Submitted to *Transportation Science*
manuscript (Please, provide the mansucript number!)

Joint Frequency-Setting and Pricing Optimization on Multi-Modal Transit Networks at Scale

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, dbertsim@mit.edu

Yee Sian Ng
Operations Research Center, Massachusetts Institute of Technology, yeesian@mit.edu

Julia Yan
Operations Research Center, Massachusetts Institute of Technology, jyyan@mit.edu

Modern public transportation systems are increasingly complex: they are operated at a large scale, must support booming urban populations, and run under tight budget constraints. Additionally, passengers are able to make choices between a variety of commuting options. We develop formulations for minimizing system wait time in multi-modal networks, while accounting for operator budget constraints, capacity constraints, and passenger preferences. Furthermore, our algorithms run to near-optimality in minutes for city-sized networks. We demonstrate the benefit of setting schedule frequencies and prices jointly through case studies on real data from Boston and Tokyo. To our knowledge, ours is the first paper that addresses joint frequency-setting and pricing optimization for public transit networks and at scale.

Key words: multi-modal transit networks, nonlinear optimization, frequency-setting, pricing
History:

1. Introduction

Public transit systems face significant operating challenges, and the question of how to efficiently manage such systems is of crucial importance. In recent years, increasing urban populations and shrinking operating budgets have both contributed to significant overcrowding and delays during peak hours. In New York City, over one-third of the subway delays have been attributed to overcrowding (Fitzsimmons et al. 2017). Meanwhile, transit networks are growing and diversifying rapidly, making it increasingly important to understand how operating decisions influence ridership and demand between different route options.

The controls available to improve transit operations are numerous. Although an optimized schedule might improve transit operations across multiple services, further coordination can be attained in capacity-constrained networks through the effectiveness of pricing schemes. In this regard, we are motivated by modern developments in private ride-sharing services such as Uber and Lyft, as well as public infrastructure such as electronic road pricing schemes and time- and distance-based pricing policies on public transit networks. In this paper, we will focus on coordinated frequency-setting and pricing in congested urban settings, while accounting for commuter preferences between multiple route choices.

1.1. Literature Review

Optimal scheduling of transit services has been well-studied in the literature. The train timetabling problem in particular has been studied under numerous variants. Szpigel (1973) first proposed an integer optimization formulation to minimize total travel time subject to coordination constraints. Since then, objectives such as profit (Brännlund et al. 1998, Yan and Chen 2002) or closeness to an ideal timetable (Caprara et al. 2002) have also been studied. The timetabling setting that is closest to our commuter-centric focus was studied by Niu and Zhou (2013), who formulated an integer optimization model that sought to minimize total commuter waiting time on a single congested urban rail line, and solved it using a genetic algorithm. Castelli et al. (2004) similarly sought to minimize total commuter waiting time, but in a multi-line setting, and solved the optimization problem using a Lagrangian-based heuristic.

One real-world drawback of timetabling is that complications such as traffic and variable dwell times make it difficult to adhere to a fixed timetable in congested urban settings. Frequency-setting addresses this challenge by determining departure rates over longer time horizons; the key difference here is that the timetabling problem is a discrete optimization problem, whereas in frequency-setting, the departures of buses and trains are modeled as continuous or fluid flows. Early work focused on analytical results on idealized single-origin-destination, infinite-capacity routes: Newell (1971) computed dispatch rates to minimize passenger waiting time, while Hurdle (1973) extended these results to a multiobjective framework that also incorporates operating costs. To account for multiple origins and destinations and capacity constraints, Schéele (1980) formulated a more general nonlinear optimization problem that set bus schedule frequencies to minimize total commuter waiting time.

The above timetabling and frequency-setting examples make demand to be fixed and known. A natural extension to these transit management problems arises when demand is not a fixed input, but rather is influenced by the controls that a transit operator applies. An early paper by Furth and Wilson (1981), which set headways to maximize social benefit, addressed the elasticity of demand by taking ridership to be a function of bus headways. However, the major simplifying assumption they take, that demand on each bus line is independent from that on other lines, is not entirely realistic. In modern transit settings, commuters are faced with a variety of substitutable route choices, and demand spilled by one route may be recaptured by another.

The general route choice setting we consider is one where each route may vary in its appeal, and commuters select the routes depending on each route's relative appeal. Common assumptions in modeling route choice are that (i) vehicles arrive randomly, and have headways following an exponential distribution with mean equal to the inverse of schedule frequency, and (ii) when presented with multiple attractive route options, commuters will choose that which has the first arriving vehicle (Spiess and Florian 1989, Nguyen et al. 1988).

Another important component to the appeal of a route, in addition to the schedule frequencies, is the price charged. A variety of pricing policies have been implemented in transit systems across the world. In New York City, the Metropolitan Transit Authority charges a flat fare for all subway and bus trips, while the Long Island Rail Road employs both zone-based fare and peak-hour congestion pricing. In traffic settings, congestion pricing has also been employed in cities such as London (Beevers and Carslaw 2005), Stockholm (Eliasson and Mattsson 2006), and Singapore (Goh 2002).

The consideration of frequency-setting and pricing in coordination motivates our use of choice modeling in representing commuters' responses to transit operating controls. The classical choice model is the multinomial logit (MNL) model, which was first proposed in McFadden (1974) in the context of choosing residential locations. Under the MNL model, commuters randomly select a route from all available route options; typically, options with higher utilities are selected more frequently. The problem of fitting transit logit models that account for features such as travel time and cost is well-studied (Nuzzolo et al. 2000, Brands et al. 2014). For a broad overview of choice modeling as it applies to route choice in transit, see Ben-Akiva and Bierlaire (1999), Prato (2009), Bovy (2009).

There are a variety of existing studies on the problem of transit management subject to route choice. Han and Wilson (1982) studied a simplified model of a congested bus network where route choices are proportional to the relative frequencies of the bus routes, and demonstrated their results on a toy six-station network. Shih and Mahmassani (1994) assumed that passengers respond to the changes in the vehicle frequencies according to a slightly modified version of the passenger assignment model of Han and Wilson (1982), and used an iterative approach to set vehicle sizes and schedule frequencies. Constantin and Florian (1995) followed the commuter route choice strategy outlined in Spiess and Florian (1989) to address commuters' responses to transit schedules, and developed a bilevel optimization framework to set schedule frequencies accordingly. Yu et al. (2009) modeled a similar problem, but additionally accounted for the fact that for buses arriving less frequently, passengers are likely to synchronize their arrivals with a given timetable. Parbo et al. (2014) used an all-or-nothing route assignment model that assigned passengers to the route of highest utility, where utility was based on total travel time, and developed timetables that would synchronize transfers between bus lines. Verbas and Mahmassani (2015) use bilevel optimization to set schedule frequencies that minimize waiting time in the upper level while modeling passenger responses to the frequencies in the lower level. Robenek et al. (2016) use integer optimization to set timetables that maximize operating profit while ensuring that passengers maintain a minimum level of utility.

The price elasticity of demand has been studied in several separate contexts from scheduling. For congestion pricing in transit, Huang (2002) assumed logit preferences that respond to price, while taking the schedule as fixed, and derives equilibrium conditions on a toy two-node system with two route options. Wu et al. (2011) and Wu et al. (2012) set tolls and public transit prices in a coordinated, multi-modal network equilibrium model to reduce total travel time in a combined traffic-transit network. In the field of airline operations, Atasoy et al. (2014) also assumed logit preferences that respond to price, and then performed fleet assignment and pricing in coordination.

1.2. Our Contribution

We provide a unified framework that allows us to:

1. Jointly optimize schedule frequencies and prices together,
2. Represent dynamic demand that responds to transit operating decisions through commuter route choice modeling, and
3. Tractably model city-scale networks and quantify the impact of policy changes to congested transit systems.

To our knowledge, ours is the first paper that addresses joint frequency-setting and pricing optimization for public transit networks and at scale. To demonstrate the flexibility of our framework, we ran it on two different networks: one from Kanagawa prefecture, Tokyo, Japan, and the other from Boston, Massachusetts. We observe that coordinated frequency-setting and pricing and the detailed modeling of commuter choices can allow for transit operations that incentivize commuters into route choices that benefit system performance, and our model provides a way of quantifying this benefit. We show in two computational case studies that schedule frequencies and prices may be found that bring the system close to its optimal potential performance, as measured by comparison to the *system optimum*.

The rest of this paper is organized as follows. Section 2 defines notation used throughout the paper. In Section 3, we elaborate on the recourse function that we use for evaluating total waiting time. In Section 4, we consider choice models for commuter preferences. In Section 5, we perform computational experiments on two real-world transit networks. Finally, we offer concluding remarks in Section 6.

2. Preliminaries

We consider the problem of scheduling transit *services*, such as buses, subways, or trains, on a network when commuters have multiple options available to them. As each service is scheduled to travel along a fixed sequence of stops, we will call these sequences of stops *lines*.

