Planning of integrated mobility-on-demand and urban transit networks

Pramesh Kumar*[a], Alireza Khani[a]

[a]*Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Twin Cities, MN*

Abstract

We envision a multimodal transportation system where Mobility-on-Demand (MoD) service is used to serve the first mile and last mile of transit trips. For this purpose, the current research formulates an optimization model for designing an integrated MoD and urban transit system. The proposed model is a mixed-integer non-linear programming model that captures the strategic behavior of passengers in a multimodal network through a passenger assignment model. It determines which transit routes to operate, the frequency of the operating routes, the fleet size of vehicles required in each transportation analysis zone to serve the demand, and the passenger flow on both road and transit networks. A Benders decomposition approach with several enhancements is proposed to solve the given optimization program. Computational experiments are presented for the Sioux Falls multimodal network. The results show a significant improvement in the congestion in the city center with the introduction and optimization of the integrated transportation system. It improves the total number of served passengers and their level of service in comparison to the base optimized transit system. Finally, managerial insights for deploying such multimodal service are provided.

Keywords: Transit network design problem (TNDP), mobility-on-demand (MoD), first mile last mile (FMLM), multimodal passenger assignment, Benders decomposition

*Corresponding author
[1]Email: kumar372@umn.edu
[2]Tel: (716) 903-2366
[3]Web: http://umntransit.weebly.com/

1. Introduction

The introduction of Mobility-on-Demand (MoD) services such as Uber, Lyft, and others as transportation alternatives has created many opportunities as well as challenges. On one hand, they provide a seamless mobility service with just a few taps on a cellphone application. On the other hand, it has increased congestion in densely populated areas due to an increase in the relocation and pickup trips made by the participating drivers in the network (Laris 2019). Furthermore, the transportation agencies envision the introduction of Autonomous Vehicles as a shared mobility service in the near future (Motavalli 2020), which would lead to severe congestion in densely populated areas as predicted by various simulation studies (Levin and Boyles 2015, Fagnant et al. 2016, Levin et al. 2017).

Public transportation, which can carry multiple passengers, is widely considered as a practical solution to the congestion problem by reducing vehicle-miles traveled (VMT) on roads (Aftabuzzaman et al. 2015). However, due to its fixed routes and schedules, limited network coverage, and waiting time, sometimes, it is less attractive to travelers in comparison to the auto mode. The limited network coverage makes it difficult or sometimes impossible to access transit service in some areas. This inaccessibility problem is also known as the *first mile/last mile (FMLM) problem for transit*. The problem is commonly faced by travelers commuting from low-density areas where transit service is not available or less frequent because of the economic in-viability of providing such service.

A few studies have argued that the Mobility-on-demand service provided using autonomous vehicles would become a competitor of public transit mode (Chen and Kockelman 2016, Levin and Boyles 2015, Mo et al. 2020), reducing its ridership, and other studies have even raised the question of whether urban mobility is possible without the classical public transit service (OECD 2015, Mendes et al. 2017). However, Salazar et al. 2018 showed that the integration of the MoD system with transit could help in achieving better results, such as a significant reduction in travel time, emissions, and costs as compared to the standalone MoD system. Through the current research, we also envisage an integrated MoD and transit system that aims to achieve the following potential benefits:

1. Providing fast and reliable mobility in low-density areas (i.e., by providing a first mile/last mile service) by means of characteristics of MoD service such as demand responsiveness, fleet repositioning, and reachability.
2. Allocation of resources from less congested areas to providing high-frequency transit service in congested areas through such integration.
3. Using existing transit infrastructure to reduce the number of vehicles needed for serving trips.

4. Reducing congestion and carbon emissions in the network, improving the mobility of travelers, and reducing the overall cost of providing transit service.

To achieve the above-mentioned benefits, we focus on the strategic planning of the transportation network that allows for intermodal trips with the first or last leg of the trips being served by the MoD service. To be specific, we try to answer questions such as which transit routes to operate when MoD vehicles are deployed to serve the FMLM connection, what should be the size of the vehicle fleet to be deployed, and what should be the frequency of operating transit routes. We attempt to answer these questions and make the following contributions through this article:

1. Propose a passenger assignment model that predicts the travel behavior of passengers in a multimodal network. This step extends the idea of a hyperpath transit assignment model proposed by Spiess and Florian 1989 to a multimodal transportation system with on-demand services.

2. Develop an optimization model to decide which transit routes to operate, frequency of operating transit routes, and MoD fleet size required to serve the FMLM of trips.

3. Develop a fast Benders decomposition implementation that uses efficient cutting planes to solve the large instances of the current problem.

4. Conduct numerical experiments to show the efficacy of the proposed model and solution methods and discuss the steps to implement such service in practice.

2. Related work

The passenger journeys that consist of auto, as well as transit mode, create a new mode of transportation known as *intermodal* or *multimodal* transportation. The research on modeling multimodal transportation has been an active area of research for several decades (Wilson 1972). Many of these studies are focused on solving the transit FMLM problem by designing a multimodal transportation system. This includes designing a demand responsive transit feeder service (Wang 2017, Maheo et al. 2017, Cayford and Yim 2004, Koffman 2004, Lee and Savelsbergh 2017, Quadrifoglio et al. 2008, Shen and Quadrifoglio 2012, Li and Quadrifoglio 2009), using park-and-ride facilities (Nassir et al. 2012, Khani et al. 2012, Webb and Khani 2020), and integrating ridesharing and transit (Masoud et al. 2017, Stiglic et al. 2018, Bian and Liu 2019, Ma et al. 2019, Chen et al. 2020, Kumar and Khani 2021).

Recently, the studies are being focused on modeling the integration of MoD and transit service for future mobility. They can be divided into two categories: simulation-based and optimization-based approaches. Under a set of assumptions on vehicle operations operations and dispatching strategies, the simulation-based studies simulate the passenger flow to assess the service quality of providing such mobility service (Gurumurthy et al. 2020). By using a four-step travel demand simulation model, Levin and Boyles 2015 predicted that the transit ridership will decrease and the number of personal vehicles will sharply increase as a result of the repositioning of vehicles resulting in congestion on the network. Vakayil et al. 2017 developed a simulation model that accounts for transit frequency, transfer costs, and MoD fleet re-balancing to use MoD as the FMLM solution to the transit mode. Their results show that such an integrated system can reduce VMT in the network by up to 50%. Mendes et al. 2017 developed an event-based simulation model to compare the performance of the MoD system with the light rail system under the same demand patterns, alignment, and operating speed. They found that 150 vehicles with 12 passenger capacity would be needed to match the 39-vehicle light rail system if operated as a demand responsive system. Similar findings were also shown by the simulation model developed by Basu et al. 2018. They showed that the introduction of MoD will act as the competitor of mass transit, however, to reduce congestion and maintain a sustainable urban transportation system, it cannot replace mass transit. Shen et al. 2018 also proposes and simulates an integrated autonomous vehicle and public transportation system based on the fixed modal split assumption. Using Singapore's organizational structure and demand characteristics, they propose to preserve high-demand bus routes while re-purposing low-demand bus routes and using shared MoD as an alternative. They found that the integrated system has the potential of serving the trips with less congestion, less passenger discomfort, and economically viable service. Wen et al. 2018 included mode choice and various vehicle capacities and hailing strategies in an agent-based model to provide insights into fleet sizing and frequency of transit routes for the integrated system. A few studies have used an optimization-based approach to developing an integrated passenger flow model. Salazar et al. 2018 developed a network flow model for intermodal service that couples the interaction between MoD and transit by maximizing social welfare. Using this model, they proposed a tolling scheme for this intermodal system that helps in reducing the travel time, costs, and emissions as compared to standalone vehicle mode. Liu et al. 2019 used Bayesian optimization to predict the mode choice of passengers in such a multimodal transportation system.

The above-cited studies show that an integrated MoD and transit system can provide an efficient mode of transportation that is sustainable, fast, eco-friendly, and economically viable. The design of such a system requires solving a *multimodal transportation network design problem* that can decide various aspects of MoD and transit modes. The problem of designing transit routes and their corresponding frequencies, which is commonly referred to as the Transit Network Design Problem (TNDP) or Line Planning Problem (LPP) in the literature, is itself a complex problem (Ceder and Wilson 1986, Baaj and Mahmassani 1991). There has been a significant amount of research in modeling TNDP and developing solution algorithms for it. For a review on transit network design literature, we refer the interested reader to Guihaire and Hao 2008. Some aspects of the multimodal network design problem have been explored in a related research problem known as *hub and arc location problem* (Mahéo et al. 2019, Campbell et al. 2005*a*,*b*). For example, Mahéo et al. 2019 proposed the design of a hub and shuttle public transit system in Canberra. They formulated a mixed-integer program to design high-frequency bus routes between key-hubs, where the first mile or last mile of trips is covered by the shuttles. However, the hub and arc location problem have a major limitation of not able to capture passenger behavior in the transit network. Recently, a couple of studies have proposed models for the transit network design in the context of integrated MoD and transit system (Manser 2017, Pinto et al. 2020, Steiner and Irnich 2020). Pinto et al. 2020 develops a bi-level optimization model to design a transit network integrated with MoD service. The upper-level optimization problem modifies the frequency of the transit routes and determines the fleet size of MoD service and the lower-level model simulates the passenger trajectories based on a simulation-based traveler assignment model. Due to the complexity of the model, they presented a heuristic approach to solving the current problem. Steiner and Irnich 2020 presents various aspects of this problem and develops a path-based mixed-integer programming model to decide which sections of the transit routes to operate and locate the transfer stops to allow for intermodal trips in the network. Due to an enormous number of possible paths in the network, they solve the current model using a branch-and-price approach.

The design of an integrated MoD and transit system is an important problem that can influence the future mobility of travelers. Recent studies have made important contributions to this complex problem but have several limitations, which we attempt to address in the current study. The motivation of the current research is outlined in the following points:

1. Before designing the integrated system, we should understand how passengers would behave in an integrated system. It is common for studies to use the classic multi-commodity flow model to predict the behavior of travelers in the network design. This may be true if passenger trajectories are completely influenced by the mobility provider. However, this is certainly not applicable in the case of transit systems when passengers try to reduce the expected travel time based on waiting time, travel time, and fare. Through this study, we extend the idea of hyperpath passenger assignment for a multimodal transportation system.

2. We develop a mixed-integer optimization model that incorporates the multimodal passenger assignment and evaluates various aspects of an integrated system. The optimization program is difficult to solve, and we need efficient techniques to solve this problem. For this purpose, an exact method based on the Benders Decomposition is proposed to solve the large-scale instances of the problem. The method improves the classic Benders decomposition strategy by precluding the infeasibility cuts and including new cuts, such as disaggregated cuts, multiple cuts, and clique/cover cuts.

The rest of the article is structured as follows. §3 discusses the notations and definitions used in this article. Then, we present the multimodal passenger assignment model, which is incorporated in the design model of the integrated MoD and transit system in §4. The solution algorithm to solve the design model is discussed in §5, which is followed by the results of numerical experiments conducted on Sioux Falls network. Finally, conclusions and recommendations for future research are presented in §7.

3. Preliminaries and Background

In this section, we get familiarize ourselves with the notations and concepts to be used in this article. Let us begin by considering a multimodal transportation network characterized by a digraph $G(N, A)$, where N denotes the set of nodes that includes road intersections N_R, transit stops/stations N_T, and centroids of traffic analysis zones Z[4] and A denotes the set of links. We associate every node $i \in N$ in the network with exactly one zone $Z(i)$. The set of links coming out and going into a node $i \in N$ are denote by $FS(i) = \{(i, j) : (i, j) \in A\}$ and $BS(i) = \{(j, i) : (j, i) \in A\}$ respectively. Let $\mathfrak{d} : N \times N \mapsto \Re_+$ be the distance function between two nodes in the network. Depending on the mode, the links are also divided into three categories, namely transit, road, and walking links represented by $A_T, A_R,$ and A_W respectively. Let $O \subset Z$ and $D \subset Z$ be the subsets of centroids representing the origins and destinations respectively. The demand between various origin-destination pairs is represented by $\{d_{od}\}_{(o,d)\in O\times D}$. The overall network can be divided into three sub-networks which are described below:

1. *Transit network*: The transit network is characterized by the subgraph $G_T(N_T, A_T)$ which consists of a set of candidate transit lines/routes denoted by the set L. The terms "route" and "line" are used interchangeably throughout this article. Each line $l \in L$ is composed of a set of stops $N_T^l \subset N_T$ which are connected by edges $A_T^l \subset A_T$. The network also consists of transfer links A_T^{tr} between two nodes if the walking distance between those is less than the acceptable walking distance ζ (say 0.75mi), i.e., $A_T^{tr} = \{(n_1, n_2) \in N \times N : n_1 \in N_T^{l_1}, n_2 \in N_T^{l_2}$ for some $l_1, l_2 \in L$ s.t. $l_1 \neq l_2$ and $\mathfrak{d}(n_1, n_2) \leq \zeta\}$.

2. *Road network*: The road network is characterized by the subgraph $G_R(N_R, A_R)$, where N_R denotes the set of nodes and A_R denotes the set of links in the road network.

