Optimizing ride-hailing with a mix of on-demand and pre-booked customers under distributional shift

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

October 29, 2024

Keywords: stochastic vehicle routing; pre-booking; partial demand uncertainty.

1 INTRODUCTION

Ride-hailing platforms (e.g., Uber and Lyft) recently started to offer pre-booking services targeting time-sensitive customers, e.g., on business trips. Customers can pay an additional pre-booking fee to reserve a ride in advance, and the operator commits to sending a driver to the agreed pick-up location at the specified time. From an operational perspective, pre-booked rides offer higher planning certainty for the operator, as the travel demand is known in advance. However, the commitment to serving pre-booked requests may force the operator to reject more profitable on-demand requests if the driver supply cannot meet overall demand. Moreover, introducing a pre-booking service may induce travel demand in areas with traditionally low driver availability, e.g., low-demand suburban areas, leading to a shift in the travel demand distribution, i.e., creating a mismatch between the trip distribution of pre-booked and on-demand requests. Consequently, implementing a pre-booking service introduces a trade-off between higher planning certainty and the rise of unfavorable rides due to shifts in the demand distribution.

Despite the extensive literature on optimizing ride-hailing systems (see, e.g., Bertsimas *et al.* 2019), only few studies have focused on mixed-service ride-hailing, i.e., with a mix of on-demand and pre-booked requests (see, e.g., Duan *et al.* 2020, Abkarian *et al.* 2022, Elting & Ehmke 2021), particularly regarding shifts in the travel demand distribution. To address this research gap, we propose a novel two-stage stochastic optimization model that maximizes the sum of pre-booking fees and the expected total profit from serving pre-booked and on-demand requests. Specifically, the operator decides to accept or reject incoming pre-booking requests in the first stage, before assigning vehicles to requests and planning routes in response to incoming on-demand requests in the second stage. Following a data-driven approach, we present a sample average approximation (SAA) formulation and a scalable heuristic algorithm. We conduct experiments based on New York City yellow taxi data and derive managerial insights indicating that the proposed mixed-service system can lead to increased profit compared to the purely on-demand baseline system, even in environments with strong distributional shifts.

2 METHODOLOGY

We consider a profit-maximizing ride-hailing operator who centrally controls a fleet of vehicles V of fixed size. A ride request is a tuple $i = (o_i, d_i, s_i, e_i, p_i)$, indicating its origin o_i, destination d_i, pick-up time window $[s_i, e_i]$, and price p_i paid to the operator. Each vehicle $v \in V$ is associated with an initial available time s_v and location o_v. Pre-booking requests enter the system before the start of the operating period, and the operator decides to accept or reject each request. By accepting a pre-booking request, the operator commits to serving it at its earliest pick-up time. During the operating period, on-demand requests enter the system and the operator decides to accept or reject each request and dispatches vehicles to serve all accepted requests, i.e., both pre-booked and on-demand. Note that the operator makes pre-booking acceptance decisions without knowledge of future on-demand requests and before determining vehicle routes.

We introduce a two-stage stochastic optimization formulation. The first stage involves acceptance decisions of pre-booking requests before the start of operations, while the second stage corresponds to the operating period and involves acceptance decisions of on-demand requests, the assignment of requests to vehicles, and routing decisions. Let D_1 be the set of all pre-booking requests and D_2 be the uncertain set of all on-demand requests that arrive throughout the operating period. Let x_i be a binary variable, taking value 1 if the operator accepts pre-booking request $i \in D_1$. Each pre-booking request $i \in D_1$ incurs an additional pre-booking fee $p_{i,1}$. The two-stage stochastic optimization problem maximizes the revenue from pre-booking fees and the expected profit from the second-stage problem:

$$\max_{\mathbf{x}\in\{0,1\}^{|D_1|}} \quad \sum_{i\in D_1} x_i\, p_{i,1} + \mathbb{E}_{D_2}[\Pi(\mathbf{x}; D_1, D_2)], \tag{1}$$

where $\Pi(\mathbf{x}; D_1, D_2)$ is the optimal objective value from the second-stage problem.

We are interested in analyzing the pre-booking acceptance decisions of the first-stage problem. To not bias our analyses, we compute an upper bound for the second stage by assuming that all on-demand requests are revealed simultaneously, corresponding to a setting with perfect foresight. Since the true underlying distribution of on-demand requests is unknown, we approximate the second-stage objective value by SAA. Let $\mathcal{K}$ be a set of scenarios describing possible realizations of the on-demand request set D_2^k, for $k \in \mathcal{K}$. We introduce the SAA model formulation:

$$\max_{\mathbf{x}\in\{0,1\}^{|D_1|}} \quad \sum_{i\in D_1} x_i\, p_{i,1} + \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \Pi(\mathbf{x}; D_1, D_2^k), \tag{2}$$

where we replace the expectation in Model (1) with the average second-stage objective value among the scenarios.