Table 1 Summary of the Notation

Symbol	Description
B	budget available for scheduling transit services (buses, subways, trains)
c^ℓ	cost to deploy a transit service along line ℓ, which can vary depending on characteristics of line ℓ such as distance or mode
K^ℓ	capacity of transit service running on line ℓ
L	number of transit lines
T	number of time periods
$\Delta(u,v,r)$	time periods it takes to travel from stop u to stop v using route r, while in-vehicle
$\texttt{legs}(u,v,r)$	the sequence of lines that a commuter would take from stop u to stop v while using route r
$\texttt{routes}(u,v)$	the set of all routes that can take a commuter from origin u to destination v
$\texttt{stops}(\ell)$	the set of all stops on line ℓ

Within the transit network, we refer to each origin-destination pair (u,v) as a *commute*. Each commute may be associated with a number of different *route* options, each of which is a different sequence of lines that can take a commuter from origin u to destination v. The travel on each individual line in the sequence is called a *leg* of the commute. The problem parameters are described in Table 1, and definitions of routes and legs are illustrated in Figure 1, a screenshot taken from Google Maps. In this example, the commute of interest is from Yokohama Station to Ofuna Station. Two route options are offered: the first route consists of two legs (Yokosuka and Tokaido), and the second route consists of a single leg (Negishi).

Given data on $\mathbf{d} = (d_t^{u,v})$, where $d_t^{u,v}$ = demand for commute (u,v) that arrives at station u at time t, we wish to construct a schedule $\mathbf{x}$ that minimizes total commuter waiting time. In practice, demand would also depend upon the choice (if any) of routes that the commuters might take through the network. We use $\boldsymbol{\theta} = (\theta_t^{u,v,r})$ to represent commuter choices, where each element $\theta_t^{u,v,r}$ represents the proportion of commuters for commute (u,v) at time t who chose to take route r.

For transit scheduling, we discretize the full schedule into time periods $t = 1,\dots,T$, and are interested in providing decisions on $\mathbf{x} = (x_t^\ell)_{t=1,\dots,T}^{\ell=1,\dots,L}$, where

$$x_t^\ell = \begin{cases} 1 & \text{if a transit service is scheduled for departure from the start of line } \ell \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the set of feasible schedules is denoted by

$$\mathcal{X}_B = \left\{ \mathbf{x} \in \{0,1\}^{T\times L} : \sum_{t=1}^{T} \sum_{\ell=1}^{L} c^\ell x_t^\ell \le B \right\}. \tag{1}$$

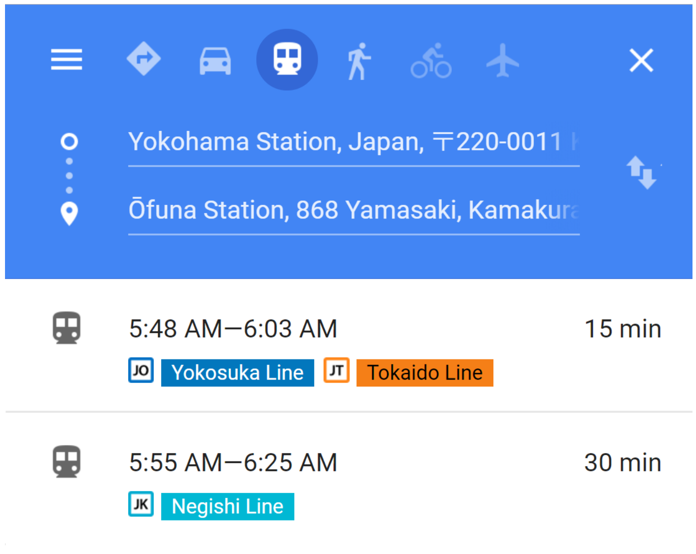

Figure 1 **A screenshot of Google Maps illustrating route options in Japanese transit. (Accessed 2017-11-03)**

The constraint in (1) enforces the requirement that the total number of services cannot exceed that allowed by the budget B. The cost c^ℓ is allowed to vary by each line ℓ, to account for the fact that different modes of transportation are associated with different costs.

The linear optimization relaxation

$$\bar{\mathcal{X}}_B = \left\{ \mathbf{x} \in [0,1]^{T \times L} : \sum_{t=1}^{T} \sum_{\ell=1}^{L} c^\ell x_t^\ell \leq B \right\}. \tag{2}$$

of the budget constraint (1) represents the set of fluid rates of departure for the transit services, rather than a concrete timetable for operators to follow. For example, a value of $x_t^\ell = 0.5$ means that a train would depart once every two time periods. The practical interpretation of the fluid model is that it primarily models waiting time due to capacity constraints, while ignoring waiting time incurred if commuters arrive between service arrivals. This simplification is useful in peak traffic hours when there is high congestion and trains arrive at a high frequency; here, the capacity restrictions, not the train interarrival wait time, are of primary concern. Another motivator of frequency-setting as opposed to timetabling is that during congested peak hours, frequency guidelines are more implementable than exact timetables. Hereafter, we focus on the relaxed constraint (2), which only requires solving a linear optimization problem.

The above formulation contains some basic operational considerations to a transit authority, but can easily be expanded to incorporate other specific constraints that might be of interest. We provide some examples below.

1. **Lower and upper bounds on schedule frequencies.** It may be desirable to ensure a minimal level of service on a particular line, or it may be impossible to operate services above a

particular level of frequency without collisions. If $\underline{B}$ and $\overline{B}$ are the service lower and upper bounds, then these requirements can be modeled with the following constraint:

$$\underline{B} \leq \mathbf{x} \leq \overline{B}.$$

2. **Service coordination.** Certain network structures may require that transit services on separate lines share common resources, such as a lane in the train tracks. Suppose that lines ℓ_1 and ℓ_2 are two such lines, and that it takes δ_1 and δ_2 time periods for trains on lines ℓ_1 and ℓ_2 to reach the point of conflict, respectively. To ensure that the shared area can only accommodate a single service for any given period, we can introduce the constraint

$$x_{t+\delta_1}^{\ell_1} + x_{t+\delta_2}^{\ell_2} \leq 1$$

for all relevant time periods t.

3. Optimization with Known Choices

To model the progress of the commuters through the network, we introduce boarding variables $\mathbf{z} = (z_t^{u,v,r,i})$, where $z_t^{u,v,r,i}$ = number of commuters traveling from origin u to destination v, taking route option r, on the i-th leg of their itinerary, boarding a service that had started at time t. For clarity of presentation, we write our formulation ignoring vehicle travel time between stations; accounting for vehicle travel time is ultimately a straightforward procedure of adjusting the time indices, but comes at the expense of significantly heftier notation. In this section, we also make the assumption that the commuter choice probabilities $\boldsymbol{\theta}$ are known, and satisfy $0 \leq \theta_t^{u,v,r} \leq 1$ and $\sum_{r \in \texttt{routes}(u,v)} \theta_t^{u,v,r} = 1$. This assumption is only for simplicity of exposition and will be revisited in Section 4. Then the transit frequency-setting problem can be stated as:

$$\min_{\mathbf{x} \in \mathcal{X}_B} Q(\mathbf{x}, \boldsymbol{\theta}), \tag{3}$$

where

$$Q(\mathbf{x}, \boldsymbol{\theta}) = \min_{\mathbf{z} \geq 0} \; J(\mathbf{z}, \boldsymbol{\theta}) \tag{4a}$$

$$\text{s.t. } O_{\ell,u,t}(\mathbf{z}) \leq K^\ell x_t^\ell \qquad \forall \ell = 1, \ldots, L, \quad \forall u = \texttt{stops}(\ell), \quad \forall t = 1, \ldots, T \tag{4b}$$

$$\sum_{t'=1}^{t} z_{t'}^{u,v,r,1} \leq \sum_{t'=1}^{t} d_{t'}^{u,v} \theta_{t'}^{u,v,r} \qquad \forall u, v = 1, \ldots, N, \quad \forall r = \texttt{routes}(u,v), \quad \forall t = 1, \ldots, T \tag{4c}$$

$$\sum_{t'=1}^{t} z_{t'}^{u,v,r,i} \leq \sum_{t'=1}^{t} z_{t'}^{u,v,r,i-1} \qquad \forall u,v=1,\ldots,N \quad \forall r=\texttt{routes}(u,v), \quad \forall i=2,\ldots,|\texttt{legs}(u,v,r)|, \quad \forall t=1,\ldots,T. \tag{4d}$$

The function $J(\mathbf{z},\boldsymbol{\theta})$ in (4a) computes the total waiting time, by adding the number of commuters who have arrived or transferred to each station and subtracting the commuters who have boarded their transit. More specifically, the total number of passengers waiting at a station u on line ℓ at time t is given by

$$AD_{\ell,u,t}(\boldsymbol{\theta}) + XD_{\ell,u,t}(\mathbf{z}) - BD_{\ell,u,t}(\mathbf{z}), \tag{5}$$

where $AD_{\ell,u,t}(\boldsymbol{\theta})$ represents "arriving" demand, $XD_{\ell,u,t}(\mathbf{z})$ represents "transferring demand", and $BD_{\ell,u,t}(\mathbf{z})$ represents "boarding demand", all cumulative up until time t. The total time passengers spend waiting across all lines, stations, and time periods is then obtained by aggregating equation (5) as follows:

$$J(\mathbf{z},\boldsymbol{\theta}) := \sum_{\ell=1}^{L} \sum_{u\in\texttt{stops}(\ell)} \sum_{t=1}^{T} AD_{\ell,u,t}(\boldsymbol{\theta}) + XD_{\ell,u,t}(\mathbf{z}) - BD_{\ell,u,t}(\mathbf{z}). \tag{6}$$

In equations (5) and (6), the arrivals quantity $AD_{\ell,u,t}(\boldsymbol{\theta})$ represents the total demand that has arrived to station u on line ℓ by time period t, and is computed as follows:

$$AD_{\ell,u,t}(\boldsymbol{\theta}) := \sum_{\{(v,r):\texttt{legs}(u,v,r)_1=\ell\}} \sum_{t'=1}^{t} d_{t'}^{u,v} \theta_{t'}^{u,v,r}, \tag{7}$$

where the condition $\{(v,r):\texttt{legs}(u,v,r)_1=\ell\}$ indicates that the first leg of the commute must be on line ℓ.