3. *Walking links*: The walking links consists of access, egress, and mode transfer links. The *access* and *egress links* are defined as $A^a = \{(n_1, n_2) \in Z \times (N_T \cup N_R) : \mathfrak{d}(n_1, n_2) \leq \zeta\}$ and $A^e = \{(n_1, n_2) \in (N_T \cup N_R) \times Z : \mathfrak{d}(n_1, n_2) \leq \zeta\}$ respectively. Similarly, the *mode transfer links* are defined as $A^m = \{(n_1, n_2) \in N_R \times N_T : \mathfrak{d}(n_1, n_2) \leq \zeta\} \cup \{(n_1, n_2) \in N_T \times N_R : \mathfrak{d}(n_1, n_2) \leq \zeta\}$. The access and egress walking links connect the centroids of various zones with the road/transit nodes and vice-versa, whereas mode transfer links are used to transfer between nodes of various modes.

[4]A *traffic analysis zone* (TAZ) or simply a zone is a geographical area where the demand is assumed to be concentrated on its centroid.

3.1. Costs

There is a subset of nodes in the network where passengers have to wait for the service. The collection of head nodes of links in the sets A^a, A_T^{tr}, and A^m constitutes the *waiting nodes* N^w. Let us assume that $c : A \mapsto \Re_+$ and $w : N^w \mapsto \Re_+$ denote the cost (e.g., walking time, in-vehicle time, and fare) associated with the links in A and waiting time associated with the nodes in N^w respectively. The cost of links is known beforehand (and is computed by adding the travel time and possible fare multiplied by the value of time). On the other hand, the wait time depends on the availability of MoD or transit service.

3.2. Waiting time computation

Unlike a personal vehicle, the MoD or transit service is not readily available, and passengers have to wait to access these services. So, it is important to quantify the expected wait time of these services, the computation of which is discussed below:

3.2.1. MoD service

We assume MoD operations in a network as a queuing system to compute the average waiting time experienced by the passengers to access such service. The average wait time may not be justified for the planning of day-to-days operations but can be used to approximate the actual wait time experienced by the passengers for long-term strategic planning of the network, which is the focus of the current study. Therefore, we consider a stationary state of an MoD system, where the number of waiting customers $\mathcal{C}$ and vacant vehicles $\mathcal{V}$ are time-invariant. Using the Cobb-Douglas production function, the matching time between the customers and the vacant vehicles can be expressed as a function of $\mathcal{C}$ and $\mathcal{V}$.

$$m^{c-v} = \mathcal{A}(\mathcal{V})^{\alpha_1}(\mathcal{C})^{\alpha_2} \tag{1}$$

where, α_1 and α_2 are defined as the elasticities of the matching function and $\mathcal{A}$ is a parameter specific to a zone, which is a function of the market area divided by the running speed in that zone (Zha et al. 2016). According to Little's law, the long-term average number of customers/drivers in a stationary system is equal to the long-term average arrival rate Q multiplied by the average wait time (w^c/w^t) that a customer/driver spends in the system before being matched (Zha et al. 2016).

$$\mathcal{V} = Qw^t \tag{2}$$

$$\mathcal{C} = Qw^c \tag{3}$$

Using (3) and assuming $\alpha_1 = \alpha_2 \approx 1$ (Douglas 1972), we can represent the stationary state ($m^{c-v} = Q$) as below:

$$Q = \mathcal{A}\mathcal{V}(Qw^c) \tag{4}$$

$$\implies w^c = \frac{1}{\mathcal{A}\mathcal{V}} \tag{5}$$

Equation (5) shows that the average waiting time of customers waiting in a zone to access the MoD service is a function of the vacant number of vehicles. To achieve the desired level of service (i.e., average waiting time), a transportation agency needs to provide $\mathcal{V}$ vehicles at any point in time.

3.2.2. Transit service

Let us now discuss the wait time computation to access transit service at the head node of an access or transfer link in the transit network. Let $f : A_T \mapsto \mathfrak{R}$ be the frequency of the transit line associated with various links of the transit network. Let $\mathfrak{g}_i(w)$ be the probability distribution function of the waiting time for line i. According to Larson and Odoni 1981, for the passengers arriving randomly at a node, the probability density function of the waiting time of line i is related to the headway or bus inter-arrival time distribution $\mathfrak{h}_i(h)$ as:

$$\mathfrak{g}_i(w) = \frac{\int_w^\infty \mathfrak{h}_i(h)dh}{\mathbb{E}[\mathfrak{h}_i]} \tag{6}$$

To evaluate the waiting time distribution, we make the following assumptions:

Assumption 1. *The inter-arrival time of a transit line $i \in L$ follows an exponential distribution with rate f_i.*

Assumption 2. *Passengers want to minimize the expected wait time to get to their destination. Therefore, at any node, passengers waiting to be served by the transit service have selected a list of attractive transit lines that can help them to get to their destination.*

Both assumption 1 and 2 are common in the transit assignment literature (e.g., see Desaulniers and Hickman 2007). By using the assumption 1 and equation (6), one can evaluate the distribution function of the wait time $\mathfrak{g}_\mathfrak{i}(w)$ as:

$$\mathfrak{g}_\mathfrak{i}(w) = f_i e^{-f_i w}, w \geq 0 \tag{7}$$

Proposition 1. *(Spiess and Florian 1989, Gentile et al. 2005) Assuming that a passenger waiting at node $n \in N^w$ is served by the set of attractive transit lines $FS^*(n)$ and let $\mathfrak{F} = \sum_{j \in FS^*(n)} f_j$. With assumptions 1 and 2, the following holds:*

1. *The probability that a passenger would choose transit line $i \in FS^*(n)$ is given by*

$$P_i = \frac{f_i}{\mathfrak{F}} \tag{8}$$

2. *The expected wait time conditional to boarding line $i \in FS^*(n)$ is given by*

$$EW_i = \frac{f_i}{\mathfrak{F}^2} \tag{9}$$

3. *The probability of wait time at node n follows an exponential distribution with rate $\mathfrak{F}$. Therefore, the expected wait time at stop n is given by $EW_n = \frac{1}{\mathfrak{F}}$.*

Proof. See Appendix A. $\square$

3.2.3. Combined MoD and transit service wait time

Before discussing the computation of the expected wait time involving both modes, we need to make an assumption about the wait time distribution of MoD service by utilizing the value of the average wait time of MoD service calculated in equation (5).

Assumption 3. *The wait time distribution of MoD service for passengers waiting at node n follows an exponential distribution with rate $f_{MoD} = \mathcal{A}_{Z(n)}\mathcal{V}_{Z(n)}$, where $Z(n)$ is the zone associated to node n and $\mathcal{V}_{Z(n)}$ is the number of vehicles deployed in zone $Z(n)$.*

A passenger waiting at the head node of an access link faces the choice between MoD or transit mode. This is because the wait time of both services can vary based on the frequency provided, and a passenger will include one or both modes in their strategy to reduce the overall expected cost. This assumption simplifies the operation of MoD service as a transit service available at any stop of the network. The following proposition evaluates the expected wait time of that passenger.

Proposition 2. *Given that the waiting time for transit and MoD mode follow an exponential distribution with rate $\mathfrak{F}$ and f_{MoD} respectively and $\mathbb{F} = \mathfrak{F} + f_{MoD}$, the following holds:*

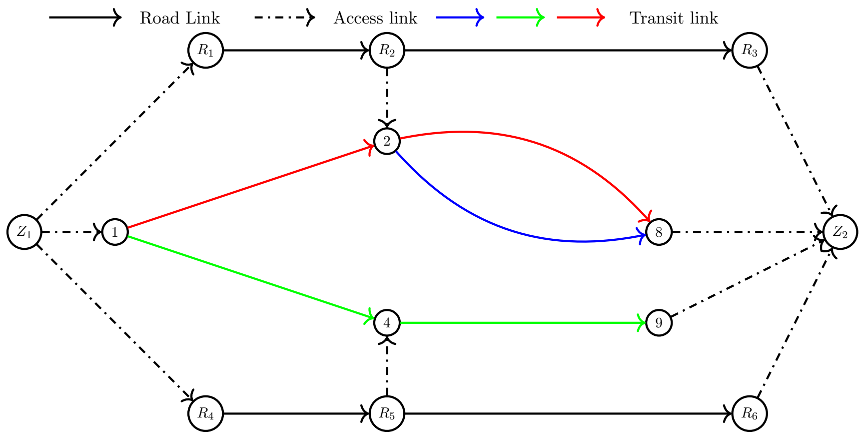
1. *The probabilities of taking transit and MoD are given by $P_{MoD} = \frac{f_{MoD}}{\mathbb{F}}$ and $P_{transit} = \frac{\mathfrak{F}}{\mathbb{F}}$ respectively.*

2. *The expected wait time of the passenger departing from an access node n served by both MoD and transit service is given by $EW_n = \frac{1}{\mathbb{F}}$*

Proof. See Appendix A. $\square$

To get more insights into the wait time computation, let us consider an example. Figure 1(a) shows an illustration of a multimodal transportation network. It consists of 2 zones, 6 nodes and 8 links as part of the road network, and 5 nodes and 10 links as part of the transit network. The transit network has 3 transit lines (color-coded) whose frequencies are shown in Figure 1(b). There are 100 and 50 vehicles deployed in zone 1 and 2 respectively. By using Prop 1, we can evaluate the probability of passengers taking various transit lines in the network. For example, the probabilities of choosing red line and green line at stop 1 are $\frac{1/6}{1/6+1/2} = 0.25$ and $\frac{1/2}{1/6+1/2} = 0.75$ respectively. The expected wait time at stop 1 is equal to $12/8 = 1.5$ minutes. Similarly, using Proposition 2, the probabilities of choosing MoD and transit at Z_1 are $\frac{0.0017*100}{0.0017*100+8/12} = 0.2$ and 0.8 respectively (assuming $\mathcal{A}_1 = 0.0017$). The overall expected wait time at node Z_1 is 1.19 minutes which is less than 1.5 minutes by only considering transit service as part of the strategy.

We further use Proposition 1 and 2 to formulate the multimodal passenger assignment model. For this purpose, we extend the frequency-based transit assignment model proposed by Spiess and Florian 1989 to a multimodal transportation system. Before moving forward, we must make the following assumptions:



(a) Multimodal network

Line	**Frequency**
Red	1/6
Green	1/2
Blue	1/3

Zone	**Vehicles**
Z_1	100
Z_2	50

(b) Frequency and number of vehicles

Figure 1: An illustrative example of a multimodal network

Assumption 4. *(a) Ridepooling is not allowed, i.e., the MoD service serves one passenger at a time.*

(b) The transit lines are assumed to have unlimited capacity.

(c) Passengers want to reduce their expected generalized travel cost consisting of travel time, wait time, and fare to get to their destination.

The ridepooling problem requires matching of customers using a specific algorithm. This is an important aspect to accurately estimate the cost of day-to-day operations. Nevertheless, ignoring ridepooling will give us an upper bound on the number of vehicles required to serve various zones. The modeling of passenger behavior while incorporating the capacity constraints (congestion) is a difficult problem. The congestion is important to consider since it causes denied boarding, which leads to increased waiting time, travel time, and discomfort. Several authors have tried to include congestion into frequency-based transit assignment models through various approaches, namely, discomfort function (Spiess and Florian 1989), effective frequency (De Cea and Fernández 1993, Cominetti and Correa 2001, Cepeda et al. 2006, Leurent et al. 2014), and failure-to-board probabilities (Kurauchi et al. 2003). Despite the effort, there is no tractable closed-form of congested frequency-based transit assignment model. On the other hand, it would not be ideal to include transit vehicle capacity constraints into the assignment program (e.g., in Szeto and Jiang 2014) because doing so may lead to unrealistic passenger behavior, which previous studies on congested frequency-based transit assignments were trying to avoid. Therefore, we use an uncapaciated assignment for the design problem. Assumption 4(c) is a common in the assignment literature. The relaxation of above assumptions are research topics in their own right, therefore, a discussion on possible ways to relax them is provided in §7. To proceed further, let us define a variable $\{g_{ik}\}_{i\in N,k\in D}$ as below:

$$g_{ik} = \begin{cases} d_{ik}, & \text{if } i \neq k, (i,k) \in O \times D \\ -\sum_{o\in O} d_{ok}, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases}$$

Furthermore, let us denote v_{ak} and W_{ik} as the flow of passengers on link $a \in A$ and waiting at node $i \in N^w$ resp. destined to $k \in D$. The assignment optimization program is presented below:

$$\underset{v,W}{\text{minimize}} \quad \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) \tag{10a}$$

$$\text{subject to} \quad \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D \tag{10b}$$

$$v_{ak} \leq f_a W_{ik}, \forall a \in FS(i) : a \in A_T, \forall i \in N^w, \forall k \in D \tag{10c}$$

$$v_{ak} \leq \mathcal{A}_{Z(i)} \mathcal{V}_{Z(i)} W_{ik}, \forall a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D \tag{10d}$$

$$v_{ak} \geq 0, \forall a \in A, \forall k \in D \tag{10e}$$

The assignment program (10) minimizes the total expected link costs and wait time at waiting nodes experienced by the passengers in a mulimodal network subject to the flow conservation constraint at each node (10b), flow proportion constraints (10c)-(10d), and the non-negativity and binary constraints (10e). The flow proportion constraints uses the probability of selecting an option $a \in FS(i)$ (if that option is a part of the strategy of the passengers traveling to destination $k \in D$) and multiplies it with the number of passengers waiting at that node. Note that the probability of selecting an option (MoD or transit line) is calculated in Proposition 2.

