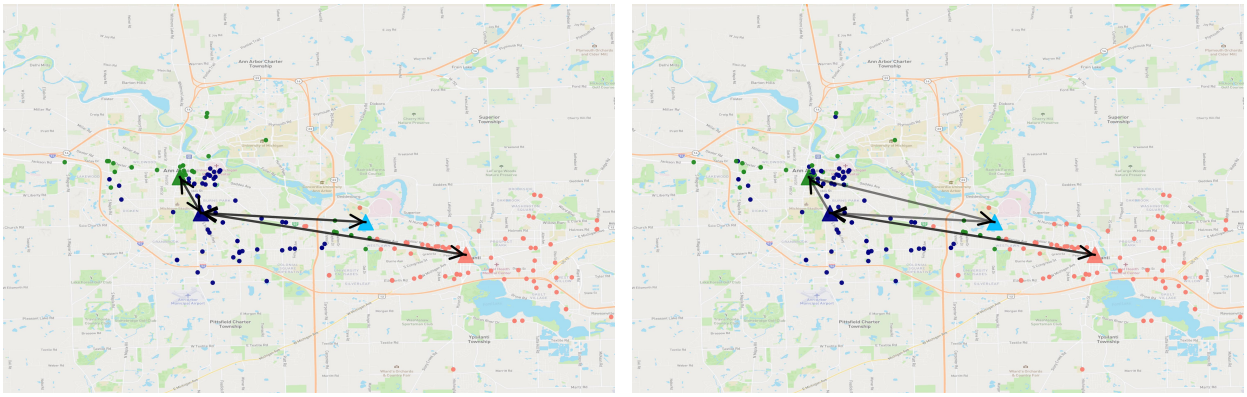
(f) $|T|$=2000, $|\Omega|$=50

$$long^{new} = \sum_{r \in R^{\mathfrak{A}}} \frac{p^r}{\sum_{r \in R^{\mathfrak{A}}} p^r} long^r,$$

where $(lat^r, long^r)$ indicates the coordination of trip r. A new destination node is introduced in a similar way. In particular, trips that originate and terminate within the same area, i.e., $\mathfrak{A}_1 = \mathfrak{A}_2$, are excluded, as such trips are assumed not to prompt the utilization of bus legs due to their short travel distance. After trip clustering, the number of stops is reduced from 1267 to 725, the number of trips is reduced from 1503 to 659, while the number of riders dropped from 2897 to 2586, taking the majority of the ridership into consideration for the subsequent multimodal transit network design problem.

Appendix D: Network Designs under Different θ^l Values

Figure 5 Network designs under different θ^l values.



(a) θ^l = 0.002 (b) θ^l = 0.005

Appendix E: Computational Performance of Algorithm 3 under Branch-and-Cut Implementation

Table 9 presents a computational comparison of the cutting-plane algorithm implemented using Gurobi's branch-and-cut framework via callback functions, both without and with RS constraints, denoted as Callback and Callback+RS, respectively. Notably, both methods reach the time limit while exploring the root node. The Callback implementation performs comparably to the iterative implementation reported in Table 5. However, incorporating RS constraints directly into the master problem in Callback+RS significantly reduces computational efficiency, resulting in substantially larger optimality gaps. In particular, no feasible solution is found within the one-hour time limit for the three largest instances. This degradation can be caused by the interplay between RS constraints and cuts added via callback, which together impede progress at the root node and prevent efficient branching.

Table 9 Computational comparison of branch-and-cut implementation (Time limit = 3600 seconds).

Trips	Scenarios	Method	Gap	# of Cuts
1000	30	Callback	7.41%	105587
		Callback+RS	18.21%	5910
1000	50	Callback	23.22%	233952
		Callback+RS	98.84%	20208
1200	30	Callback	15.93%	200248
		Callback+RS	\	18006
1200	50	Callback	11.18%	149956
		Callback+RS	\	8252
1500	30	Callback	11.60%	163872
		Callback+RS	\	11058
1500	50	Callback	11.22%	182002
		Callback+RS	\	9990