The transferring demand quantity $XD_{\ell,u,t}(\mathbf{z})$ represents the total number of passengers who have arrived to station u on line ℓ by time period t, having transferred over from another line. It is computed as follows:

$$XD_{\ell,u,t}(\mathbf{z}) := \sum_{\{(w,v,r,i)\in\texttt{xfrthru}(\ell,w)\}} \sum_{t'=1}^{t} z_{t'}^{w,v,r,i-1}, \tag{8}$$

where the set $\texttt{xfrthru}(\ell,u)$ represents all of the commute-route-legs that make a transfer through station u on line ℓ. Specifically, when considering a commute (w,v), route option $r\in\texttt{routes}(w,v)$, on the ith leg of the commute, (w,v,r,i) must satisfy the criteria that (i) $i\geq 2$, (ii) the transfer station connecting the $(i-1)$th and ith legs of the itinerary is station w on line ℓ.

The boarding demand quantity $BD_{\ell,u,t}(\mathbf{z})$ represents the total number of passengers who have managed to board a train at station u on line ℓ by time period t. It is computed as follows:

$$BD_{\ell,u,t}(\mathbf{z}) := \sum_{\{(w,v,r,i)\in \texttt{brdat}(\ell,u)\}} \sum_{t'=1}^{t} z_{t'}^{w,v,r,i}, \tag{9}$$

where the set $\texttt{brdat}(\ell, u)$ represents all of the commute-route-legs that require boarding a vehicle at station u on line ℓ. This includes commutes where the origin station is u and the origin line is ℓ, as well as commutes where station u and line ℓ are the location of some later transfer.

The function $O_{\ell,u,t}(\mathbf{z})$ in (4b) computes the occupancy of a service as it passes through stop u on line ℓ, having started its run at time t. It is computed as the following summation:

$$O_{\ell,u,t}(\mathbf{z}) := \sum_{\{(w,v,r,i)\in \texttt{passthru}(\ell,u)\}} z_t^{w,v,r,i}, \tag{10}$$

where the set $\texttt{passthru}(\ell, u)$ represents all of the commute-route-legs that pass through station u on line ℓ. Specifically, when considering a commute (w, v), route option $r \in \texttt{routes}(w, v)$, and on the ith leg of the commute, (w, v, r, i) must satisfy the criteria that (i) the ith element of $\texttt{legs}(w, v, r)$ is ℓ, (ii) the transfer station connecting the $(i-1)$th and ith legs of the itinerary is *at or before* station w on line ℓ, and (iii) the transfer station connecting the ith and $(i+1)$th legs of the itinerary is *after* station w on line ℓ. With this definition of $O_{\ell,u,t}(\mathbf{z})$, constraint (4b) then ensures that the number of commuters on board a transit vehicle must be within the vehicle capacity K^ℓ.

Finally, the boarding and transfer constraints (4c) and (4d) enforce the requirement that commuters cannot embark on the ith leg of their commute until they have completed all previous legs.

Both $J(\mathbf{z}, \boldsymbol{\theta})$ and $O_{l,u,t}(\mathbf{z})$ are linear functions of the decision variables $\mathbf{z}$ in the case where the commuter choice probabilities $\boldsymbol{\theta}$ are known. Therefore, equation (3) is a linear optimization problem when the commuter choice probabilities are known.

4. Optimization with Design-Dependent Choices

In this section, we examine a framework where addition to setting the schedule frequencies, the service operator can set prices $\mathbf{p} = (p^{u,v,r})$, where $p^{u,v,r}$ represents the price charged for commute (u, v) and route r. The indices for $\mathbf{p}$ are intentionally detailed to accommodate the variety of pricing policies that could be implemented. Such policies can be described with the addition of the appropriate constraints and auxiliary variables, some examples of which are given below.

1. **Flat fare.** If the transit operator wants to charge a flat fare for all commuters entering the system, this can be accomplished by adding the constraint

$$p^{u,v,r} = f \tag{11}$$

for all commutes (u, v) and routes r, where f represents the value of the flat fare.

2. **Line-based fare.** If the commuters are charged for each service they use, then auxiliary variables f^ℓ are introduced to represent the fare charged for each service. In addition, the constraints

$$p^{u,v,r} = \sum_{\ell \in \texttt{legs}(u,v,r)} f^\ell \tag{12}$$

should be added for every commute (u,v) and route r.

3. **Distance-based fare.** A more equitable pricing policy than the line-based pricing would be to weight the fare charged on each line by the distance that the commuter traveled along that line. Denoting that distance by the constant $h^{u,v,r,\ell}$ for commute (u,v), route r, and line ℓ, the constraints

$$p^{u,v,r} = \sum_{\ell \in \texttt{legs}(u,v,r)} h^{u,v,r,\ell} f^\ell \tag{13}$$

then represent this distance-based pricing policy.

In a further extension to the framework in Section 3, we also allow that the route choices are not fixed, but can depend on the decisions made by the service operator, namely, the optimization decision variables. We write the choice probabilities as $\boldsymbol{\theta}(\mathbf{x},\mathbf{p})$ to show that they can depend on the schedule frequencies $\mathbf{x}$ and prices $\mathbf{p}$. The schedule frequencies and prices can then be set through the objective function

$$\min_{\mathbf{x} \in \bar{\mathcal{X}}_B, \mathbf{p} \geq 0} Q(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x},\mathbf{p})). \tag{14}$$

The choice probabilities $\boldsymbol{\theta}(\mathbf{x},\mathbf{p})$ are influenced by the utility that a commuter would gain from taking each route option, which we express as $\mu_t^{u,v,r}(\mathbf{x},\mathbf{p})$ for commute (u,v) and route choice r under schedule frequencies $\mathbf{x}$ and prices $\mathbf{p}$. To specify $\mu_t^{u,v,r}(\mathbf{x},\mathbf{p})$, we assume that the crucial attributes in a commuter's utility function are *time*, *cost*, and *comfort*. The time component can be described as time spent waiting to board a service, plus time spent on the service. The time spent waiting to board a service can often include not just time spent waiting for the next service arrival, but also any waiting time that is incurred because the service cannot fit any more commuters. We quantify comfort using the percent occupancy of the transit service, since that is the primary feature of commuter comfort that is endogenous to the model.

Concretely, assuming that services at time t on line ℓ are arriving uniformly at rate x_t^ℓ, an arriving commuter should expect on average to wait $\frac{1}{2x_t^\ell}$ for a service to arrive. This waiting time is incurred at every leg $\ell \in \texttt{legs}(u,v,r)$. We represent the remaining time of the commute, which is spent on the service, with the term $\Delta(u,v,r)$. For the cost attribute, we have previously defined $p^{u,v,r}$ as the price charged to this commuter. Assuming that the commuter utility function is linear in the time, cost, and comfort attributes, the form of $\mu_t^{u,v,r}(\mathbf{x},\mathbf{p})$ is then given by:

$$\mu_t^{u,v,r}(\mathbf{x},\mathbf{p}) := -\beta_1 \left(\sum_{\ell \in \texttt{legs}(u,v,r)} \frac{1}{2x_t^\ell} + \Delta(u,v,r) \right) - \beta_2 p^{u,v,r} - \beta_3 \left(\sum_{\ell \in \texttt{legs}(u,v,r)} \psi \left(\frac{O_{\ell,u,t}(\mathbf{z})}{\bar{K}^\ell x_t^\ell} \right) \right), \tag{15}$$

where β_1 represents marginal utility of time, β_2 represents the marginal utility of money, and β_3 represents the marginal utility of comfort. It is of course possible to weight the different time components of the utility function separately if commuters treat waiting time differently from in-vehicle time. The last term $\sum_{\ell \in \texttt{legs}(u,v,r)} \psi(\frac{O_{\ell,u,t}(\mathbf{z})}{\bar{K}^\ell x_t^\ell})$ represents our quantification of comfort from the percent occupancies of the relevant transit services, where $\bar{K}^\ell \leq K^\ell$ is a threshold "soft" capacity. This is in contrast to constraint (4b), which is a "hard" capacity constraint. The discomfort function $\psi : \mathbb{R} \to \mathbb{R}$ largely follows the definition in İmre and Çelebi (2017), and has been modified to be convex and differentiable:

$$\psi(\kappa) = \begin{cases} \kappa & \text{if } \kappa \leq 1 \\ e^{\kappa - 1} & \text{if } \kappa > 1. \end{cases} \tag{16}$$

Intuitively, the discomfort increases linearly with the occupancy $O_{\ell,u,t}(\mathbf{z})$ up until the occupancy reaches the soft threshold $\bar{K}^\ell x_t^\ell$, and exponentially thereafter. The discomfort term poses some particular challenges which we will address in Section 4.2.