4. Design of an integrated MoD and transit system

In this section, we present an optimization model incorporating the assignment program proposed in previous section for the design of an integrated MoD and transit system. The optimization program is formalized as a Mixed Integer Non-linear Program (MINLP). In this model, we determine which transit routes to keep operating among the current transit routes in the city network, decide the optimal frequency of those operating routes, and finally, determine the fleet size of vehicles required to provide MoD service in various zones. Note that one can also include new candidate transit routes as part of the design plan. The sets, parameters, and decision variables for the optimization model are summarized in Table 1.

The design of an integrated transit and MoD system should consider both passenger and operator perspectives. The operator's perspective is to provide the service at minimum cost, and the passengers' perspective is to minimize the overall cost of travel (including travel time, wait time, and

Table 1: Sets, decision variables and parameters used in the design model

<u>Sets</u>

$$
\begin{array}{rcl}
\mathfrak{B} & \triangleq & \text{Set of binary values} \\
L & \triangleq & \text{Set of candidate transit lines} \\
\Theta = \{2,3,4,6,12\} & \triangleq & \text{Set of possible frequencies of a line (buses/hr)} \\
\Omega = \{0.01,50,100,200,500\} & \triangleq & \text{Set of possible number of vehicles deployed in a zone}
\end{array}
$$

<u>Parameters</u>

$$
\begin{array}{rcl}
\bar{B} & \triangleq & \text{Total number of buses available} \\
\bar{F} & \triangleq & \text{Total number of vehicles available}
\end{array}
$$

<u>Decision Variables</u>

$$
\begin{array}{rcl}
x_l & = & \begin{cases} 1, & \text{if line } l \in L \text{ is decided to keep operating} \\ 0, & \text{otherwise} \end{cases} \\
y_{lf} & = & \begin{cases} 1, & \text{if frequency } f \in \Theta \text{ is adopted for line } l \in L \\ 0, & \text{otherwise} \end{cases} \\
\mathcal{N}_{zn} & = & \begin{cases} 1, & \text{if a fleet of size } n \in \Omega \text{ is deployed in zone } z \in Z \\ 0, & \text{otherwise} \end{cases} \\
v_{ak} & = & \text{Flow of passengers on link } a \in A \text{ destined to } k \in D \\
W_{ik} & = & \text{Wait time of passengers waiting at node } i \in N^w \text{ destined to } k \in D
\end{array}
$$

fare). Based on these perspectives, the design optimization model is presented as (11). The objective function is the sum of the total expected travel cost and wait time experienced by the passengers in the network. The mapping $\mathcal{B} : L \times \Theta \mapsto \mathbb{N}$ used in (11b) is defined as $\mathcal{B}(l, f) = \left(f \times \sum_{a \in A_T^l} 2t_a\right)$, which describes that the number of buses required to provide frequency $f \in \Theta$ for a line $l \in L$ is equal to the product of the frequency and round trip travel time. (11b) constrain the total number of buses needed to be less than or equal to $\bar{B}$, which can be evaluated for a given budget. (11c) describes the flow conservation constraints at every node for every destination. For a given MoD and bus fleet assignment, (11d)-(11e) describe the passenger flow on each link based on the frequency of the bus route and MoD service. A frequency value can be assigned to a route if that route is decided to keep operating as constrained by (11f). (11g) describe that exactly one of the fleet sizes can be adopted for each zone. (11h) constrain the required number of vehicles to be less than or equal to $\bar{F}$. Finally, (11i), (11j) and (11k)-(11m) are the non-negativity constraints of the flow, wait time being free variables, and binary constraints of design variables respectively. One can also incorporate other constraints related to the budget of operating MoD and transit service but for the sake of simplicity, we do not include them here.

$$
\begin{aligned}
\underset{v,W,x,y,\mathcal{N}}{\text{minimize}} \quad & \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) && \text{(11a)}\\
\text{subject to} \quad & \sum_{l\in L}\sum_{f\in\Theta} \mathcal{B}(l,f)\times y_{lf} \leq \bar{B} && \text{(11b)}\\
& \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i\in N, \forall k\in D && \text{(11c)}\\
& v_{ak} \leq \left(\sum_{f\in\Theta} f y_{l(a)f}\right) W_{ik}, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(11d)}\\
& v_{ak} \leq \mathcal{A}_{Z(i)}\left(\sum_{n\in\Omega} n\mathcal{N}_{Z(i)n}\right) W_{ik}, \forall a\in FS(i): a\in A_R, \forall i\in N^w, \forall k\in D && \text{(11e)}\\
& \sum_{f\in\Theta} y_{lf} = x_l, \forall l\in L && \text{(11f)}\\
& \sum_{n\in\Omega} \mathcal{N}_{zn} = 1, \forall z\in Z && \text{(11g)}\\
& \sum_{z\in Z}\left(\sum_{n\in\Omega} n\mathcal{N}_{zn}\right) \leq \bar{F} && \text{(11h)}\\
& v_{ak} \geq 0, \forall a\in A, \forall k\in D && \text{(11i)}\\
& W_{ik} \text{ free }, \forall i\in N^w, \forall k\in D && \text{(11j)}\\
& x_l \in \mathfrak{B}, \forall l\in L && \text{(11k)}\\
& y_{lf} \in \mathfrak{B}, \forall f\in\Theta, \forall l\in L && \text{(11l)}\\
& \mathcal{N}_{zn} \in \mathfrak{B}, \forall n\in\Omega, z\in Z && \text{(11m)}
\end{aligned}
$$

The optimization program (11) is a mixed-integer non-linear program (MINLP). The non-linearity arise from the constraints (11d)-(11e). It is computationally difficult to solve this program for large instances, which can be attributed to the integer constraints (11k)-(11m) and the bilinear constraints (11d)-(11e). The bilinear constraints are particularly difficult to handle due to the non-convex nature even if the integrality constraints of the involved variables are relaxed. Fortunately, in this case, the non-convexity arises due to the product of continuous and binary variables, which can be exactly relaxed by employing McCormick relaxations. Let $t_{faik} = y_{l(a)f} W_{ik}, \forall f\in\Theta, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D$ and $\omega_{ink} = \mathcal{N}_{Z(i)n} W_{ik}, \forall n\in\Omega, \forall i\in N^w\cap N_R, \forall k\in D$. Further, let us assume that there exists a finite upper and lower bound on the variable W_{ik} , i.e., $\underline{W}_{ik} \leq W_{ik} \leq \overline{W}_{ik}$. Then, t_{faik} and ω_{ink} can be expressed as the set of linear constraints

(12a)-(12d) and (13a)-(13d) respectively:

$$\begin{aligned}
\overline{W}_{ik} - W_{ik} + t_{faik} - \overline{W}_{ik} y_{l(a)f} &\geq 0 && \text{(12a)}\\
\overline{W}_{ik} y_{l(a)f} - t_{faik} &\geq 0 && \text{(12b)}\\
t_{faik} - \underline{W}_{ik} y_{l(a)f} &\geq 0 && \text{(12c)}\\
W_{ik} - \underline{W}_{ik} - t_{faik} + \underline{W}_{ik} y_{l(a)f} &\geq 0 && \text{(12d)}
\end{aligned}$$

$$\begin{aligned}
\overline{W}_{ik} - W_{ik} + \omega_{ink} - \overline{W}_{ik} \mathcal{N}_{Z(i)n} &\geq 0 && \text{(13a)}\\
\overline{W}_{ik} \mathcal{N}_{Z(i)n} - \omega_{ink} &\geq 0 && \text{(13b)}\\
\omega_{ink} - \underline{W}_{ik} \mathcal{N}_{Z(i)n} &\geq 0 && \text{(13c)}\\
W_{ik} - \underline{W}_{ik} - \omega_{ink} + \underline{W}_{ik} \mathcal{N}_{Z(i)n} &\geq 0 && \text{(13d)}
\end{aligned}$$

5. Solution methodology

After relaxing the bilinear constraints (11d)-(11e), the resulting model is a Mixed Integer Linear Program (MILP). The program is still difficult to solve efficiently for large instances. However, the structure of the problem allows us to use decomposition techniques such as Benders decomposition to efficiently solve it. In this section, we present the details of the Benders reformulation for this problem, along with the proposed algorithmic enhancements.

5.1. Benders Reformulation

Benders decomposition (Geoffrion 1972) is an elegant way of solving a large scale MILP by iteratively solving two simpler subproblems: the relaxed master problem (RMP), which is a relaxation of the original problem and a subproblem (SP) which provides inequalities/cuts to strengthen the RMP. The subproblem should possess strong duality properties. Let us consider the network design problem described in the previous section. For a given feasible value of design decision variables $\hat{x}, \hat{y}, \hat{\mathcal{N}}$ and with $\underline{W}_{ik} = 0$ (wait time cannot be negative), we can rewrite the original problem as a *Benders subproblem* (14).

$$
\begin{aligned}
z^{SP}(\hat{x},\hat{y},\hat{\mathcal{N}}) = \min_{v,W,\omega,t} \quad & \sum_{k\in D}\left(\sum_{a\in A} c_a v_{ak} + \sum_{i\in N^w} W_{ik}\right) && \text{(14a)}\\
\text{subject to} \quad & \sum_{a\in FS(i)} v_{ak} = \sum_{a\in BS(i)} v_{ak} + g_{ik}, \forall i\in N, \forall k\in D && \text{(14b)}\\
& v_{ak} \le \left(\sum_{f\in\Theta} f t_{faik}\right), \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(14c)}\\
& W_{ik} - t_{faik} \le \overline{W}_{ik}(1-\hat{y}_{l(a)f}), \forall f\in\Theta, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(14d)}\\
& t_{faik} \le \overline{W}_{ik}\hat{y}_{l(a)f}, \forall f\in\Theta, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(14e)}\\
& W_{ik} - t_{faik} \ge 0, \forall f\in\Theta, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(14f)}\\
& v_{ak} \le \mathcal{A}_{Z(i)}\left(\sum_{n\in\Omega} n\omega_{ink}\right), \forall a\in FS(i): a\in A_R, \forall i\in N^w, \forall k\in D && \text{(14g)}\\
& W_{ik} - \omega_{ink} \le \overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n}), \forall n\in\Omega, \forall i\in N^w\cap N_R, \forall k\in D && \text{(14h)}\\
& \omega_{ink} \le \overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}, \forall n\in\Omega, \forall i\in N^w\cap N_R, \forall k\in D && \text{(14i)}\\
& W_{ik} - \omega_{ink} \ge 0, \forall n\in\Omega, \forall i\in N^w\cap N_R, \forall k\in D && \text{(14j)}\\
& v_{ak} \ge 0, \forall a\in A, \forall k\in D && \text{(14k)}\\
& t_{faik} \ge 0, \forall f\in\Theta, \forall a\in FS(i): a\in A_T, \forall i\in N^w, \forall k\in D && \text{(14l)}\\
& \omega_{ink} \ge 0, \forall n\in\Omega, \forall i\in N^w\cap N_R, \forall k\in D && \text{(14m)}
\end{aligned}
$$

Let $\mathcal{X}^{SP} = \{(v,W,\omega,t) : (14b)-(14m)\}$ be the feasible region of the Benders subproblem (14). Further, let us denote $\mathcal{X}^{MA} = \{(v,W) : (10b)-(10e)\}$ as the feasible region of the multimodal assignment linear program. We can show the following result:

Proposition 3. *The projection of the feasible region of the subproblem* (14) *on to the space of* v *and* W *is same as the feasible region of the multimodal assignment problem* (10) *i.e.,*

$$proj_{v,W}\mathcal{X}^{SP} = \mathcal{X}^{MA}$$

Proof. See Appendix A. □

Proposition 3 shows that one can use the efficient Spiess and Florian 1989's primal-dual algorithm designed for the transit assignment problem to solve the current Benders subproblem. To speed up the process of Benders decomposition by avoiding feasibility cuts, we need to put some restrictions so that the subproblem (14) is always feasible.

Proposition 4. *Given that* $0 \notin \Omega$ *and* $G_R(N_R, A_R)$ *is connected, then* $\mathcal{X}^{SP}$ *is non-empty for any given feasible value of* $\hat{x}, \hat{y}, \hat{\mathcal{N}}$*.*

Proof. See Appendix A. $\square$

Proposition 4 makes the Benders subproblem feasible for any feasible value of $\hat{x}, \hat{y}, \hat{\mathcal{N}}$. This is an important result to make the Benders decomposition implementation faster.