We define the second-stage problem on a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where each request $i \in D_1 \cup D_2$ is represented by a node. An arc $(i,j) \in \mathcal{A}$ represents the possibility for a vehicle to pick up customer j immediately after serving request i. Each arc $(i,j) \in \mathcal{A}$ has a travel time T_{ij} representing the time to serve request i and to drive from i's destination d_i to j's origin o_j. Each arc $(i,j) \in \mathcal{A}$ also has a profit π_{ij} corresponding to the fare paid by j minus the cost of driving from d_i to o_j (dead-heading) and from o_j to d_j (occupied trip). In addition, each vehicle $v \in V$ is represented by a node in $\mathcal{V}$, and an arc $(v,i) \in \mathcal{A}$ represents the possibility for request $i \in D_1 \cup D_2$ to be the first request picked up by vehicle $v \in V$. Each arc $(v,i) \in \mathcal{A}$ has a travel time T_{vi} from the initial location of vehicle v to the origin o_i of request i, and a profit π_{vi}, i.e., the fare paid by i minus the cost of driving from o_v to o_i and from o_i to d_i.

The second-stage vehicle dispatching problem assigns a sequence of requests to each vehicle, maximizing the operator's profit subject to routing and pick-up time window constraints. Let y_i be a binary decision variable that takes value 1 if the operator accepts the on-demand request $i \in D_2$. We represent routing decisions by binary variables z_{vi} and z_{ij}, where $z_{vi} = 1$ if request i assigned to vehicle v as a first request in the sequence, and $z_{ij} = 1$ if request i is picked up by a vehicle immediately after request j. Variable $t_i \in [s_i, e_i]$ models the pick-up time of request i. Given pre-booking and on-demand request sets, D_1 and D_2, and first-stage decisions $\mathbf{x}$, we formulate the second-stage problem as follows:

$$\Pi(\mathbf{x}; D_1, D_2) = \max \quad \sum_{(i,j)\in\mathcal{A}} \pi_{ij}\, z_{ij} \tag{3}$$

$$\text{s.t.} \quad y_i = x_i \qquad \forall i \in D_1 \tag{4}$$

$$y_j = \sum_{i \in N^-(j)} z_{ij} \qquad \forall j \in D_1 \cup D_2 \qquad (5)$$

$$\sum_{j \in N^+(i)} z_{ij} \leq y_i \qquad \forall i \in D_1 \cup D_2 \qquad (6)$$

$$\sum_{i \in N^+(v)} z_{vi} \leq 1 \qquad \forall v \in V \qquad (7)$$

$$t_j - t_i \geq (s_j - e_i) + (T_{ij} - (s_j - e_i))z_{ij} \qquad \forall (i,j) \in \mathcal{A} \qquad (8)$$

$$s_i \leq t_i \leq e_i \qquad \forall i \in D_1 \cup D_2 \cup V \qquad (9)$$

$$z_{ij} \in \{0,1\}, \quad y_j \in \{0,1\} \qquad \forall (i,j) \in \mathcal{A} \qquad (10)$$

where $N^+(i) = \{j \in \mathcal{V} : (i,j) \in \mathcal{A}\}$ is the set of nodes that are reachable by the arcs going out of i and $N^-(j) = \{i \in \mathcal{V} : (i,j) \in \mathcal{A}\}$ is the set of nodes that can reach j by its incoming arcs. Constraints (4) are linking constraints and ensure that pre-booking requests are served (or not) in accordance with first-stage decisions. Constraints (5)–(7) ensure the conservation of flow and that each request is assigned to at most one vehicle. Constraints (8) ensure that the pick-up time windows are satisfied, where we assume, w.l.o.g., that $t_v = s_v = e_v$ for vehicle $v \in V$, and we enforce that wait times for pre-booked customers are equal to zero by setting $t_i = s_i = e_i$ for all $i \in D_1$. Constraints (9)–(10) define the variable domains.

The proposed SAA model is a challenging optimization problem as it requires the solution of integer subproblems that must satisfy linking constraints with respect to first-stage decisions. To address these challenges, we develop a scalable heuristic algorithm that separately solves approximations of the second-stage subproblems, where we fix customer pick-up times, and remove the linking constraints related to first-stage decisions. To define the approximate second-stage subproblem, we rely on a directed dispatching graph $\widehat{\mathcal{G}}$. We place an arc (i,j) in $\widehat{\mathcal{G}}$ if $e_i + T_{ij} \leq e_j$, which ensures that pick-up times are fixed to $t_i = e_i$ for all $i \in D_1 \cup D_2$, resulting in an acyclic graph. Then, each resulting approximate subproblem can be solved as a K-disjoint shortest path problem (K-DSPP), which can be solved in polynomial time, e.g., as proposed in Schiffer *et al.* (2021).

Our heuristic algorithm separately solves the approximate subproblem corresponding to each scenario and then determines first-stage decisions by combining the individual solutions to the subproblems. At a high level, the heuristic algorithm consists of three main steps:

1. **Solving the approximate subproblems:** We separately solve the approximate subproblem for each scenario.
2. **Consensus fixing:** We determine a consensus among the subproblems regarding first-stage decisions by majority voting.
3. **Arc weight adjustment and re-solving:** We adjust arc weights, i.e., profits, based on the first-stage decisions and re-solve each subproblem, with the aim of enforcing consistent first-stage decisions among the scenarios. Specifically, for each pre-booking request $j \in D_1$, we update the profit π_{ij} associated with arc (i,j) by adding a sufficiently large constant if j is accepted. Otherwise, if j is rejected, we subtract a sufficiently large constant from π_{ij}.