4.1. Discrete Choice Models in Transit Assignment

With the routes and the associated utilities in hand, one possible choice model is the *first-choice* model (McBride and Zufryden 1988), which is also called the *pure characteristics* model in the economics literature (Berry and Pakes 2007, Dobson and Kalish 1988) or the *all-or-nothing* model in the traffic literature. Under this model, the commuter chooses the route that provides her with highest utility, so that the choice probability is given by

$$\theta_t^{u,v,r}(\mathbf{x}, \mathbf{p}) := \mathbb{1}(\mu_t^{u,v,r}(\mathbf{x}, \mathbf{p}) = \max_{r' \in \texttt{routes}(u,v)} \mu_t^{u,v,r'}(\mathbf{x}, \mathbf{p})). \tag{17}$$

However, this assumption that commuters behave exactly according to the specified utility function as a monolith is not entirely realistic, and is particularly unstable to problem parameters. We therefore turn to the multinomial logit model (McFadden 1974), where the setup is similar to that in the first-choice model, but the utility function $\mu_t^{u,v,r}(\mathbf{x}, \mathbf{p})$ is also associated with a Gumbel-distributed noise parameter. Due to the noise parameter, commuters will probabilistically pick routes with the choice probabilities given by:

$$\theta_t^{u,v,r}(\mathbf{x}, \mathbf{p}) := \frac{\exp(\mu_t^{u,v,r}(\mathbf{x}, \mathbf{p}))}{\sum_{r' \in \texttt{routes}(u,v)} \exp(\mu_t^{u,v,r'}(\mathbf{x}, \mathbf{p}))}. \tag{18}$$

Finally, for purposes of comparison, we introduce the concept of the *system optimum*. In this model, we allow the choice probabilities $\boldsymbol{\theta}$ to be any valid probability distribution over the route options. In this case, prices are irrelevant, and schedule frequencies are set according to the formulation

$$\min_{\mathbf{x} \in \bar{\mathcal{X}}_B, \boldsymbol{\theta}} \quad Q(\mathbf{x}, \boldsymbol{\theta}) \tag{19a}$$

$$\text{s.t} \quad \sum_{r \in \texttt{routes}(u,v)} \theta_t^{u,v,r} = 1 \qquad \forall u, v = 1, \ldots, N \quad \forall t = 1, \ldots, T \tag{19b}$$

$$\boldsymbol{\theta} \geq 0. \tag{19c}$$

The interpretation of (19) is that the system operator is able to direct passengers to take whatever route would benefit the system most. For example, if a particular route is especially congested, the system operator could direct passengers to alternative routes. Such a dictatorial policy would be impossible to implement, but the objective value of (19) serves as a useful lower bound for the best waiting time that any transit policy could achieve.

The multinomial logit model (18) suffers from a property known as the independence of irrelevant alternatives (IIA). An alternative choice model is the nested logit model, which is a generalization of the multinomial logit that has been shown by Williams (1977) to be consistent if the IIA assumption does hold. Following the example in Lo et al. (2004), one approach is to classify all the route choices into a small set of "travel modes" based on the combination of services used by that route. Under this model, commuters make route choices based on a two-level choice model: the first level decides the mode, and the second level decides the route options available under that mode. The resulting choice probabilities will then look like

$$\theta_t^{u,v,r}(\mathbf{x}, \mathbf{p}) := \frac{\exp(\alpha_1 \bar{\mu}_t^{u,v,m}(\mathbf{x}, \mathbf{p}))}{\sum_{m' \in \texttt{modes}(u,v)} \exp(\alpha_1 \bar{\mu}_t^{u,v,m'}(\mathbf{x}, \mathbf{p}))}, \tag{20}$$

where $\texttt{modes}(u,v)$ is the set of travel modes available to go from u to v, α_1 is the coefficient of perceptional variation between different travel modes, and $\bar{\mu}_t^{u,v,m}$ determines the utility for commuters to choose travel mode m as

$$\bar{\mu}_t^{u,v,m}(\mathbf{x}, \mathbf{p}) = \frac{1}{\alpha_2} \ln \left(\sum_{r \in \texttt{routes}(u,v;m)} \exp(\alpha_2 \mu_t^{u,v,r}(\mathbf{x}, \mathbf{p})) \right), \tag{21}$$

where $\texttt{routes}(u,v;m)$ is the set of routes under travel mode m, and α_2 is the corresponding perception variation parameter. In practice, both α_1 and α_2 will have to be estimated from commuter survey data. The gradient $\nabla \theta_t^{u,v,r}(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ for both the multinomial logit (18) and the nested logit (20) can be easily derived using a recursive application of the chain rule, or computed using automatic differentiation (Bartholomew-Biggs et al. 2000).

4.2. Solution Algorithm

The transit frequency-setting problem with known choices (3) is a linear optimization problem, and can be solved efficiently. Similarly, the calculation of the system optimum (19) is a linear optimization problem. However, the transit frequency-setting and pricing problem (14), with the choice probabilities given according to the multinomial logit model (18) or the nested logit model

(20), is a nonconvex optimization problem. To solve this problem, we use a first-order method which replaces the nonlinear function with a series of locally linear approximations. Around a particular point $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$, the choice probability can be approximated as

$$\hat{\theta}_t^{u,v,r}(\mathbf{x}, \mathbf{p}; \bar{\mathbf{x}}, \bar{\mathbf{p}}) \approx \theta_t^{u,v,r}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) + \nabla\theta_t^{u,v,r}(\bar{\mathbf{x}}, \bar{\mathbf{p}})'(\mathbf{x} - \bar{\mathbf{x}}, \mathbf{p} - \bar{\mathbf{p}}), \tag{22}$$

provided the point $(\mathbf{x}, \mathbf{p})$ is close to the original point $(\bar{\mathbf{x}}, \bar{\mathbf{p}})$. Equation (22) is linear in the decision variables $\mathbf{x}$ and $\mathbf{p}$. Substituting the linearized choice probability approximations $\hat{\boldsymbol{\theta}}$ in (22) for the multinomial logit probabilities (18) in formulation (14) turns the problem into a linear optimization problem. Then, beginning with feasible starting schedule frequencies and prices $(\mathbf{x}^{(0)}, \mathbf{p}^{(0)})$, a new $(\mathbf{x}^{(i)}, \mathbf{p}^{(i)})$ can be produced for each iteration $i = 1, \ldots,$ by solving the problem

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{p}} \quad & Q(\mathbf{x}, \hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{p}; \mathbf{x}^{(i-1)}, \mathbf{p}^{(i-1)})) && \text{(23a)}\\
\text{s.t.} \quad & \mathbf{x} \in \bar{\mathcal{X}}_B && \text{(23b)}\\
& \mathbf{p} \geq 0 && \text{(23c)}\\
& \mathbf{x}^{(i-1)} - \eta \leq \mathbf{x} \leq \mathbf{x}^{(i-1)} + \eta && \text{(23d)}\\
& \mathbf{p}^{(i-1)} - \gamma \leq \mathbf{p} \leq \mathbf{p}^{(i-1)} + \gamma, && \text{(23e)}
\end{aligned}$$

where η and γ are constants referring to the step-sizes for schedule frequencies and prices, respectively. Formulation (23) solves the optimization problem with design-dependent choices when the frequencies $\mathbf{x}$ and prices $\mathbf{p}$ lie within the small intervals $[\mathbf{x}^{(i-1)} - \eta, \mathbf{x}^{(i-1)} + \eta]$ and $[\mathbf{p}^{(i-1)} - \gamma, \mathbf{p}^{(i-1)} + \gamma]$, in which the choice probabilities can be linearly approximated. In this way, we solve the nonconvex optimization problem using a series of linear optimization problems. The constants η and γ should be chosen to be small enough that the approximation (22) is reasonably accurate within the frequency and price intervals, but large enough that progress can be made quickly. Advantages of such first-order methods in general include their speed and simplicity of implementation. Furthermore, using a first-order approximation of the nonlinear terms rather than using point estimates between iterations is more accurate. However, it is possible in nonconvex problems such as ours to converge to suboptimal local extrema. To guard against this, we repeat our procedure at multiple random starting points and select the frequencies and prices producing the best objective value. For more on first-order methods, see Boyd and Vandenberghe (2004).