Let $\{\mu_{ik}\}$, $\{\lambda^1_{aik}\}$, $\{\lambda^2_{faik}\}$, $\{\lambda^3_{faik}\}$, $\{\lambda^4_{faik}\}$, $\{\lambda^5_{aik}\}$, $\{\lambda^6_{nik}\}$, $\{\lambda^7_{nik}\}$, and $\{\lambda^8_{nik}\}$ be the dual variables associated with the constraints (14b) -(14j) respectively. Then, the dual of the subproblem DSP can be stated as below:

$$\begin{aligned}
z^{DSP}(\hat{x},\hat{y},\hat{\mathcal{N}}) = \max_{\mu,\lambda} \quad & \sum_{k\in D}\Bigg[\sum_{i\in N}\mu_{ik}g_{ik} + \sum_{i\in N^w}\sum_{a\in FS(i):a\in A_T}\sum_{f\in\Theta}\big(\overline{W}_{ik}(1-\hat{y}_{l(a)f})\lambda^2_{faik} \\
& +\overline{W}_{ik}\hat{y}_{l(a)f}\lambda^3_{faik}\big)\big) + \sum_{i\in N^w\cap N_R}\sum_{n\in\Omega}\Big(\overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n})\lambda^6_{nik} \\
& +\overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}\lambda^7_{nik}\Big)\Bigg] && \text{(15a)} \\
\text{subject to} \quad & \mu_{ik}-\mu_{jk}+\lambda^1_{aik}+\lambda^5_{aik}\le c_a, \forall a=(i,j)\in A, \forall k\in D && \text{(15b)} \\
& -\sum_{f\in\theta}\sum_{\substack{a\in FS(i):\\ a\in A_T}}\big(\lambda^2_{faik}+\lambda^4_{faik}\big)+\sum_{n\in\Omega}\big(\lambda^6_{nik}+\lambda^8_{nik}\big)=1,\forall,\forall i\in N^w,\forall k\in D && \text{(15c)} \\
& -f\lambda^1_{aik}-\lambda^2_{faik}+\lambda^3_{faik}-\lambda^4_{faik}\le 0,\forall f\in\Theta,\forall a\in FS(i):a\in A_T,\forall i\in N^w,\forall k\in D && \text{(15d)} \\
& -n\lambda^5_{aik}-\lambda^6_{nik}+\lambda^7_{nik}-\lambda^8_{nik}\le 0,\forall n\in\Omega,\forall i\in N^w\cap N_R,\forall k\in D && \text{(15e)} \\
& \lambda^1_{aik},\lambda^5_{aik}\le 0,\forall a\in FS(i),\forall i\in N^w,\forall k\in D && \text{(15f)} \\
& \lambda^2_{faik},\lambda^3_{faik}\le 0,\lambda^4_{faik}\ge 0,\forall f\in\Theta,\forall a\in FS(i):a\in A_T,\forall i\in N^w,\forall k\in D && \text{(15g)} \\
& \lambda^6_{nik},\lambda^7_{nik}\le 0,\lambda^8_{nik}\ge 0,\forall n\in\Omega,\forall i\in N^w\cap N_R,\forall k\in D && \text{(15h)}
\end{aligned}$$

Let us denote the feasible region of DSP as $\Pi = \{(\mu,\lambda^1,\lambda^2,\lambda^3,\lambda^4,\lambda^5,\lambda^6,\lambda^7,\lambda^8) : (15b)-(15h)\}$. Note that Π does not depend on the value of $x, y, \mathcal{N}$. From Proposition 4, we know that SP is always feasible for any given feasible value of $(\hat{x},\hat{y},\hat{\mathcal{N}})$, then by linear programming duality, DSP should be bounded. The implication is that the polyhedron describing Π is bounded and can be described as the convex hull of a set of extreme points only (from Minkowski-Weyl's theorem on characterization of polyhedra (Conforti et al. 2014, Chapter 3)). Let $\{(\mu^\pi,(\lambda^1)^\pi,(\lambda^2)^\pi,(\lambda^3)^\pi,(\lambda^4)^\pi,(\lambda^5)^\pi,(\lambda^6)^\pi,(\lambda^7)^\pi,(\lambda^8)^\pi)\}_{\pi\in\mathcal{K}}$ be the set of extreme points of polytope Π, where $\mathcal{K}$ represents the set of indices of extreme points. By applying an outer linearization procedure to the inner (sub) problem of the original problem, we can restate it as (16), which is referred to as the *Benders Master problem* (MP).

Theorem 1. *(Benders 1962) The problem* (11) *can be reformulated as below:*

$$\underset{x,y,\mathcal{N},\eta}{\text{minimize}} \quad \eta \tag{16a}$$

$$\text{subject to} \quad \sum_{l\in L}\sum_{f\in\Theta}\mathcal{B}(l,f)\times y_{lf} \le \bar{B} \tag{16b}$$

$$\sum_{f\in\Theta} y_{lf} = x_l, \forall l \in L \tag{16c}$$

$$\sum_{n\in\Omega} \mathcal{N}_{zn} = 1, \forall z \in Z \tag{16d}$$

$$\sum_{z\in Z}\left(\sum_{n\in\Omega} n\mathcal{N}_{zn}\right) \le \bar{F} \tag{16e}$$

$$\begin{aligned}\eta \ge \sum_{k\in D}\Bigg[&\sum_{i\in N}(\mu_{ik})^\pi g_{ik} + \sum_{i\in N^w}\ \sum_{a\in FS(i):a\in A_T}\ \sum_{f\in\Theta}\big(\overline{W}_{ik}(1-\hat{y}_{l(a)f})(\lambda^2_{faik})^\pi \\ &+\overline{W}_{ik}\hat{y}_{l(a)f}(\lambda^3_{faik})^\pi)\big) + \sum_{i\in N^w\cap N_R}\ \sum_{n\in\Omega}\Big(\overline{W}_{ik}(1-\hat{\mathcal{N}}_{Z(i)n})(\lambda^6_{nik})^\pi \\ &+\overline{W}_{ik}\hat{\mathcal{N}}_{Z(i)n}(\lambda^7_{nik})^\pi\Big)\Bigg], \forall \pi \in \mathcal{K}\end{aligned} \tag{16f}$$

$$x_l \in \mathfrak{B}, \forall l \in L \tag{16g}$$

$$y_{lf} \in \mathfrak{B}, \forall f \in \Theta, \forall l \in L \tag{16h}$$

$$\mathcal{N}_{in} \in \mathfrak{B}, \forall n \in \Omega, i \in Z \tag{16i}$$

Proof. See Benders 1962. □

5.2. Classic Benders decomposition implementation

The issue with the Benders reformulation is that there could be a large number of extreme points of the polyhedron associated with the feasible region of DSP, therefore, one applies an iterative process of solving two problems, namely, the relaxed master problem (RMP) and the subproblem (SP) repeatedly. The relaxed master problem is the master problem with constraints (16f) being defined only for a subset of extreme points, i.e., $\mathcal{K}' \subset \mathcal{K}$. The overall implementation of the classic Benders Decomposition is summarized in Algorithm 1. We start by finding the feasible value of $(x^0, y^0, \mathcal{N}^0)$. This can be done by solving (16) without (16f) and including a constraint $\eta \ge 0$. Then, in each iteration t, the algorithm solves RMP with the given set of extreme points and then SP with the current value of $(x^t, y^t, \mathcal{N}^t)$. Since RMP is relaxation and SP is solved for a feasible value $(x^t, y^t, \mathcal{N}^t)$, they provide a lower bound and upper bound respectively to the original problem. The subproblem also provides inequalities (optimality cuts) to strengthen the formulation of RMP in each iteration. Thus, it is guaranteed to have non-decreasing lower bounds. In our case, there are no feasibility cuts since our subproblem is always feasible (Proposition 4). The algorithm terminates when both the upper bound and lower bound are close to each other.

Algorithm 1 Classic Benders decomposition implementation

1: (*Initialize*) Let $t = 0, UB = -\infty, LB = \infty, \mathcal{K}^{'} = \phi$. Assume an initial feasible value $(x^0, y^0, \mathcal{N}^0)$. Solve the SP (14), obtain the optimal dual solution and append that to set $\mathcal{K}^{'}$.
2: **while** $UB - LB > \epsilon$ **do** ▷ ϵ is the tolerance parameter
3: Set $t = t + 1$. Solve RMP (16) and obtain its optimal solution $(x^t, y^t, \mathcal{N}^t)$.
4: Set LB $= \eta$
5: Solve SP (14) for $(x^t, y^t, \mathcal{N}^t)$, obtain dual solutions and append that to $\mathcal{K}^{'}$.
6: Set $UB = \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak}^t + \sum_{i \in N^w} W_{ik}^t \right)$

5.3. Enhanced Benders decomposition implementation

The classic Benders decomposition may take prohibitive computational effort to converge, thus making it difficult to solve the problem for large instances. The slow convergence can be attributed to the low strength of the optimality cuts, degeneracy in the subproblem, no guarantee of non-decreasing upper bounds in each iteration, or not formulating the problem "properly" (Saharidis and Ierapetritou 2010, Tang et al. 2013, Magnanti and Wong 1981). To accelerate the Benders decomposition algorithm, we make use of several enhancements that are described below:

5.3.1. Use of multiple cuts via disaggregated cuts

For this design problem, we can further utilize the decomposable structure of the Benders subproblem (11) as it is decomposable for each destination $k \in D$. That is, we can solve several (smaller) subproblems and generate multiple optimality cuts for the master problem. The disaggregated cuts have a higher probability of finding facet-defining inequalities characterizing Π. For this purpose, we modify RMP to allow for the disaggregated cuts as (17):

$$
\begin{aligned}
&\underset{x,y,\mathcal{N},\eta}{\text{minimize}} \quad && \sum_{k \in D} \eta_k && \text{(17a)}\\
&\text{subject to} && \text{(16b)} - \text{(16e)} && \text{(17b)}\\
& && \eta_k \geq \Bigg[\sum_{i \in N} (\mu_{ik})^\pi g_{ik} + \sum_{i \in N^w} \sum_{a \in FS(i): a \in A_T} \sum_{f \in \Theta} \big(\overline{W}_{ik} (1 - \hat{y}_{l(a)f}) (\lambda_{faik}^2)^\pi && \\
& && + \overline{W}_{ik} \hat{y}_{l(a)f} (\lambda_{faik}^3)^\pi \big) \big) + \sum_{i \in N^w \cap N_R} \sum_{n \in \Omega} \Big(\overline{W}_{ik} (1 - \hat{\mathcal{N}}_{Z(i)n}) (\lambda_{nik}^6)^\pi && \\
& && + \overline{W}_{ik} \hat{\mathcal{N}}_{Z(i)n} (\lambda_{nik}^7)^\pi \Big) \Bigg], \forall \pi \in \mathcal{K}_k, \forall k \in D && \text{(17c)}\\
& && \text{(16g)} - \text{(16i)} && \text{(17d)}
\end{aligned}
$$

Note that by adding the disaggregated cuts for every destination, we can get back the optimality cuts defined in the classic Benders relaxed master problem.

5.3.2. Use of multiple cuts via multiple solutions

To further improve the convergence of the algorithm, Beheshti Asl and MirHassani 2019 used a strategy known as multiple cuts via multiple solutions. When solving the RMP, any commercial solver such as AIMMS or GUROBI can be asked to generate multiple solutions of an integer program (optimal as well as suboptimal) by using `pool solution option`. These multiple solutions can be used to generate multiple classic (16f) or disaggregated cuts (17c) to be added in next iteration of RMP. This strategy is expected to decrease the overall iterations and possibly the solution time of the algorithm.

5.3.3. Use of clique/cover cuts

Due to the limited availability of bus and vehicle fleet, one can use the clique/cover cuts to tighten the feasible region of the master problem.

Proposition 5. *For every $n \in \Omega$, if $\lfloor \frac{\bar{F}}{n} \rfloor < |Z|$ then the clique inequality $\sum_{z \in Z} \mathcal{N}_{zn} \leq \lfloor \frac{\bar{F}}{n} \rfloor$ is valid for* (11).

Proof. See Appendix A. □

If for any $n \in \Omega$, we have $\lfloor \frac{\bar{F}}{n} \rfloor > |Z|$, then the inequality $\sum_{z \in Z} \mathcal{N}_{zn} \leq \lfloor \frac{\bar{F}}{n} \rfloor$ will be redundant and therefore, we do not add it to the model.

The inequality which constrain the number of buses (11b) is a Knapsack constraint. A set $C \subseteq L \times \Theta$ is a *cover* for inequality (11b) if $\sum_{(l,f) \in C} \mathcal{B}(l, f) > \bar{B}$ and it is *minimal cover* if $\sum_{(l,f) \in C \setminus \{(l',f')\}} \mathcal{B}(l, f) \leq \bar{B}$, for all $(l', f') \in C$

Proposition 6. *For any minimal cover $C \subseteq L \times \Theta$, the inequality $\sum_{(l,f) \in C} y_{lf} \leq |C| - 1$ is valid for* (11).

Proof. See Appendix A. □

To generate some of the minimal cover cuts, one can use the heuristic given in Algorithm 2. In this algorithm, for each frequency $f \in \Theta$, we keep the list of lines G for which the number of buses required to provide the frequency f does not exceed $\bar{B}$. Then, any line which is not in G, along with G forms a minimal cover.