3 RESULTS

We adopt the data set from NYC Taxi & Limousine Commission (2010) for our case study, focusing on the Yellow Cab rides of Tuesday, January 26, 2010, from 9:00 a.m. to 3:00 p.m. Rides from and to Manhattan constitute 92.17% of all rides in the considered time horizon. In order to study distributional shifts in the travel demand, we also consider rides from or to the boroughs of Brooklyn, Queens, and The Bronx. We analyze the effect of distributional shifts by sampling pre-booking requests from a different distribution than on-demand requests.

Specifically, given the number of pre-booking requests m and given the share of pre-booking requests ν that are outside of Manhattan, i.e., with pick-up or drop-off locations outside of Manhattan, we construct D_1 by sampling $m\nu$ requests from boroughs outside Manhattan and $m(1-\nu)$ requests from within Manhattan. Values of ν close to 0 correspond to settings with weak distributional shifts while values close to 1 correspond to strong distributional shifts.

We compare the performance of the SAA model against two baseline policies. First, the *accept-all* policy accepts all incoming pre-booking requests regardless of their attributes. Second, the *reject-all* policy corresponds to the setting in which the operator does not offer a pre-booking service. In this case, we assume that customers who would have pre-booked a ride request on demand instead, ensuring that the total number of requests remains consistent across policies.

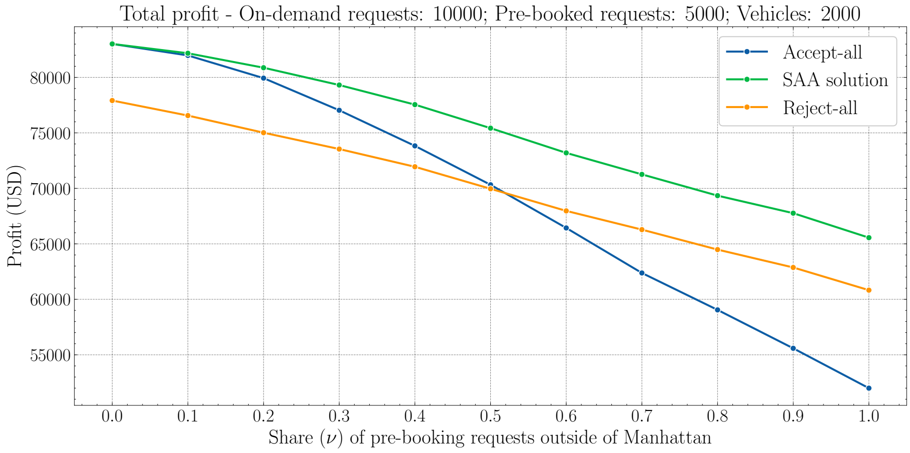

Figure 1 – *Out-of-sample performance of the SAA model*

Figure 1 shows the total profit from the SAA model and the baseline policies for increasing ν. We observe several interesting insights. First, the SAA solutions have significantly higher profit than the reject-all baseline, regardless of the parameter value ν. Accordingly, optimizing the mixed-service system based on the SAA model leads to an increase in profit compared to the purely on-demand system, regardless of how strong the distributional shift is. Second, greedily accepting all pre-booking requests can lead to a significant decrease in profit compared to the purely on-demand system, especially in regimes with stronger distributional shifts ($\nu \geq 0.52$). Third, in settings with weak distributional shifts, the greedy accept-all policy is more profitable than the reject-all policy, showing performance comparable to the SAA solutions. Further results and managerial insights will be presented at the conference, including analyses of structural properties, sensitivity analyses, and results from different heuristic policies.

References

Abkarian, Hoseb, Mahmassani, Hani S., & Hyland, Michael. 2022. Modeling the mixed-service fleet problem of shared-use autonomous mobility systems for on-demand ridesourcing and carsharing with reservations. *Transportation Research Record*, **2676**(8), 363–375.

Bertsimas, Dimitris, Jaillet, Patrick, & Martin, Sébastien. 2019. Online vehicle routing: The edge of optimization in large-scale applications. *Operations Research*, **67**(1), 143–162.

Duan, Leyi, Wei, Yuguang, Zhang, Jinchuan, & Xia, Yang. 2020. Centralized and decentralized autonomous dispatching strategy for dynamic autonomous taxi operation in hybrid request mode. *Transportation Research Part C: Emerging Technologies*, **111**, 397–420.

Elting, Steffen, & Ehmke, Jan Fabian. 2021. Potential of shared taxi services in rural areas – a case study. *Transportation Research Procedia*, **52**, 661–668.

Schiffer, Maximilian, Hiermann, Gerhard, Rüdel, Fabian, & Walther, Grit. 2021. A polynomial-time algorithm for user-based relocation in free-floating car sharing systems. *Transportation Research Part B: Methodological*, **143**, 65–85.