As previously mentioned, the term $O_{\ell,u,t}(\mathbf{z})$ in the utility function (15) poses a particular challenge. Using the full occupancy expression would make the utility functions extremely dense functions of the model decision variables $\mathbf{z}$. Rather than use the full expression, we take its value at iteration i to be the value of of the occupancy calculated from step $i-1$. Since the region in constraints (23d) and (23e) is small, the previous values are good estimates of the values in the current iteration.

To summarize, our model optimizes frequencies and prices in order to minimize waiting time due to congestion. Commuters respond to the frequencies and prices with choice probabilities that are determined by the utility they might derive from each route option. These choice probabilities are also decision variables in the optimization formulation, allowing commuters to respond dynamically to the decisions made by the transit operator. The full formulation is a nonconvex optimization model, but is solved efficiently using first-order methods.

5. Computational Experiments

We now turn to computational experiments on two real-world transit networks for which we obtained data. The first is a subset of the train network in Kanagawa prefecture, which is part of the metropolitan area surrounding Tokyo, Japan. The second is the subway and bus systems in Boston, Massachusetts, which are run by the Massachusetts Bay Transportation Authority (MBTA). These two transit systems are of particular interest due to their high utilization by a booming urban population, as well as the multiplicity of route choices available to commuters.

All methods were implemented using the optimization package `JuMP` (Lubin and Dunning 2015) in the `Julia` programming language (Bezanson et al. 2017), and solved using Gurobi 7.5 (Gurobi Optimization, Inc. 2016). Computational experiments were run on eight cores of a computer with a 16-core Intel Xeon E5-2650 CPU, 3.40 GHz processor, and 64GB of memory.

Before discussing our computational experiments in detail, we note that the purpose of the experiments is not to provide a policy prescription, but to show that our methods can be applied to realistic data. Due to the limited data we had available, we use the multinomial logit choice model (18), and describe our method for generating utility function coefficients in Appendix A. We do not claim that parameters such as relative costs, capacities, or coefficients in the utility function, are accurate, merely that they are reasonable enough to draw insight.

5.1. Case Study: Tokyo

Through East Japan Railway Company (2016), Odakyu Electric Railway Co. (2016), and City of Yokohama (2011), we obtained transit demand data for a subset of the train system in Kanagawa Prefecture, which is part of the Greater Tokyo Area in Japan. The transit network, displayed in Figure 2, is comprised of 57 stations along five lines: Blue, Tokaido, Negishi, Odakyu, and Yokohama. The demand data includes the precise time, origin, and destination of each commute for a typical evening, from 5-9pm. We divided time periods to be of length 15 minutes. Travel times along the network were obtained through queries to Google Maps. Route options were precomputed for the network using shortest-path-based methods. In general, it appeared that route options after the second-shortest path tended to involve much higher in-vehicle times and numbers of transfers, so these were discarded.

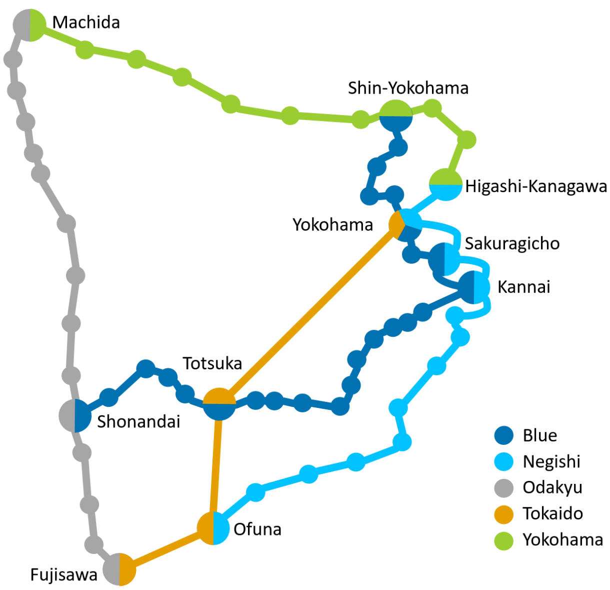

Figure 2 **(Color online) A map of the Kanagawa trains network.**

As shown in Figure 2, the Blue and Negishi (cyan) lines are local trains with closely-spaced stops, while the Tokaido (orange) line is an express line that connects some of the stations on the Blue and Negishi lines. Some example commutes and associated travel times are shown in Table 2, illustrating that the Tokaido line is significantly faster than both the Blue and Negishi lines, and the Blue line is slightly faster than the Negishi line.

Table 2 **Travel times for various origin-destination pairs on the Kanagawa network**

Origin	Destination	Line	Travel Time
Ofuna	Yokohama	Tokaido	16 min
		Negishi	34 min
Totsuka	Yokohama	Tokaido	10 min
		Blue	27 min
Kannai	Yokohama	Blue	5 min
		Negishi	7 min

In this case study, we first focus on an illustrative subset of our Kanagawa data that comprises the Blue, Negishi, and Tokaido lines during the rush hour from 5-6pm, where the transit dynamics are relatively simple and allow for straightforward interpretation of the optimal decision variables. We then move to the full network of all five lines for the full 5-9pm time range in our dataset to show tractability on a large network. Our findings are outlined as follows:

- In Section 5.1.1, we examine *optimal frequency-setting* on a subset of the Kanagawa network and show that it can have a significant impact on reducing total waiting time. We also show that the impact of frequency-setting on reducing waiting times becomes more pronounced as commuters' sensitivity to congestion increases.
- In Section 5.1.1, we also examine *optimal pricing* on a subset of the Kanagawa network and show that pricing is more effective in reducing waiting times when commuters are sensitive to congestion. Nonetheless, it is less effective than frequency-setting in reducing total waiting time.
- In Section 5.1.2, we perform coordinated frequency-setting and pricing and demonstrate that our methods are practically useful. We compared system performance for the full Kanagawa network under a variety of pricing policies and evaluate their social impact, and show that our methods are tractable for a city-size network with quality solutions obtained within minutes. We also show that *coordinated frequency-setting and pricing* significantly outperforms frequency-setting alone.

5.1.1. The effect of congestion on a small network In this section, we focus on a subnetwork comprising the Blue, Negishi, and Tokaido lines, where the dynamics are relatively simple. Of the 70,599 commuters on this subnetwork, 24,226 (34.3%) have multiple commuting options available to them. We refer to them as *multiple-option commuters*. The remaining 46,373 (65.7%) passengers only have one route option available to them, and we refer to them as *single-option commuters*. Among single-option commuters, the Blue line is the most highly sought-after line (22,206 passengers), followed by the Negishi line (15,162 passengers), and finally by the Tokaido line (14,021 passengers). Note that these numbers do not add up to the 46,373 single-option commuter total because some passengers must transfer between multiple lines to get from their origin to their destination. Although single-option commuters do not have preferences to be modeled, their presence on the trains impacts the congestion of each line and the decisions that are made to alleviate this congestion. The parameter $\bar{K}^\ell$ in the utility function (15) was set to $0.8K$, meaning that commuters begin to experience heightened disutility once the occupancy of the vehicle grows past 80% its total capacity.

In our first experiment, we examine the impact of commuters' sensitivity to congestion on optimal frequency-setting without pricing. For a range of budgets and β_3 parameters, we ran our solution algorithm from thirty initial starting points where the frequencies were set randomly and the prices were set to be identical for each line. Each instance terminated in under a minute, with many as quickly as several seconds. The optimal waiting times averaged across the total commuter demand are reported in the third column of Table 3, the next three columns break down the waiting time experienced across the three different lines, and the final three columns show the total number of trains allocated to each line. As commuters become more sensitive to discomfort due to congestion,

Table 3 Frequency-setting on a small network under varying budgets and commuter sensitivities to congestion.

Budget (trains)	β_3	Wait / Commuter (min) *Overall*	*Blue*	*Tokaido*	*Negishi*	Commuter / Train *Blue*	*Tokaido*	*Negishi*	Number of Trains *Blue*	*Tokaido*	*Negishi*
30	0	5.84	2.66	0.80	1.02	1,003	944	1,093	7.1	9.0	13.9
	1	5.79	2.53	0.79	1.05	974	957	1,100	7.4	8.7	13.8
	2	5.76	2.39	0.88	1.04	945	972	1,107	7.8	8.4	13.7
36	0	3.65	1.92	0.32	0.62	838	750	942	8.5	11.4	16.1
	1	3.34	1.60	0.35	0.67	780	764	968	9.3	11.0	15.6
	2	2.04	1.28	0.42	0.70	716	772	1,016	10.4	10.8	14.8
42	0	2.04	0.77	0.29	0.37	707	583	856	12.5	11.7	17.7
	1	1.87	0.71	0.23	0.36	692	575	876	12.7	12.0	17.3
	2	1.52	0.49	0.18	0.39	616	676	859	12.0	12.5	17.5

the average waiting time per commuter drops. This drop is driven largely by the commuters on the Blue line, which is the most highly sought-after line.