Algorithm 2 Cover cut generation heuristic

procedure
  Compute the value of mapping $\mathcal{B}(l, f)$ for all $(l, f) \in L \times \Theta$.
  $CC \leftarrow \phi$
  for $f \in \Theta$ **do**
    $G \leftarrow []; temp \leftarrow 0$
    for $l \in L : \mathcal{B}(l, f)$ in an ascending order **do**
      $temp = temp + \mathcal{B}(l, f)$
      if $temp \leq \bar{B}$ **then**
        append (l, f) to G
      else
        break
    for $l \in L \backslash G$ **do**
      $C \leftarrow G \cup \{(l, f)\}$
      append C to CC
return CC

Furthermore, one can use other efficient techniques to produce maximal clique or minimal cover cuts for the problem.

5.3.4. Other recommendations

One of the problems with the Benders subproblem (14) is that it assumes the value of $\overline{W}_{ik}$ as a given upper bound. The value of $\overline{W}_{ik}$ is a big-M introduced to relax the non-linearity in the original model. If the value of the big-M is not chosen properly, then one can face serious issues with the convergence of the algorithm. For example, choosing $\overline{W}_{ik} < W_{ik}$ can make the subproblem infeasible, and choosing $\overline{W}_{ik}$ too high would generate weak optimality cuts, which would increase the computational time of the algorithm. One way to avoid this issue is to solve the assignment problem (10) for given design variables $(x, y, \mathcal{N})$ and compute the optimal value of W_{ik} and use that as an upper bound. Further improvements in the Benders decomposition method can involve the use of *pareto-optimal* cuts proposed by Magnanti and Wong 1981. They help in avoiding the generation of multiple optimality cuts for a degenerate subproblem. We tried this strategy, however, we did not find any significant improvement in the solution time using these cuts, therefore, we do not discuss it here. Finally, when RMP is loaded with a large number of cuts we recommend removing the non-active cuts from the model by checking the slack value. There is no guarantee that they will not be generated again, but it will be faster to solve the RMP. The overall steps of the Benders implementation with possible acceleration techniques are summarized in Algorithm 3.

Algorithm 3 Enhanced Benders decomposition implementation

(*Initialize*) Let $t = 0, UB = -\infty, LB = \infty, \mathcal{K}_k^{'} = \phi, \forall k \in D$.
Prepare the master problem with clique and cover inequalities.
Assume an initial feasible value $(x^0, y^0, \mathcal{N}^0)$. Solve the SP (14), obtain the optimal dual solutions and append that to the set $\mathcal{K}_k^{'}, \forall k \in D$.
while $UB - LB > \epsilon$ **do** ▷ ϵ is the tolerance parameter
    Set $t = t + 1$. Solve RMP (16), obtain its optimal solution $(x_0^t, y_0^t, \mathcal{N}_0^t)$ and other optimal/-suboptimal solutions $\{(x_s^t, y_s^t, \mathcal{N}_s^t)\}_{1 \leq s \leq l}$, where l is specified by the user.
    Set LB $= \sum_{k \in D} \eta_k$
    for $s = 0, 1, ..., l$ **do**
        Solve SP (14) for $(x_s^t, y_s^t, \mathcal{N}_s^t)$, obtain dual solution and append that to $\mathcal{K}_k^{'}, \forall k \in D$.
    Set $UB = \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak}^t + \sum_{i \in N^w} W_{ik}^t \right)$

6. Computational results

In this section, we present the computational study based on the model (11), (16), and acceleration techniques presented in §5.3. We start by describing the details of the experiment used to show the application of the proposed method. Then, we present the details of the network design results in §6.2, which is followed by the comparison of the computational performance of the solver, the classic Benders implementation, and the enhanced Benders techniques described in §6.3. Finally, we discuss the results of the sensitivity analysis on two important parameters in the model, namely, the available fleet of buses $\bar{B}$ and vehicles $\bar{F}$ in §6.4, which is followed by the comparison of the performance of optimized existing transit system and proposed integrated system in §6.5.

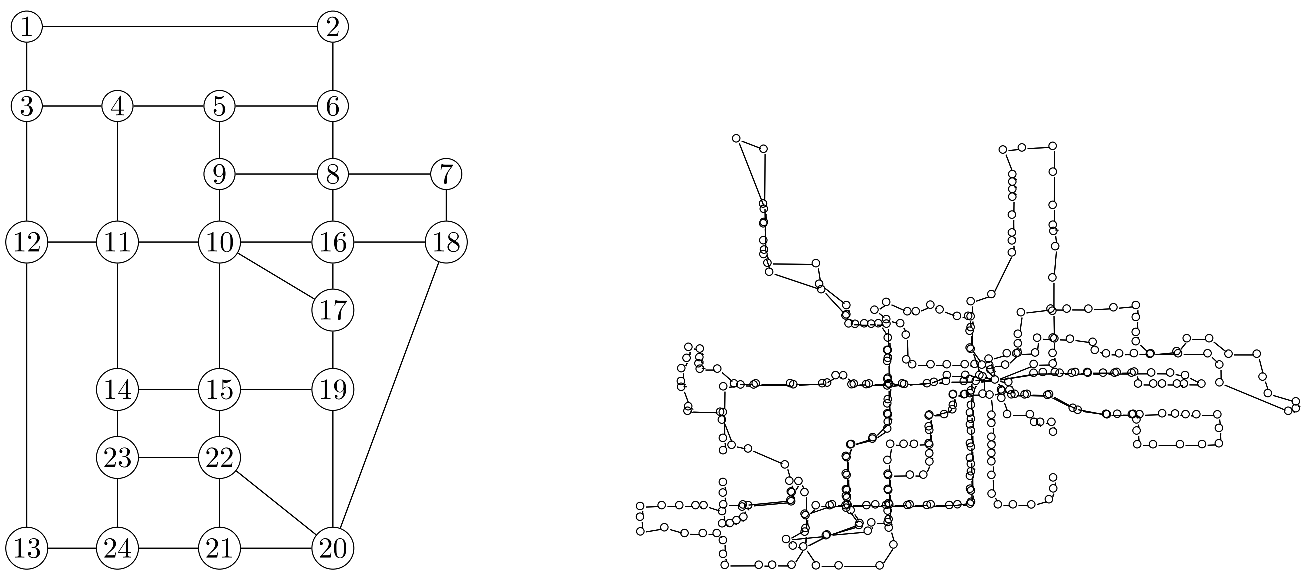
6.1. Experiment details

The computational experiments are based on the Sioux Falls road and transit network. The road network has 24 nodes, whereas the static transit network has 384 stops. A walking distance of 0.5 miles is used to create walking links. An illustration of the two networks is shown in Figure 2 and the number of different types of links in the network is given in Table 2. There are 12 candidate transit routes in the transit network. We consider the set of possible frequencies as $\Theta = \{2, 3, 4, 6, 12\}$ buses/hr to be assigned to any candidate transit route and possible vehicle fleet size to be assigned to any zone as $\Omega = \{0.01, 50, 100, 200, 500\}$. The vehicle fleet size value 0.01 is a dummy element to represent that no vehicles are assigned in a zone and that zone can be served by transit service only. A time-based fare of \$0.21/ min and a base fare of \$0.8 is assumed for the MoD service, whereas the transit fare is assumed to be a fixed value of \$2. To convert the monetary costs into time units, the value of travel time equal to 23\$/hr is used. The value of parameter $\mathcal{A}$ used in the wait time computations of MoD service is assumed to be equal to 0.0017 for all the zones (Yin 2019). The available number of buses $\bar{B}$ and vehicles $\bar{F}$ are assumed to be 70 and 3,000 respectively. There are 576 O-D pairs in the network with a total number of trips equal to 36,060. All implementations

Table 2: Number of different types of links in the Sioux Falls multimodal network

Link type	Number of links
Access (A^a)	243
Egress (A^e)	243
Road (A_R)	76
Transit (A_T)	398
Transit transfer (A_T^{tr})	368
Mode transfer (A^m)	152

are coded in Python 3.8 using Gurobi 9.0.1 as the optimization solver. The tests were executed on Intel(R) i7-7700 CPU running at 3.6 GHz with 32 GB RAM under a Windows operating system.



(a) Road network

(b) Transit network

Figure 2: Sioux Falls network

6.2. Network design results

We solve the network design problem (11) for the instance explained in §6.1. The selected transit routes with their optimal frequency are given in Table 3. Out of 12 candidate routes, 6 routes are decided to keep operating. The transit network with active and inactive routes is shown in Figure 3. We observe that most of the routes are located in the central region of the network. All the routes have been assigned the highest frequency i.e., 12 buses/hr, except route 8, which has been assigned a frequency of 3 buses/hr. To provide this service, 69 buses are required. The average number of vehicles deployed in each zone is given in Table 4. In the optimal allocation of vehicles, it is decided not to deploy any vehicles in 10 zones out of 24 zones. Most of the zones have been allocated 200 vehicles providing an average wait time of 3 minutes. We further observe that the vehicles are deployed in the outskirt zones of the network where transit routes are not located.

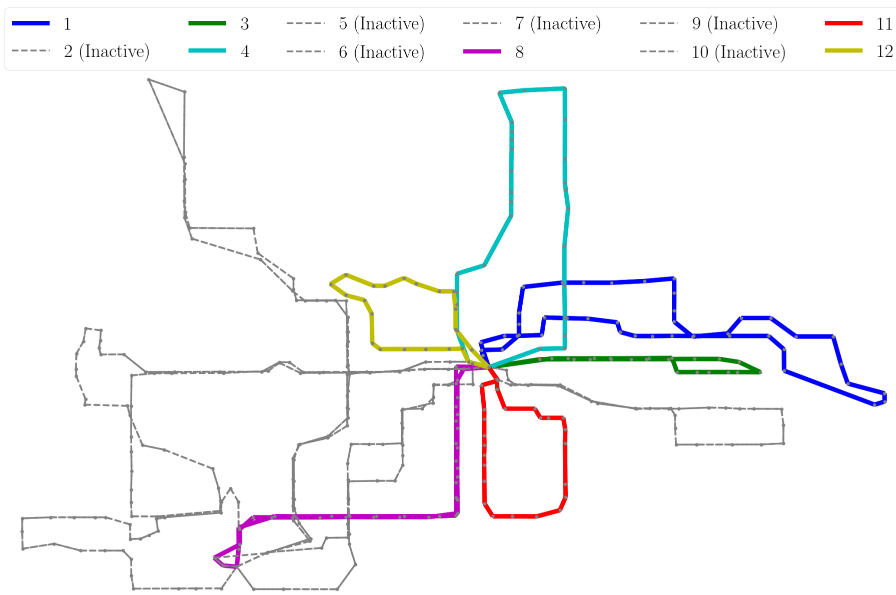


Figure 3: Transit routes (inactive routes are shown by dashed gray color)

Table 3: Selected transit routes with their optimal frequency

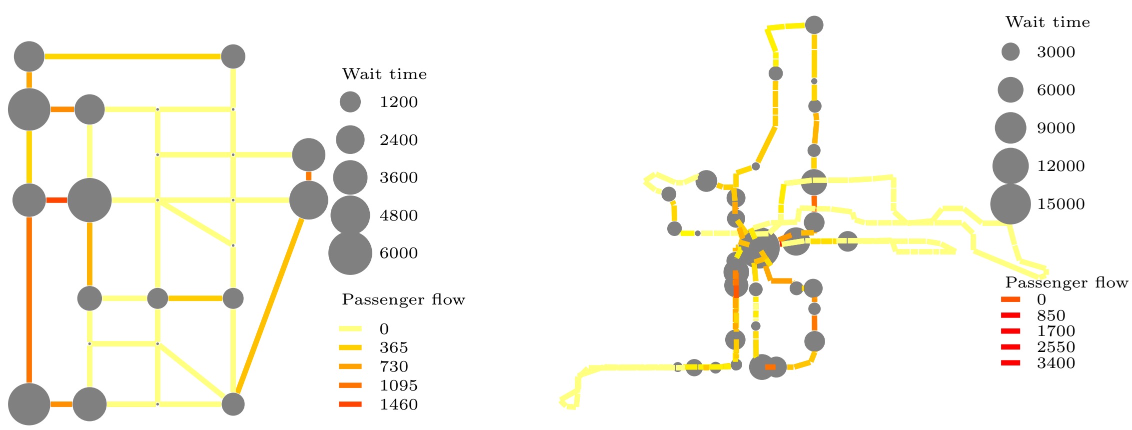
Route	**Located?**	**Optimal frequency** (buses/hr)	**Average wait time** (min)
1	Yes	12	5
2	No	-	-
3	Yes	12	5
4	Yes	12	5
5	No	-	-
6	No	-	-
7	No	-	-
8	Yes	3	20
9	No	-	-
10	No	-	-
11	Yes	12	5
12	Yes	12	5

The total time spent in the system is equal to 11,301 passenger-hours including 8,673 passenger-hours of travel time on various links, 1,881 passenger-hours of wait time spent on the transit network, and 747 passenger-hours of wait time spent on the road network. We found that more passengers

Table 4: vehicle allocation to different zones

Zone	Vehicles	Avg. Wait time (min)	Zone	Vehicles	Avg. Wait time (min)
1	200	3	13	200	3
2	100	6	14	200	3
3	200	3	15	200	3
4	200	3	16	-	-
5	-	-	17	-	-
6	-	-	18	200	3
7	200	3	19	200	3
8	-	-	20	200	3
9	-	-	21	-	-
10	-	-	22	-	-
11	200	3	23	-	-
12	500	1.2	24	200	3

take transit than MoD service. The share of passengers using the road, transit, and multimodal service are 23 %, 61 %, and 16 % respectively. The passenger flow on various links and wait time on various nodes of the road and transit networks (resp.) are visualized in Figure 4(a) and (b) respectively. We observe that most of the passenger trips in the central zones are made using transit network, whereas the trips on the outskirts of the network are made using both MoD and multimodal service. The figures further show that the congestion in the central zones is significantly improved with the resulting network design.