This phenomenon illustrates that optimal frequency-setting is more effective when the commuters are more sensitive to congestion. Intuitively, as service frequency increases on a line, all else being equal, more commuters will opt to use that service. A line that improves its service frequency must then be prepared to serve not only its existing commuters, but also the others who switch to its improved service. However, this effect is controlled by the commuters' disutility for congestion, since as the vehicles become more crowded, commuters will ultimately choose alternatives that are less crowded even if they are slower or less frequent. This effect is seen in the commuters-per-train columns of Table 3, which show that as the commuters become more sensitive to congestion, relatively fewer commuters choose to take the Blue line and instead move to the Tokaido or Negishi lines.

Table 4 Price-setting on a small network under varying budgets and commuter sensitivities to congestion

Budget (trains)	β_3	Wait/Commuter (min)	Pricing Premium *Blue*	*Negishi*	*Tokaido*	Number of Trains *Blue*	*Negishi*	*Tokaido*
30	0	6.38	2.05	0.00	1.25	12.8	9.3	7.9
	1	6.37	1.77	0.00	1.20			
	2	6.37	1.44	0.00	1.15			
60	0	4.86	1.00	0.00	1.50	15.4	11.1	9.5
	1	4.86	1.05	0.00	0.75			
	2	4.85	1.05	0.00	0.75			
90	0	3.83	1.00	0.00	0.70	18.0	13.0	11.0
	1	3.82	1.05	0.00	0.75			
	2	3.81	1.05	0.00	0.75			

In our next experiment, we illustrate the impact of commuters' sensitivity to congestion on pricing. The setup was identical as before, except that frequencies were fixed and pricing was the only lever to manage congestion. The frequencies were fixed roughly in proportion to the demand for each line, with the Blue, Negishi, and Tokaido lines receiving 21.4%, 15.5%, and 13.1% of the budget respectively. The optimal waiting times averaged across the total commuter demand are reported in the third column of Table 4, and the next three columns show the pricing premium set on each line relative to the Negishi line, which is consistently the cheapest line. The final three columns show the fixed number of trains that were allocated to each line. As in the frequency-setting case before, the waiting time declines as commuters become more sensitive to congestion although the effect is substantially less pronounced than before. Furthermore, for the lower budgets $B = 30$ and $B = 60$, the prices generally decline with increased β_3, an effect that is seen most clearly at the lowest budget. This illustrates that when commuters are sensitive to congestion, their natural tendency to avoid congestion allows the transit operator to achieve lower waiting times at lower prices.

5.1.2. Practical policy evaluation on the full network Returning to the full Kanagawa network and the full 5-9pm time range, we now demonstrate the tractability of our methods for even larger network sizes and longer time scales. Of the 362,477 commuters on this subnetwork, 183,681 (50.7%) are multiple-option commuters, and of the 201 origin-destination pairs with nonzero demand, 98 (48.8%) of them are multiple-option commutes. Table 5 shows the median run time for each budget, over 30 different random starting positions $(\mathbf{x}^{(0)}, \mathbf{p}^{(0)})$. Unsurprisingly, frequency-setting alone appears to be the easiest problem; however, even for distance- and line-pricing, typical run times are within several minutes.

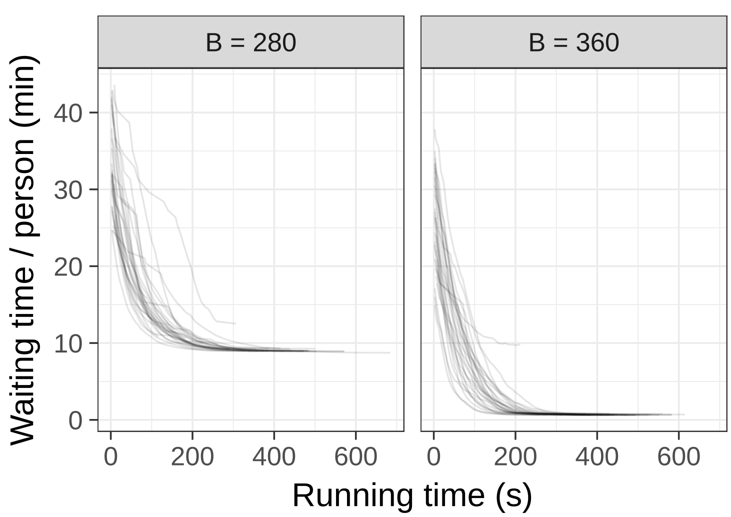

Figure 3 **Optimization progress for multiple different random starting points and two example budgets, on the full Kanagawa network and using frequency-setting and line pricing. Each line corresponds to the objective function value from a different random starting point.**

Figure 3 shows the progress of the objective function over time for a selection of budgets B when optimizing for both schedule frequencies and prices jointly, under a distance-based pricing policy. Each line on the plot corresponds to one of the 30 different random starting positions. Some of the runs converge to suboptimal extrema. However, in the majority of cases, high-quality solutions are obtained within approximately 200 seconds, and the remainder of the time is spent on small refinements with incremental improvement to the objective. This indicates that quality solutions should be obtainable with a relatively modest number of random starts, and significant time can be saved by terminating immediately after the objective plateaus.

Table 5 **Median run time over thirty iterations for various budgets and pricing policies.**

Budget (trains)	**Median Run Time (s)** *Frequencies*	*Frequencies, Line Prices*	*Frequencies, Dist. Prices*
280	260	430	386
320	224	554	386
360	241	477	436
400	234	218	238

The waiting times per commuter and mean utilities for various budget levels B and pricing policies are shown in Table 6. For comparison, the waiting times for line-pricing (cf. equation (12)) and distance-pricing (cf. equation (13)) policies are compared to those under a pure frequency-setting policy, where all prices are fixed to a flat fare, and the system optimum (cf. equation (19)). These are the respective upper and lower bounds on optimal policy performance. The mean utilities for the flat-fare policy are evaluated with the price of \$2.50. Note that under a flat-fare policy with the MNL choice model, the choice probabilities within the network are independent of price, so this price was set to an arbitrary but reasonable value to contextualize the utility values.

Both the line-pricing and distance-pricing policies produce substantial decreases in waiting time: for example, at $B = 280$, the distance-pricing policy is able to reduce the waiting time per person by 1.48 minutes, and the line-pricing policy is able to lower this further by 0.44 minutes, for a total improvement of 1.92 minutes per commuter. However, this improvement does not come for free at budget values of 360 trains and below, where both line pricing and distance pricing result in lower mean utility values for the passengers. At the highest budget of 400 trains, low prices are sufficient to manage waiting times, and so line pricing and distance pricing both show higher utility values than the flat-fare policy. These results show how a transit operator might weigh whether gains in system efficiency are worth any extra disutility associated with the variable pricing schemes. Here, distance-based pricing provides a comparable waiting time to line-based pricing, while providing a lower mean commuter utility.

Table 6 Optimal waiting times per commuter and mean utilities under the system optimum (SO), frequency-setting without pricing (F), coordinated frequency-setting and line pricing (F-LP), and coordinated frequency-setting and distance pricing (F-DP). The network of study was the full Kanagawa network.

Budget (trains)	**Wait / Commuter (min)** *SO*	*F*	*F-LP*	*F-DP*	**Mean Utility** *F*	*F-LP*	*F-DP*
280	6.73	10.22	8.30	8.74	-39.0	-43.9	-39.3
320	2.24	5.25	3.41	3.85	-36.5	-40.2	-37.3
360	0.06	1.67	0.43	0.78	-34.4	-39.7	-35.1
400	0.00	0.06	0.00	0.00	-33.8	-32.0	-32.2

5.2. Case Study: Boston

We now turn to a transit setting in Boston, Massachusetts. The motivation of this second study is to show applicability for different transit networks and to contrast with the Kanagawa results. In Boston, there are fewer route choices available than in Kanagawa; the potential gains are therefore smaller, but still substantial. Transit in the greater Boston area consists of both buses and the subway, run by the MBTA. Buses are typically numbered, and the subway system consists of five intersecting lines called the Red, Orange, Green, Blue, and Silver lines. With this network, we move beyond networks with trains as the single mode of transit, to a fully multi-modal setting.

The MBTA provided us with private data on the number of entries and exits hourly at every station, for both the buses and the subway system. We ran a generative model based on the approach in Bertsimas and Yan (2018) to produce peak-hour origin-destination demand matrices on weekdays. Again, we focus on the rush-hour from 4-7pm when the system is at capacity, where there is the potential for pricing mechanisms to influence commuting behavior and reduce network congestion.

We focus on the core of the metropolitan area, which accounts for the main volume of commuters and where multiple alternatives in route choices exist. To this end, we include bus service 1, as well as the Red, Orange, and Green subway services. This corresponds to a transit network that comprises 74 stops along 4 services. A map showing the core component of our subnetwork is shown in Figure 4. Since the subway and bus lines in Boston interact across a smaller area than in the Kanagawa network, there are fewer multiple-option commuters. In order to model more route choices for the commuters, we assumed that certain stations that were within a 0.5 mile walk of each other were substitutable for each other. For example, much of the Green and Orange lines lie close together, so commuters might have the option of taking either line. These substitutable stations are shown in dotted boxes in Figure 4. After considering these substitutable stations, we found that of the 95,760 commuters in this network during the 4-7pm rush hour period, 28,020 (29.2%) of them are multiple-option commuters, and of the 362 origin-destination pairs with nonzero demand, 100 (27.6%) of them are multiple-option commutes.