(a) Flow and wait time of passengers in the road network

(b) Flow and wait time of passengers in the transit network

Figure 4: Flow and wait time (pass-min) of passengers in the network

6.3. Computational performance

In this section, we compare the computational performance of various models and implementation techniques. We consider the following approaches to compare:

1. Solving model (11) using Gurobi solver

Table 5: Computational performance

Method	Iterations	Computational time (s)	Gap (%)
Gurobi bilinear	-	Timed out*	13.3
Gurobi	-	Timed out*	0.62
Classic	734	Timed out*	0.16
Classic + Clique/Cover	510	7,440	0
Classic + Multiple	500	6,524	0
Classic + Clique/Cover + Multiple	423	5,566	0
Disaggregate	31	381	0
Disaggregate + Clique/Cover	30	347	0
Disaggregate + Multiple	28	535	0
Disaggregate + Multiple + Clique/Cover	27	498	0

*Note: Maximum time limit = 3 hours

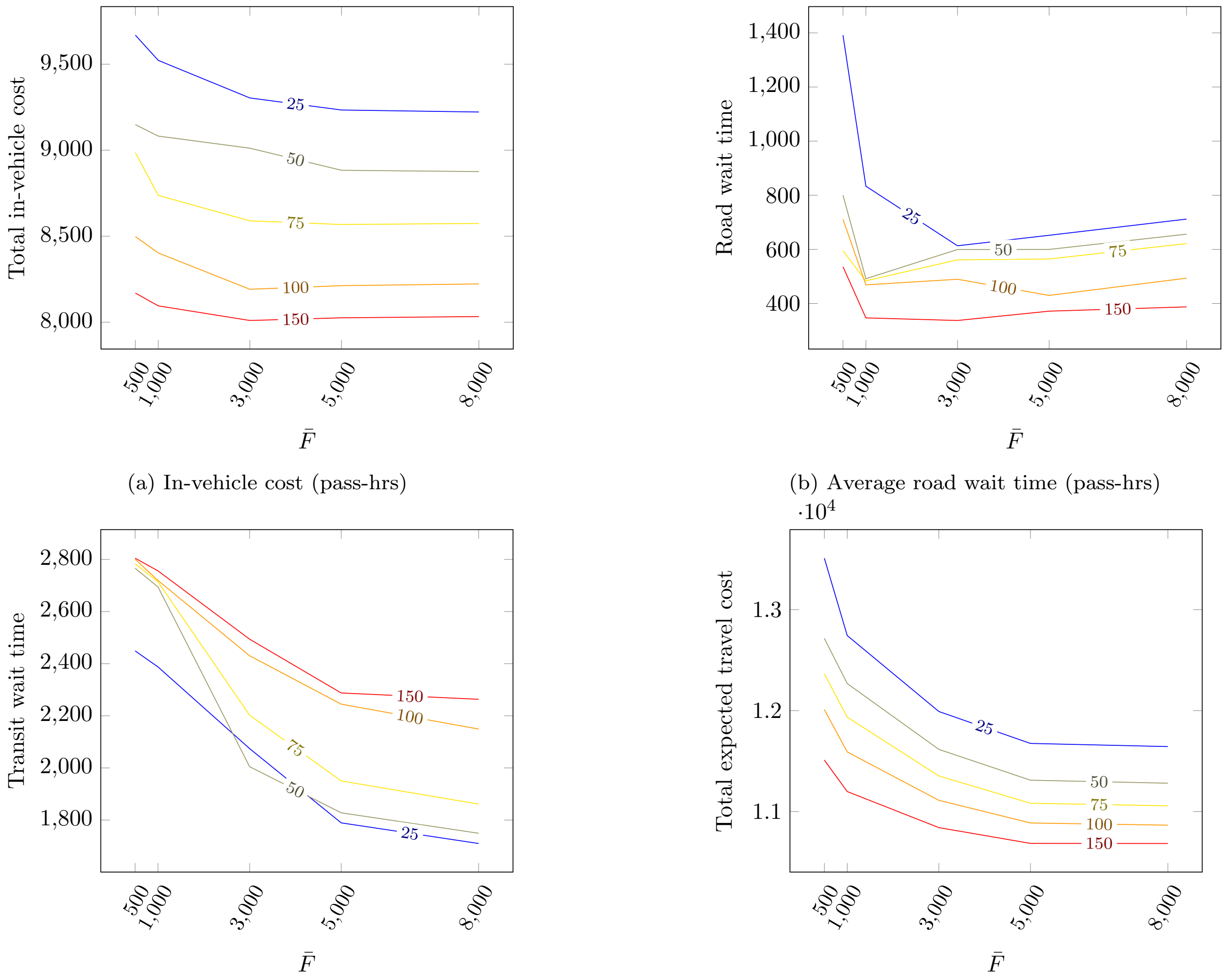
2. Solving model (16) using Gurobi solver
3. Solving model (16) using classic Benders decomposition (Algorithm 1)
4. Solving model (16) using Benders decomposition with clique/cover cuts (§5.3.3)
5. Solving model (16) using Benders decomposition with multiple cuts via multiple solutions (§5.3.2)
6. Solving model (16) using Benders decomposition with both clique/cover and multiple cuts via multiple solutions
7. Solving model (16) using Benders decomposition with disaggregated cuts (§5.3.1)
8. Solving model (16) using Benders decomposition with disaggregated and clique/cover cuts
9. Solving model (16) using Benders decomposition with disaggregated and multiple cuts via multiple solutions
10. Solving model (16) using Benders decomposition with disaggregated, clique/cover, and multiple cuts via multiple solutions

To solve the bilinear model (11), we set the Gurobi parameter `NonConvex = 2`. For Benders decomposition with multiple cuts via multiple solutions, we set the Gurobi parameters `PoolSolutions = 2`, `PoolGap = 0.01`, `PoolSearchMode = 2`. For all above tests, the maximum time limit was set to 3 hours.

The computational performance of every method is shown in Table 5. The iterations are counted as the number of times RMP is solved, the computational time is recorded in seconds, and Gap is defined as $(UB - LB) * 100/UB$. The bilinear model (11) is hard to solve, and Gurobi took 3 hours to reach the optimality gap of 13.3 %. The rest of the results are discussed for the optimization model (16). Other than Gurobi and classic Benders decomposition, all the methods coverage to the optimal solution. Gurobi and classic Benders decomposition reached an optimality gap of 0.62% and 0.16% respectively. This means both methods reached very close to the optimal solution in 3 hours. The hybrid approach of classic Benders decomposition with both clique/cover cuts and multiple cuts via multiple solutions outperforms the classic Benders decomposition with clique/-cover cuts or multiple cuts via multiple solutions only. The disaggregated Benders decomposition is computationally more efficient than any classic Benders approach with cut improvements. The disaggregated cuts with other cuts show further improvement in the solution time and the number of iterations to converge to the optimal solution. The Benders decomposition using disaggregated, clique/cover, and multiple cuts via multiple solutions outperforms other methods in terms of the number of iterations to converge to an optimal solution, whereas Benders decomposition with disaggregated and clique/cover cuts outperform other methods in terms of computational time. This may be because the multiple cuts are generated by solving several subproblems, which takes more computational time, but the generated cuts may not be as effective. Overall, the experiments performed for this section show that the computational methods presented in this study are quite efficient in solving the current problem exactly.

6.4. Sensitivity analysis on parameters

The availability of buses and vehicles can result in different network design results. Hence, we choose to perform a sensitivity analysis on the available bus fleet $\bar{B}$ and vehicle fleet $\bar{F}$. We solve the model (16) with varying bus fleet size of 25, 50, 75, 100, and 150 and varying vehicle fleet size of 500, 1000, 3000, 5000, 8000. Figure 5 and 6 show the sensitivity analysis results based on contour plots. The x-axis shows the varying vehicle fleet sizes, and contours represent varying bus fleet sizes. Figure 5(a), (b), (c), and (d) show the in-vehicle cost, average road wait time, average transit wait time, and total expected travel cost in passenger-hours respectively. We can observe that the in-vehicle cost decreases with the increase in the number of available vehicles. The effect of increasing the number of buses is more than the increase in the number of vehicles. Moreover, the in-vehicle cost is not affected by increasing the number of vehicles to more than 5,000. The average road wait time decreases with the increase in the number of available vehicles. It also decreases with the increase in the number of available buses due to mode shift. The passenger-hours spent as the wait time in the transit network increases with the increase in the number of available vehicles as well as buses. This is because more passengers take the transit mode as more buses are made available. The overall expected travel cost also reduces with the increase in the bus and vehicle fleet. However, the effect of an increase in the vehicle fleet size of more than 5,000 is negligible. Figure 6 shows the mode share as a function of the available bus and vehicle fleet size. As expected, the transit and MoD share increase with the increase in the available bus and vehicle fleet size respectively. The share of multimodal service increases with the number of vehicles and buses up to 5,000 and 75 respectively but declines after that. The decline in multimodal share is because of the reduced wait time for both services, which drives passengers to use single mode rather than multiple modes.



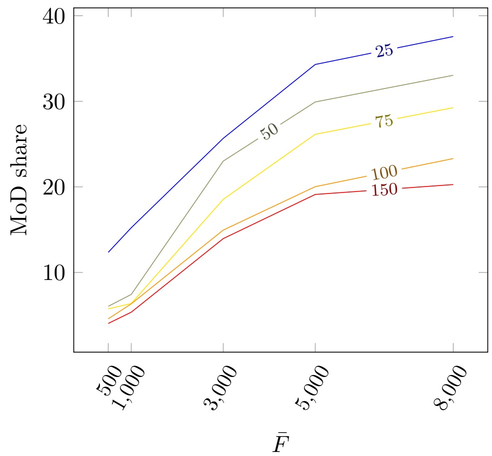
(a) In-vehicle cost (pass-hrs)

(b) Average road wait time (pass-hrs)

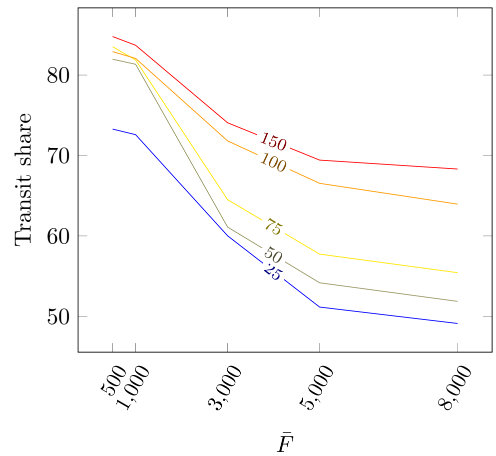
(c) Average transit wait time (pass-hrs)

(d) Total expected travel cost (pass-hrs)

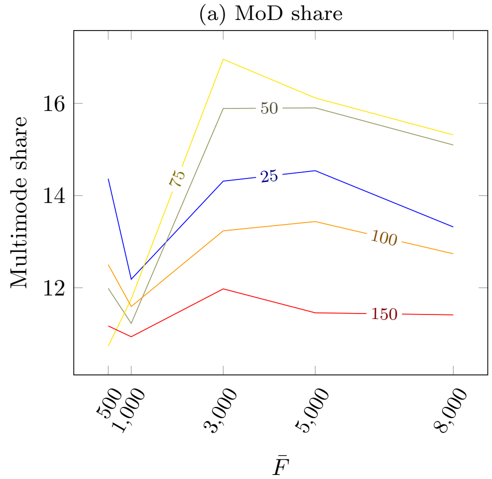
Figure 5: Sensitivity of parameters $\bar{F}$ and $\bar{B}$ on different costs (contour represents varying bus fleet sizes)



(a) MoD share



(b) Transit share



(c) Multimode share

Figure 6: Sensitivity of parameters $\bar{F}$ and $\bar{B}$ on mode share (contour represents varying bus fleet sizes)

6.5. Comparison of optimized base transit system with proposed integrated system

In this section, we present a comparison of the operation of the "optimized base transit system" corresponding to the existing transit system with optimized frequencies versus the design of the integrated system evaluated in §6.2. For the optimized base case, we solve the optimization program (18) for the instance described in §6.1. The results of optimized frequencies of various routes are given in Table 6. The network provides an average wait time of 18 minutes to the passengers.

$$
\begin{aligned}
\underset{v,W,y}{\text{minimize}} \quad & \sum_{k \in D} \left(\sum_{a \in A} c_a v_{ak} + \sum_{i \in N_T^w} W_{ik} \right) && \text{(18a)} \\
\text{subject to} \quad & \sum_{l \in L} \sum_{f \in \Theta} \mathcal{B}(l,f) \times y_{lf} \leq \bar{B} && \text{(18b)} \\
& \sum_{a \in FS(i)} v_{ak} = \sum_{a \in BS(i)} v_{ak} + g_{ik}, \forall i \in N, \forall k \in D && \text{(18c)} \\
& v_{ak} \leq \left(\sum_{f \in \Theta} f y_{l(a)f} \right) W_{ik}, \forall a \in FS(i) : a \in A_T, \forall i \in N^w, \forall k \in D && \text{(18d)} \\
& v_{ak} \geq 0, \forall a \in A, \forall k \in D && \text{(18e)} \\
& W_{ik} \text{ free }, \forall i \in N_T^w, \forall k \in D && \text{(18f)} \\
& y_{lf} \in \mathfrak{B}, \forall f \in \Theta, \forall l \in L && \text{(18g)}
\end{aligned}
$$

Table 6: Routes with their optimal frequency (optimized base case)

Route	**Optimal frequency** (buses/hr)	**Average wait time** (min)
1	4	15
2	2	30
3	6	10
4	12	5
5	3	20
6	2	30
7	3	20
8	3	20
9	3	20
10	2	30
11	12	5
12	6	10

The results comparing the performance of the optimized base transit system and integrated system are provided in Table 7. The base transit system has 12 two-way bus services operated by a bus fleet of 69 buses, whereas the new integrated system has 6 two-way bus services operated by 69 buses. Along with 69 buses, the new integrated system deploys 3,000 vehicles to serve the demand. The deployment of these extra vehicles can be costly to the transportation agencies. However, they provide several benefits. First, the optimized base transit system is not able to serve 13 % of the demand due to the non-availability of transit service in 2 zones in the network. On the other hand, the first mile and last mile of these zones are covered by vehicles in the integrated system. Second, the average in-vehicle travel time of passengers using the integrated system is only 14.43 minutes in comparison to the 21.16 minutes for passengers using the base transit system. However, the average wait time of the integrated system users is increased slightly in comparison to the base transit system. This is due to the increased number of transfers to access MoD and transit service.

Table 7: Comparison of optimized base transit system and integrated system

	Optimized base transit system	**Integrated system**
Number of active routes	12	6
Number of buses used	69	69
Number of vehicles used	0	3,000
Satisfied demand (%)	87	100
Average in-vehicle time (min/passenger)	21.16	14.43
Average wait time (min/passenger)	2.88	4.37

6.6. Managerial insights for implementing such service

For implementing the proposed model in practice, we need to follow the following procedure. First, we divide the region into zones. Second, we collect the peak hour demand data in the region. Third, solve the proposed design model for varying fleet sizes. This step will be similar to the sensitivity analysis given in Section 6.4. This analysis will help us decide the optimal fleet size of buses and vehicles for our service. It will also provide us with the allocation of vehicles and buses for different zones and bus routes respectively. This allocation is designed for peak hours. We can reduce the vehicle operation in non-peak hours. For the bus service, scheduling needs to be performed to publish a schedule for the service.

7. Conclusions and Future Research

Advances in mobility services have paved the way for the development of a new type of MoD service, which can help in serving the first mile and last mile of transit trips. Such a system requires rethinking the design of a transit system that allows for intermodal trips with MoD as the first or last leg of trips. We developed a mixed-integer non-linear program (MINLP) to design such system. The MINLP was relaxed to a mixed-integer linear program with the help of McCormick relaxations. To solve the resulting MILP model efficiently, we proposed the Benders decomposition method with several enhancements. These enhancements include the use of disaggregated cuts, clique/cover cuts, and multiple cuts via multiple solutions. The numerical results show that disaggregated cuts with clique/cover cuts and multiple cuts via multiple solutions are efficient techniques to solve the current problem. Furthermore, the experiment results on the Sioux Falls network show that the congestion in the city center is improved with such design as most of the passengers were found to take the transit in that region. The sensitivity analysis on bus fleet size and vehicle fleet size shows that the passenger hours spent in the system as in-vehicle time and wait time reduces with an increase in the number of available buses and vehicles. The share of multimodal service was observed to be highest for the vehicle fleet size and bus fleet size of 3,000 and 75 respectively. We also compared the proposed integrated system with the optimized base transit system. We found that the integrated system can be costly due to the deployment of vehicles, but it reduces the passenger in-vehicle time and serves more demand than the optimized base case.

This research can be expanded in multiple directions. First, ridepooling was not allowed in the current study. Further research is needed to explore the ideas of including the matching of passengers for ridepooling, which will further reduce the size of the vehicle fleet required to provide the service. Second, better calibration of parameter $\mathcal{A}$ used in the wait time computation of MoD service is needed. The data from ridehailing services can be used for this purpose.

Acknowledgments

This research is conducted at the University of Minnesota Transit Lab, currently supported by the following, but not limited to, projects:

- National Science Foundation, award CMMI-1831140
- Freight Mobility Research Institute (FMRI), Tier 1 Transportation Center, U.S. Department of Transportation: award RR-K78/FAU SP#16-532 AM2 and AM3
- Minnesota Department of Transportation, Contract No. 1003325 Work Order No. 44 and 111
- University of Minnesota Office of Vice President for Research, COVID-19 Rapid Response Grants

References

Aftabuzzaman, M., Currie, G. and Sarvi, M. 2015, ‘Evaluating the Congestion Relief Impacts of Public Transport in Monetary Terms’, *Journal of Public Transportation* **13**(1), 1–24.

Baaj, M. H. and Mahmassani, H. S. 1991, ‘An AI-based approach for transit route system planning and design’, *Journal of advanced transportation* **25**(2), 187–209.

Basu, R., Araldo, A., Akkinepally, A. P., Nahmias Biran, B. H., Basak, K., Seshadri, R., Deshmukh, N., Kumar, N., Azevedo, C. L. and Ben-Akiva, M. 2018, ‘Automated Mobility-on-Demand vs. Mass Transit: A Multi-Modal Activity-Driven Agent-Based Simulation Approach’, *Transportation Research Record* **2672**(8), 608–618.
URL: *https://doi.org/10.1177/0361198118758630*

Beheshti Asl, N. and MirHassani, S. A. 2019, ‘Accelerating benders decomposition: multiple cuts via multiple solutions’, *Journal of Combinatorial Optimization* **37**(3), 806–826.
URL: *https://doi.org/10.1007/s10878-018-0320-8*

Benders, J. F. 1962, ‘Partitioning procedures for solving mixed-variables programming problems’, *Numerische Mathematik* **4**(1), 238–252.

Bian, Z. and Liu, X. 2019, ‘Mechanism design for first-mile ridesharing based on personalized requirements part I: Theoretical analysis in generalized scenarios’, *Transportation Research Part B: Methodological* **120**, 147–171.
URL: *https://doi.org/10.1016/j.trb.2018.12.009*

Campbell, J. F., Ernst, A. T. and Krishnamoorthy, M. 2005*a*, ‘Hub arc location problems: Part I - Introduction and results’, *Management Science* **51**(10), 1540–1555.

Campbell, J. F., Ernst, A. T. and Krishnamoorthy, M. 2005*b*, ‘Hub Arc location problems: Part II - Formulations and optimal algorithms’, *Management Science* **51**(10), 1556–1571.

Cayford, R. and Yim, Y. B. Y. 2004, ‘Personalized Demand-Responsive Transit Service’.

Ceder, A. and Wilson, N. H. M. 1986, ‘Bus network design’, *Transportation Research Part B: Methodological* **20**(4), 331–344.
URL: *http://www.sciencedirect.com/science/article/pii/0191261586900470*

Cepeda, M., Cominetti, R. and Florian, M. 2006, ‘A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria’, *Transportation research part B: Methodological* **40**(6), 437–459.

Chen, S., Wang, H. and Meng, Q. 2020, ‘Solving the first-mile ridesharing problem using autonomous vehicles’, *Computer-Aided Civil and Infrastructure Engineering* **35**(1), 45–60.

Chen, T. D. and Kockelman, K. M. 2016, 'Management of a shared autonomous electric vehicle fleet: Implications of pricing schemes', *Transportation Research Record* **2572**(2572), 37–46.

Cominetti, R. and Correa, J. 2001, 'Common-lines and passenger assignment in congested transit networks', *Transportation Science* **35**(3), 250–267.

Conforti, M., Cornuejols, G. and Zambelli, G. 2014, *Integer Programming*, Graduate Texts in Mathematics, Springer International Publishing.
URL: *https://books.google.com/books?id=KUHUoAEACAAJ*

De Cea, J. and Fernández, E. 1993, 'Transit assignment for congested public transport systems: an equilibrium model', *Transportation science* **27**(2), 133–147.

Desaulniers, G. and Hickman, M. D. 2007, 'Chapter 2 Public Transit', *Handbooks in Operations Research and Management Science* **14**(C), 69–127.

Douglas, G. W. 1972, 'Price Regulation and Optimal Service Standards: The Taxicab Industry', *Journal of Transport Economics and Policy* **6**(2), 116–127.
URL: *http://www.jstor.org/stable/20052271*

Fagnant, D. J., Kockelman, K. M. and Bansal, P. 2016, 'Operations of Shared Autonomous Vehicle Fleet for Austin, Texas, Market', *Transportation Research Record* **2563**(1), 98–106.
URL: *https://doi.org/10.3141/2536-12*

Gentile, G., Nguyen, S. and Pallottino, S. 2005, 'Route choice on transit networks with online information at stops', *Transportation science* **39**(3), 289–297.

Geoffrion, A. M. 1972, 'Generalized Benders decomposition', *Journal of Optimization Theory and Applications* **10**(4), 237–260.
URL: *https://doi.org/10.1007/BF00934810*

Guihaire, V. and Hao, J. K. 2008, 'Transit network design and scheduling: A global review', *Transportation Research Part A: Policy and Practice* **42**(10), 1251–1273.
URL: *http://dx.doi.org/10.1016/j.tra.2008.03.011*

Gurumurthy, K. M., Kockelman, K. M. and Zuniga-Garcia, N. 2020, 'First-Mile-Last-Mile Collector-Distributor System using Shared Autonomous Mobility', *Transportation Research Record* **0**(0), 0361198120936267.
URL: *https://doi.org/10.1177/0361198120936267*

Khani, A., Lee, S., Hickman, M., Noh, H. and Nassir, N. 2012, 'Intermodal Path Algorithm for Time-Dependent Auto Network and Scheduled Transit Service', *Transportation Research Record: Journal of the Transportation Research Board* **2284**(2284), 40–46.
URL: *http://trrjournalonline.trb.org/doi/10.3141/2284-05*

Koffman, D. 2004, *Operational Experiences with Flexible Transit Services*, National Academies Press.

Kumar, P. and Khani, A. 2021, 'An algorithm for integrating peer-to-peer ridesharing and schedule-based transit system for first mile/last mile access', *Transportation Research Part C: Emerging Technologies* **122**, 102891.
URL: *https://www.sciencedirect.com/science/article/pii/S0968090X20307919*

Kurauchi, F., Bell, M. G. and Schmöcker, J. D. 2003, 'Capacity Constrained Transit Assignment with Common Lines', *Journal of Mathematical Modelling and Algorithms* **2**(4), 309–327.

Laris, M. 2019, 'Uber and Lyft concede they play role in traffic congestion in the District and other urban areas'.
URL: *https://www.washingtonpost.com/transportation/2019/08/06/uber-lyft-concede-they-play-role-traffic-congestion-district-other-urban-areas/*

Larson, R. C. and Odoni, A. R. 1981, *Urban operations research*, number Monograph, 2nd editio edn, Dynamic Ideas.

Lee, A. and Savelsbergh, M. 2017, 'An extended demand responsive connector', *EURO Journal on Transportation and Logistics* **6**(1), 25–50.
URL: *http://dx.doi.org/10.1007/s13676-014-0060-6*

Leurent, F., Chandakas, E. and Poulhès, A. 2014, 'A traffic assignment model for passenger transit on a capacitated network: Bi-layer framework, line sub-models and large-scale application', *Transportation Research Part C: Emerging Technologies* **47**(P1), 3–27.
URL: *http://dx.doi.org/10.1016/j.trc.2014.07.004*

Levin, M. W. and Boyles, S. D. 2015, 'Effects of autonomous vehicle ownership on trip, mode, and route choice', *Transportation Research Record* **2493**, 29–38.

Levin, M. W., Kockelman, K. M., Boyles, S. D. and Li, T. 2017, 'A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application', *Computers, Environment and Urban Systems* **64**, 373–383.
URL: *http://dx.doi.org/10.1016/j.compenvurbsys.2017.04.006*

Li, X. and Quadrifoglio, L. 2009, 'Optimal Zone Design for Feeder Transit Services', *Transportation Research Record: Journal of the Transportation Research Board* **2111**(1), 100–108.