The optimization formulation (3) largely treats buses and trains the same; the main difference is that the capacity K^ℓ tends to be smaller for buses, and that buses tend to be operated at lower costs. In the computational experiments that follow, we determine pricing and frequency-setting for both trains and buses jointly, assuming that buses operate at 30% of the capacity and at 30% of the cost of trains.

As before, we consider the pricing policies corresponding to line-based fare (12) and distance-based fare (13), and compared them to a frequency-setting policy with a flat fare. Again, all policies were compared to the lower bound of the system optimum (19). For each model, we ran the first-order method (23) with 30 different random starting points and chose the schedule frequencies and prices with the lowest waiting time. To compare the models with different operating budgets, we ran it across a range of budgets from the equivalent of 45 to 60 trains, and present the waiting times per commuter and mean utilities in Table 7. Again, the mean utilities for the flat-fare policy are evaluated with the price of $2.50.

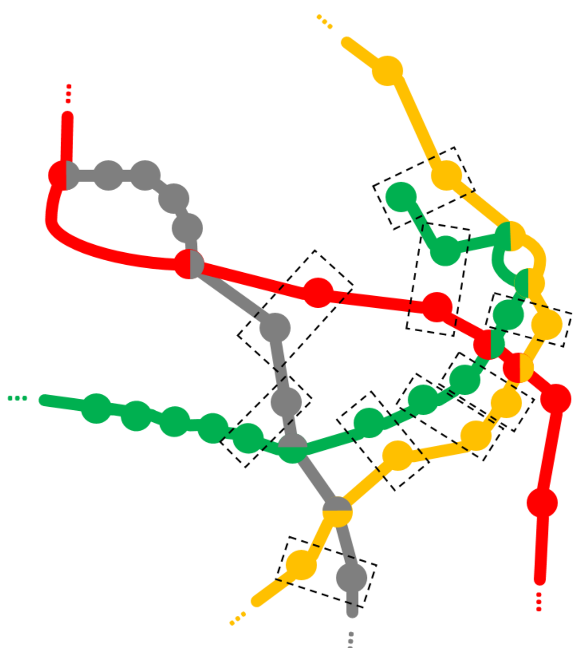

Figure 4 **(Color online) A subset of the MBTA network used in Section 5.2, showing parts of the Red, Green, Orange, and 1 bus (gray) lines. Stations further outside of the metropolitan core of the network, and some 1 bus stations, are omitted for clarity.**
To generate more route options for commuters, some stations that were within a 0.5 mile walk of each other were assumed to be substitutable for each other. Stations that were viewed as substitutable are surrounded with dotted boxes.

Comparing the performance of the models, there is a consistent gap between the frequency-setting model alone and the system optimum. As was the case with the Kanagawa network, both line pricing and distance pricing improve upon the model with only frequency-setting, although this time, the line-pricing and distance-pricing policies are comparable in performance. Distance pricing also comes with a slight reduction in utility as compared to the flat fare. However, line pricing is able to find a lower set of differentiated prices that also improve system performance, and thus is associated with the highest utility values. This differs from what we saw in the Kanagawa network, showing that there is no individually dominant pricing scheme across multiple networks. The improvements in waiting time for joint frequency-setting and pricing are also less substantial than they were in the case of the Kanagawa network, which is likely due to the fact that there are both relatively fewer route options and fewer multi-option commuters in Boston. Nevertheless, the addition of distance pricing decreases waiting time by 5.5% in the case of $B = 55$, showing that coordinated frequency-setting and pricing can provide gains even on simpler networks.

Table 7 Optimal waiting times per commuter and mean utilities under the system optimum (SO), frequency-setting without pricing (F), coordinated frequency-setting and line pricing (F-LP), and coordinated frequency-setting and distance pricing (F-DP). The network of study was the Boston network.

Budget	Wait / Commuter (min)				Mean Utility		
(trains)	*SO*	*F*	*F-LP*	*F-DP*	*F*	*F-LP*	*F-DP*
45	7.49	9.86	9.48	9.54	-40.0	-38.6	-44.5
50	4.24	6.08	5.80	5.79	-38.5	-37.5	-43.6
55	2.28	3.85	3.67	3.64	-36.3	-35.3	-40.4
60	0.82	2.00	1.73	1.82	-33.8	-32.6	-38.4

6. Conclusion

In this paper, we have presented a general framework for jointly optimizing schedule frequencies and prices on multi-modal transit networks. With modern developments in transit services, it has grown of increasing importance to other cities to use whatever means available to them to operate under tight constraints. Our framework accounts for the feedback loop that occurs when passengers respond to frequency-setting and pricing decisions made by the transit operator through the multinomial logit model. The passenger choices induce demand for transit services, and the transit operator seeks to efficiently service this demand by minimizing waiting time in the transit system.

We solve our formulation using first-order methods, and demonstrate its tractability on two real-world transit networks in Tokyo, Japan and Boston, Massachusetts. To our knowledge, ours is the first paper that addresses joint frequency-setting and pricing optimization for public transit networks and at scale. As the computational experiments are performed on a mixture of synthetic and real data, it is difficult to make direct comparisons with the current system. Nonetheless, we are able to evaluate the performance of multiple frequency-setting and pricing policies against a lower bound given by the system optimum, illustrating the potential benefits that joint frequency-setting and pricing can have in inducing passengers to make decisions that benefit the system as a whole. In future work it may be useful to consider how pricing may impact farebox recovery and transit ridership, as well as to compare the impact of varied pricing schemes across more cities.

Acknowledgments

The authors are thankful to collaborators from Mitsubishi Electric Corporation for many productive discussions, and also for providing data from Japan. Research partially supported by a grant from Mitsubishi Electric Corporation. The authors also thank the anonymous reviewers whose comments substantially improved the manuscript.

Appendix A: Input Parameter Estimation

The MNL model requires estimation of the $\boldsymbol{\beta}$ coefficients that represent the commuter's valuation of time and money, cf. equation (15). Although estimation of commuter utility functions is not the focus of this paper, we estimated $\boldsymbol{\beta}$ coefficients in a reasonable and intuitive way, guided by survey data and basic economic principles (Layard et al. 2008).

For ease of presentation, the model outlined in Section 4 assumes a homogeneous population with a shared utility function. However, generalizing to multiple segments of the population that are each described by a different utility function is straightforward, and it is the model we use in our computational experiments. In this section, we will take $\boldsymbol{\beta}$ to be a $G \times 2$ vector, where G is the number of commuter segments, and β_1^g and β_2^g represent the values of time and money for commuter segment g, respectively. The commuter segments are taken to be the different income levels in census surveys.

To estimate reasonable $\boldsymbol{\beta}$ values, we propose the following relationships. First, for simplicity, we assume that β_1^g and β_3^g, the marginal utilities for time and comfort, are the same for all income groups: extra time or comfort makes everyone equally happy, but they vary in their ability and willingness to pay for it. We arbitrarily set $\beta_1^g = 1\ \forall g = 1, \dots, G$. The value of money, β_2^g, should clearly vary with income, so we propose the relationship

$$\frac{\beta_2^g}{\beta_2^{g'}} = \frac{y_{g'}}{y_g}, \tag{24}$$

where y_g represents the income of income group g. Equation (24) indicates that someone with \$10,000 of income experiences 10 times the incremental utility for an extra dollar as compared to someone with \$100,000.

To relate β_1^g and β_2^g, we turn to a travel survey by Nie et al. (2010). When survey-takers were asked to list the most important factors in their route choice, 81.3% answered that time was important, 31.5% answered that cost was important, and 24.7% answered that comfort was important. The numbers do not add up to 100% because survey-takers were allowed to list multiple important factors. There were additional factors of reliability, safety, and emissions that could influence the route choice, but since those are exogenous to the model, we focus solely on the time and cost factors. The answers to the travel survey motivated us to weight the utility coefficients accordingly as follows:

$$\frac{\sum_{g=1}^{G} \pi_g \beta_2^g}{\sum_{g=1}^{G} \pi_g \beta_1^g} = \frac{81.3}{31.5}, \tag{25}$$

where π_g represents the proportion of the city population that is at income level g. Similarly, we took the marginal value of comfort to be $\frac{24.7}{31.5} = 0.784$.

Equations (24) and (25) are together G equations in G unknowns (recalling that we have set $\beta_1^g = 1$). Boston-area income levels and distributions were obtained from the publicly-available U.S. Census (U.S. Census Bureau 2015), and Tokyo-area data was obtained from a Japanese survey (Statistics Bureau 2013). Solving equations (24) and (25) produced a range of β_2^g values roughly between 0.5 and 15, which we used in the computational experiments that follow. Again, the purpose of this section was only to discuss a reasonable way of generating input parameters, and we do not claim that these values are exact.