Liu, Y., Bansal, P., Daziano, R. and Samaranayake, S. 2019, 'A framework to integrate mode choice in the design of mobility-on-demand systems', *Transportation Research Part C: Emerging Technologies* **105**, 648–665.
URL: *http://www.sciencedirect.com/science/article/pii/S0968090X18313718*

Ma, T. Y., Rasulkhani, S., Chow, J. Y. and Klein, S. 2019, 'A dynamic ridesharing dispatch and idle vehicle repositioning strategy with integrated transit transfers', *Transportation Research Part E: Logistics and Transportation Review* **128**(July), 417–442.
URL: *https://doi.org/10.1016/j.tre.2019.07.002*

Magnanti, T. L. and Wong, R. T. 1981, 'Accelerating Benders Decomposition: Algorithmic Enhancement and Model Selection Criteria', *Operations Research* **29**(3), 464–484.
URL: *https://doi.org/10.1287/opre.29.3.464*

Maheo, A., Kilby, P. and Van Hentenryck, P. 2017, 'Benders decomposition for the design of a hub and shuttle public transit system', *Transportation Science* **53**(1), 77–88.

Mahéo, A., Kilby, P. and Van Hentenryck, P. 2019, 'Benders decomposition for the design of a hub and shuttle public transit system', *Transportation Science* **53**(1), 77–88.

Manser, P. 2017, Public Transport Network Design in a World of Autonomous Vehicles, PhD thesis, Master thesis.

Masoud, N., Nam, D., Yu, J. and Jayakrishnan, R. 2017, 'Promoting Peer-to-Peer Ridesharing Services as Transit System Feeders', *Transportation Research Record: Journal of the Transportation Research Board* **2650**, 74–83.
URL: *http://trrjournalonline.trb.org/doi/10.3141/2650-09*

Mendes, L. M., Bennàssar, M. R. and Chow, J. Y. 2017, 'Comparison of light rail streetcar against shared autonomous vehicle fleet for Brooklyn–Queens connector in New York City', *Transportation Research Record* **2650**(1), 142–151.
URL: *https://doi.org/10.3141/2650-17*

Mo, B., Cao, Z., Zhang, H., Shen, Y. and Zhao, J. 2020, 'Dynamic Interaction between Shared Autonomous Vehicles and Public Transit: A Competitive Perspective', **100084**.
URL: *http://arxiv.org/abs/2001.03197*

Motavalli, J. 2020, 'Who Will Own the Cars That Drive Themselves?'.
URL: *https://www.nytimes.com/2020/05/29/business/ownership-autonomous-cars-coronavirus.html*

Nassir, N., Khani, A., Hickman, M. and Noh, H. 2012, 'Algorithm for Intermodal Optimal Multi-destination Tour with Dynamic Travel Times', *Transportation Research Record: Journal of the Transportation Research Board* **2283**, 57–66.
URL: *http://trrjournalonline.trb.org/doi/10.3141/2283-06*

OECD 2015, 'Urban Mobility System Upgrade: How shared self-driving cars could change city traffic', *Corporate Partnership Board Report* pp. 1–36.
URL: *http://www.internationaltransportforum.org/Pub/pdf/15CPB_Self-drivingcars.pdf*

Pinto, H. K., Hyland, M. F., Mahmassani, H. S. and Verbas, I. Ö. 2020, ‘Joint design of multimodal transit networks and shared autonomous mobility fleets’, *Transportation Research Part C: Emerging Technologies* **113**(June), 2–20.

Quadrifoglio, L., Dessouky, M. M. and Ordóñez, F. 2008, ‘A simulation study of demand responsive transit system design’, *Transportation Research Part A: Policy and Practice* **42**(4), 718–737.

Saharidis, G. K. and Ierapetritou, M. G. 2010, ‘Improving benders decomposition using maximum feasible subsystem (MFS) cut generation strategy’, *Computers and Chemical Engineering* **34**(8), 1237–1245.

Salazar, M., Rossi, F., Schiffer, M., Onder, C. H. and Pavone, M. 2018, On the Interaction between Autonomous Mobility-on-Demand and Public Transportation Systems, *in* ‘2018 21st International Conference on Intelligent Transportation Systems (ITSC)’, pp. 2262–2269.

Shen, C.-W. and Quadrifoglio, L. 2012, ‘Evaluation of Zoning Design with Transfers for Paratransit Services’, *Transportation Research Record: Journal of the Transportation Research Board* **2277**(1), 82–89.

Shen, Y., Zhang, H. and Zhao, J. 2018, ‘Integrating shared autonomous vehicle in public transportation system: A supply-side simulation of the first-mile service in Singapore’, *Transportation Research Part A: Policy and Practice* **113**(March), 125–136.

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

Steiner, K. and Irnich, S. 2020, ‘Strategic Planning for Integrated Mobility-on-Demand and Urban Public Bus Networks’, *Transportation Science* **54**(6), 1616–1639.
URL: *https://doi.org/10.1287/trsc.2020.0987*

Stiglic, M., Agatz, N., Savelsbergh, M. and Gradisar, M. 2018, ‘Enhancing urban mobility: Integrating ride-sharing and public transit’, *Computers and Operations Research* **90**, 12–21.
URL: *http://dx.doi.org/10.1016/j.cor.2017.08.016*

Szeto, W. Y. and Jiang, Y. 2014, ‘Transit route and frequency design: Bi-level modeling and hybrid artificial bee colony algorithm approach’, *Transportation Research Part B: Methodological* **67**, 235–263.
URL: *http://dx.doi.org/10.1016/j.trb.2014.05.008*

Tang, L., Jiang, W. and Saharidis, G. K. 2013, ‘An improved Benders decomposition algorithm for the logistics facility location problem with capacity expansions’, *Annals of Operations Research* **210**(1), 165–190.

Vakayil, A., Gruel, W. and Samaranayake, S. 2017, Integrating shared-vehicle mobility-on-demand systems with public transit, Technical report, Transportation Research Board, Washington, D.C.

Wang, H. 2017, 'Routing and Scheduling for a Last-Mile Transportation System', *Transportation Science* **53**(December 2018), trsc.2017.0753.
URL: *http://pubsonline.informs.org/doi/10.1287/trsc.2017.0753*

Webb, A. and Khani, A. 2020, 'Park-and-Ride Choice Behavior in a Multimodal Network with Overlapping Routes', *Transportation Research Record* **2674**(3), 150–160.
URL: *https://doi.org/10.1177/0361198120908866*

Wen, J., Chen, Y. X., Nassir, N. and Zhao, J. 2018, 'Transit-oriented autonomous vehicle operation with integrated demand-supply interaction', *Transportation Research Part C: Emerging Technologies* **97**(January), 216–234.
URL: *https://doi.org/10.1016/j.trc.2018.10.018*

Wilson, W. H. 1972, 'Statewide Intermodal Transportation Planning in the Less Urbanized State', *Highway Research Record* (401).

Yin, Y. 2019, 'Macroscopic modeling of ridesourcing systems - Regulations and Fundamental Diagram'.
URL: *https://www.youtube.com/watch?v=oiBhwJl5xXc&t=717s*

Zha, L., Yin, Y. and Yang, H. 2016, 'Economic analysis of ride-sourcing markets', *Transportation Research Part C: Emerging Technologies* **71**, 249–266.
URL: *http://dx.doi.org/10.1016/j.trc.2016.07.010*

Appendix A Proofs of various propositions

Proof of Proposition 1

(Although the proof can be found in Spiess and Florian 1989 or Gentile et al. 2005, we repeat it here because some of the details presented here will be used in proving the next proposition.) The probability of choosing line $i \in FS(n)$ is equal to the probability of waiting time for line $i \in FS(n)$ to be less than or equal to waiting time of other lines $j \neq i$, i.e.,

$$P_i = Prob(w_i \leq \min_{j \neq i} w_j) = \int_0^\infty \mathfrak{g}_i(w) \Pi_{j \neq i} Prob(w_j \geq w) dw = \int_0^\infty \gamma_i(w) dw \tag{19}$$

where, $\gamma_i(w) = g_i(w) \Pi_{j \neq i} Prob(w_j \geq w) = f_i e^{-f_i w} \Pi_{j \neq i} e^{-f_j w} = f_i e^{-(\sum_j f_j) w}$. The value of $\gamma_i(w)$ can be interpreted as the probability density function of the waiting time at the stop n conditional to boarding line i. Using (19), the probability of choosing line $i \in FS(n)$ can be evaluated as:

$$P_i = \int_0^\infty f_i e^{-(\sum_j f_j) w} dw = \frac{f_i}{\mathfrak{F}} \tag{20}$$

The expected wait time conditional to boarding line i is:

$$EW_i = \int_0^\infty w\gamma_i(w)dw = \int_0^\infty wf_i e^{-(\sum_j f_j)w}dw = \frac{f_i}{\mathfrak{F}^2} \tag{21}$$

Summing over all the lines $FS(n)$ gives us the expected wait time at the stop, i.e.,

$$EW_n = \sum_{i\in FS(n)} \int_0^\infty w\gamma_i(w)dw = \int_0^\infty w \sum_i \gamma_i(w)dw \tag{22}$$

where, $\sum_i \gamma_i(w)$ is the probability density function of the waiting time at stop n.

$$\sum_{i\in FS(n)} \gamma_i(w) = \sum_{i\in FS(n)} f_i e^{-(\sum_j f_j)w} = \mathfrak{F}e^{-\mathfrak{F}w}, w \geq 0 \tag{23}$$

Therefore, the expected wait time at stop n if given by $EW_n = \frac{1}{\mathfrak{F}}$

Proof of Proposition 2

The probability of taking transit is given by:

$$P_{transit} = \int_0^\infty (\sum_i \gamma_i(w))Prob(w \leq w_{MoD})dw \tag{24}$$

$$P_{transit} = \int_0^\infty \mathfrak{F}e^{-\mathfrak{F}w} \times e^{-(f_{MoD})w}dw = \frac{\mathfrak{F}}{\mathbb{F}} \tag{25}$$

$$\tag{26}$$

Similarly, the probability of taking MoD is given by $P_{MoD} = \frac{\mathfrak{F}}{\mathbb{F}}$. The expected wait time of the passenger departing from an access node n is given by:

$$EW_n = \int_0^\infty w\left(\mathfrak{F}e^{-(f_{MoD}+\mathfrak{F})} + f_{MoD}e^{-(f_{MoD}+\mathfrak{F})}\right)dw = \frac{1}{\mathbb{F}} \tag{27}$$

Proof of Proposition 3 To prove this, we need to show that (a) $proj_{v,W}\mathcal{X}^{SP} \subseteq \mathcal{X}^{MA}$ and (b) $\mathcal{X}^{MA} \subseteq proj_{v,W}\mathcal{X}^{SP}$. Let us first start by proving (b). Let $(v, W) \in \mathcal{X}^{MA}$. For all $a \in FS(i) : a \in A_T, \forall i \in N^w, \forall k \in D$, we have $v_{ak} \leq f_a W_{ik}$. Let $y_{l(a)f} = 1$ if the frequency of the line associated to arc a is $f \in \Theta$ and 0, otherwise. Let $t_{faik} = \hat{y}_{l(a)f}W_{ik}$, then $f_a W_{ik} = \sum_{f\in\Theta} f\hat{y}_{l(a)f}W_{ik} = \sum_{f\in\Theta} ft_{faik}$, which is same as (14c). Also, $t_{faik} = \hat{y}_{l(a)f}W_{ik}$ can be expressed as (14d)- (14f) and (14l). Using a similar argument, we can show that for all $a \in FS(i) : a \in A_R, \forall i \in N^w, \forall k \in D$, the inequality $v_{ak} \leq f_a W_{ik}$ can be expressed as (14g)-(14j) and (14m). This shows that $\mathcal{X}^{MA} \subseteq proj_{v,W}\mathcal{X}^{SP}$. To prove part (a), let $(v, W, \omega, t) \in \mathcal{X}^{SP}$, then using Fourier-Motzkin elimination, we have $(v, W) \in \mathcal{X}^{MA}$ (Conforti et al. 2014, Chapter 3).

Proof of Proposition 4 The set $\mathcal{X}^{SP}$ can be empty in two cases i.e., when there is no flow balance ($\sum_{k\in D}\sum_{i\in N} g_{ik} \neq 0$) or there does not exist a directed path from any node $i \in N$ to any destination k. However, it is not possible to have any of these cases because from the definition of g_{ik}, we have $\sum_{k \in D} \sum_{i \in N} g_{ik} = 0$ and since there is always at least 0.01 vehicle assigned to all the zones and the road network is connected, there always exists a path from any node $i \in N$ to any destination $k \in D$. Therefore, $\mathcal{X}^{SP} \neq \phi$.

Proof of Proposition 5 Due to limited fleet available, one can allocate $n \in \Omega$ vehicles in at most $\lfloor \frac{\bar{F}}{n} \rfloor$ zones.

Proof of Proposition 6 The proposition follows from the definition of minimal cover that we cannot provide the number of buses required in the minimal cover.