References

Atasoy, Bilge, Matteo Salani, Michel Bierlaire. 2014. An integrated airline scheduling, fleeting, and pricing model for a monopolized market. *Computer-Aided Civil and Infrastructure Engineering* **29**(2) 76–90.

Bartholomew-Biggs, Michael, Steven Brown, Bruce Christianson, Laurence Dixon. 2000. Automatic differentiation of algorithms. *Journal of Computational and Applied Mathematics* **124**(1-2) 171–190.

Beevers, Sean D, David C Carslaw. 2005. The impact of congestion charging on vehicle emissions in london. *Atmospheric Environment* **39**(1) 1–5.

Ben-Akiva, Moshe, Michel Bierlaire. 1999. Discrete choice methods and their applications to short term travel decisions. *Handbook of Transportation Science* **23** 5–33.

Berry, Steven, Ariel Pakes. 2007. The pure characteristics demand model. *International Economic Review* **48**(4) 1193–1225.

Bertsimas, Dimitris, Julia Yan. 2018. From physical properties of transportation flows to demand estimation: An optimization approach. *Transportation Science* **52**(4) 1002–1011.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, Viral B Shah. 2017. Julia: A fresh approach to numerical computing. *SIAM review* **59**(1) 65–98.

Bovy, Piet HL. 2009. On modelling route choice sets in transportation networks: a synthesis. *Transport reviews* **29**(1) 43–68.

Boyd, Stephen, Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Brands, Ties, Erik de Romph, Tim Veitch, Jamie Cook. 2014. Modelling public transport route choice, with multiple access and egress modes. *Transportation Research Procedia* **1**(1) 12–23.

Brännlund, Ulf, Per Olov Lindberg, Andreas Nou, J-E Nilsson. 1998. Railway timetabling using lagrangian relaxation. *Transportation science* **32**(4) 358–369.

Caprara, Alberto, Matteo Fischetti, Paolo Toth. 2002. Modeling and solving the train timetabling problem. *Operations research* **50**(5) 851–861.

Castelli, Lorenzo, Raffaele Pesenti, Walter Ukovich. 2004. Scheduling multimodal transportation systems. *European Journal of Operational Research* **155**(3) 603–615.

City of Yokohama. 2011. Number of passengers getting on and off the yokohama municipal subway. URL `http://www.city.yokohama.lg.jp/koutuu/kigyo/jigyou/koukoku/pdf/jyoukoukyakusu.pdf`. [Online; accessed 9-September-2017].

Constantin, Isabelle, Michael Florian. 1995. Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *International Transactions in Operational Research* **2**(2) 149–164.

Dobson, Gregory, Shlomo Kalish. 1988. Positioning and pricing a product line. *Marketing Science* **7**(2) 107–125.

East Japan Railway Company. 2016. Passengers at jr east stations. URL `http://www.jreast.co.jp/passenger/`. [Online; accessed 9-September-2017].

Eliasson, Jonas, Lars-Göran Mattsson. 2006. Equity effects of congestion pricing: quantitative methodology and a case study for stockholm. *Transportation Research Part A: Policy and Practice* **40**(7) 602–620.

Fitzsimmons, Emma G, Ford Fessenden, K K Rebecca Lai. 2017. Every New York City subway line is getting worse. Here's why. *New York Times* URL `https://www.nytimes.com/interactive/2017/06/28/nyregion/subway-delays-overcrowding.html`.

Furth, Peter G, Nigel HM Wilson. 1981. Setting frequencies on bus routes: Theory and practice. *Transportation Research Record* **818**(1981) 1–7.

Goh, Mark. 2002. Congestion management and electronic road pricing in Singapore. *Journal of Transport Geography* **10**(1) 29–38.

Gurobi Optimization, Inc. 2016. Gurobi optimizer reference manual. URL `http://www.gurobi.com`.

Han, Anthony F, Nigel HM Wilson. 1982. The allocation of buses in heavily utilized networks with overlapping routes. *Transportation Research Part B: Methodological* **16**(3) 221–232.

Huang, Hai-Jun. 2002. Pricing and logit-based mode choice models of a transit and highway system with elastic demand. *European Journal of Operational Research* **140**(3) 562–570.

Hurdle, VF. 1973. Minimum cost schedules for a public transportation routei. theory. *Transportation Science* **7**(2) 109–137.

İmre, Şükrü, Dilay Çelebi. 2017. Measuring comfort in public transport: A case study for istanbul. *Transportation Research Procedia* **25** 2441–2449.

Layard, Richard, Guy Mayraz, Stephen Nickell. 2008. The marginal utility of income. *Journal of Public Economics* **92**(8) 1846–1857.

Lo, Hong K, Chun-Wing Yip, Quentin K Wan. 2004. Modeling competitive multi-modal transit services: a nested logit approach. *Transportation Research Part C: Emerging Technologies* **12**(3-4) 251–272.

Lubin, Miles, Iain Dunning. 2015. Computing in operations research using julia. *INFORMS Journal on Computing* **27**(2) 238–248.

McBride, Richard D, Fred S Zufryden. 1988. An integer programming approach to the optimal product line selection problem. *Marketing Science* **7**(2) 126–140.

McFadden, Daniel. 1974. The measurement of urban travel demand. *Journal of Public Economics* **3**(4) 303–328.

Newell, Gordon F. 1971. Dispatching policies for a transportation route. *Transportation Science* **5**(1) 91–105.

Nguyen, Sang, Eugenio Morello, Stefano Pallottino. 1988. Discrete time dynamic estimation model for passenger origin/destination matrices on transit networks. *Transportation Research Part B: Methodological* **22**(4) 251–260.

Nie, Yu Marco, Xing Wu, Joseph Zissman, Chanhan Lee, Michael Haynes, et al. 2010. Providing reliable route guidance: Phase ii .

Niu, Huimin, Xuesong Zhou. 2013. Optimizing urban rail timetable under time-dependent demand and oversaturated conditions. *Transportation Research Part C: Emerging Technologies* **36** 212–230.

Nuzzolo, Agostino, Umberto Crisalli, Francesca Gangemi. 2000. A behavioural choice model for the evaluation of railway supply and pricing policies. *Transportation Research Part A: Policy and Practice* **34**(5) 395–404.

Odakyu Electric Railway Co. 2016. Odakyu electric railway one-day passengers. URL `https://www.odakyu.jp/company/railroad/users/`. [Online; accessed 9-September-2017].

Parbo, Jens, Otto Anker Nielsen, Carlo Giacomo Prato. 2014. User perspectives in public transport timetable optimisation. *Transportation Research Part C: Emerging Technologies* **48** 269–284.

Prato, Carlo Giacomo. 2009. Route choice modeling: past, present and future research directions. *Journal of choice modelling* **2**(1) 65–100.

Robenek, Tomáš, Yousef Maknoon, Shadi Sharif Azadeh, Jianghang Chen, Michel Bierlaire. 2016. Passenger centric train timetabling problem. *Transportation Research Part B: Methodological* **89** 107–126.

Schéele, Siv. 1980. A supply model for public transit services. *Transportation Research Part B: Methodological* **14**(1) 133–146.

Shih, Mao Chang, Hani S Mahmassani. 1994. Vehicle sizing model for bus transit networks. *Transportation Research Record* (1452) 35–41.

Spiess, Heinz, Michael Florian. 1989. Optimal strategies: a new assignment model for transit networks. *Transportation Research Part B: Methodological* **23**(2) 83–102.

Statistics Bureau. 2013. Housing and Land Survey. URL `http://www.e-stat.go.jp/SG1/estat/NewList.do?tid=000001063455`. [Online; accessed 9-September-2017].

Szpigel, Bernardo. 1973. Optimal train scheduling on a single line railway. *Operations Research* **34**(4) 343–352.

U.S. Census Bureau. 2015. American Community Survey. URL `https://censusreporter.org/tables/B19013/`. [Online; accessed 5-April-2017].

Verbas, İ Ömer, Hani S Mahmassani. 2015. Integrated frequency allocation and user assignment in multimodal transit networks: Methodology and application to large-scale urban systems. *Transportation Research Record: Journal of the Transportation Research Board* (2498) 37–45.

Williams, Huw CWL. 1977. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and planning A* **9**(3) 285–344.

Wu, Di, Yafeng Yin, Siriphong Lawphongpanich. 2011. Pareto-improving congestion pricing on multimodal transportation networks. *European Journal of Operational Research* **210**(3) 660–669.

Wu, Di, Yafeng Yin, Siriphong Lawphongpanich, Hai Yang. 2012. Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transportation Research Part B: Methodological* **46**(9) 1273–1287.

Yan, Shangyao, Hao-Lei Chen. 2002. A scheduling model and a solution algorithm for inter-city bus carriers. *Transportation Research Part A: Policy and Practice* **36**(9) 805–825.

Yu, Bin, Zhongzhen Yang, Jinbao Yao. 2009. Genetic algorithm for bus frequency optimization. *Journal of Transportation Engineering* **136**(6) 576–583.